PUBLICACCESS.EU STUDY

ENSURING INTEGRATED ACCESS TO ALL PUBLICLY AVAILABLE EU DOCUMENTS

FINAL version

Author:
AION CS, s. r. o.
Nám. T. G. Masaryka 1280
760 01 Zlín
Czech Republic

Document date: December 2, 2016
Document version: 2.1

Contents

1. Executive summary

Problem statement and structure of the study

Publicly available documents from institutions, agencies and bodies of the EU are scattered across a large number of different repositories – different document sources. In order to find required information, citizens usually have to perform searches in a number of separate document sources, which can be time-consuming and cumbersome.

The main aim of the present study, launched as part of a multi-stage project under the banner of **'PublicAccess.eu'**, is to provide insight into how this fragmentation of the EU information ecosystem can be tackled. To this end, the study provides an overview of the current situation, using as source material a number of document sources belonging to a range of EU institutions, agencies and bodies, and proposes a number of options for the final **integrated access solution**, which would provide citizens with integrated access to all publicly available EU documents in an easy, clear and understandable way. The study is complementary to a number of ongoing related initiatives led by EU institutions in the area of access to documents, notably the landscaping exercise for the legislative domain (a comprehensive overview of tools, data standards and specifications used in the EU legislative process) and the European Commission's Better Regulation initiative, although its scope goes beyond documents related to the legislative process, which are the focus of these initiatives.

The study consists of two main parts. The first part is analytical and it contains detailed overviews of **27 document sources** belonging to **14 different EU institutions, agencies and bodies**. Based on the analytical part, the second part provides four alternatives for the possible integrated access solution, with "big picture" descriptions showing how each of them may be developed and further elaborated. Furthermore, the study is accompanied by a database in which the findings are presented in more detail and organised using an ontological approach.

Core analytical findings

The analysis has revealed that overall there are more than **4.5 million documents** in all analysed document sources (this number includes only one language version of the document; the number of documents including all language versions is considerably higher). The volume per document source ranges from **320** (ERC Executive Agency) to **3.1 million** (EUR-Lex and TED). The analysis has also shown that there is significant fragmentation as regards the content and context of documents, most evident in the quality of **metadata**, which can be improved by the use of **controlled vocabularies**, and as regards the existence of tools allowing **machine readability**.

A **vocabulary** is a file or a list of organised and named metadata entries which serves as a description of document properties, allowing better searchability and classification of documents. The analysis has shown that a total of **146** different vocabularies are used in all the document sources investigated. While this high number of vocabularies is not an issue in itself, what is problematic is the fact that frequently the vocabularies and their entries are not specifically defined (e.g. as in the case of the Metadata Registry used in EUR-Lex). The names of the entries in vocabularies are often not self-explanatory, proper clarification is missing and thus they are sometimes not easily understandable. Moreover, different vocabularies may describe the same type of information, but names of their entries across document sources sometimes differ. Even in cases where different vocabularies provide the entry with the same name, it is not possible to rely on the fact that these same names represent the same values. This means that a significant harmonisation effort would be required to bring these different vocabularies from separate document sources into one integrated access solution.

As regards **attributes**, i.e. the **metadata** captured in the form of specific values of variables, a total of **61 attributes** are used in the document sources of the EU institutions and agencies which have been investigated. The analysis has shown that apart from the meta-information from the group 'Date', all other groups of metadata attributes can be considered as dedicated to specific document sources. Therefore, similarly to the vocabularies, a large amount of documentary work focused on metadata harmonisation would be required before the implementation of any integrated access solution.

Finally, for each document source a summary concerning the **possible re-use options** in the potential integrated access solution has been included. In this context the study focuses on the existence of tools allowing **machine readability** and on **how the content of documents is provided** (e.g. as attachments or in HTML format).

As for **machine readability**, only portals managed by the Publications Office (EUR-Lex and TED) provide an API, while there are only five other document sources that provide the RSS function. Improvements in this area would require a significant amount of effort from the administrators of the various document sources, but the investment would be justified not only by the increase in re-usability for the integrated access solution, but also for 3rd party systems.

As to **how the content is provided**, in the vast majority of cases documents are available only as file attachments (PDF, DOC, etc.). However, this method does not allow for optimal access to the content of documents and for this reason conversion into more flexible formats would be desirable.

Four possible integrated access solutions

Bearing in mind that the overall goal of an integrated access solution would be to establish a new level of integrated access service focused on end-users, it would have to combine the advantages of a single search in a number of document sources with a comprehensive web portal functionality allowing quick and easy display of documents. In this context, and based on the analytical findings outlined above, the study provides a high-level overview of some key elements of the possible architecture of the integrated access solution. Firstly, the need for the **automated retrieval of documents from current systems** is presented, together with the necessity to define the overall scope and the method for transmitting documents or their metadata to the integrated access solution. Secondly, **automated reprocessing of current metadata into a single ontology** is discussed as an optimal way to let the different document sources communicate with each other. Thirdly, **automated content (text) parsing** allowing a move from file attachments (in PDF, DOC, etc.) to structured formats is explained as another important building block of the future architecture. Next, **intuitive management** of the content and context of documents is discussed, and finally, the need for a **modern front-end web and mobile application** based on users' needs is emphasised, providing users with top-of-the-range functionalities.

Taking into account the above requirements, the study proposes the following integrated access alternatives:

- Alternative No. 1: Web application built upon the common repository of the Publications Office (the CELLAR)

This solution is based on the use of the common repository of the Publications Office, the CELLAR, as the common repository of the future integrated access solution. At least the context of the documents, i.e. a standardised core set of descriptive metadata including a link to the content of the document itself, should be transmitted to the CELLAR by means of the IMMC Exchange Protocol. Each institution would need to establish a transfer channel for pushing documents (content/context) from its document source into the CELLAR.

The information required by the integrated access solution would need to be defined by a common knowledge model. The core component of this solution would be the synchronisation process between the CELLAR and the integrated access solution database, i.e. the replication of content in a structured format and of the related subsets of the metadata that are needed to comply with the common knowledge model, from the CELLAR to the integrated access solution database. Documents that are not physically stored in the CELLAR but referred to by a link would need to be crawled by the integrated access solution during the synchronisation process.

Finally, a new front-end web application providing content from the integrated access solution database would have to be built based on the common knowledge model.

The solution based on the CELLAR has moderate costs of development and deployment, with easy addition of new document sources. It would, however, require further metadata harmonisation within EU institutions, and the development of transmission channels would require a significant level of commitment from the relevant institutions, agencies and bodies. Implementation of this solution would also take longer than most of the remaining solutions proposed.

- Alternative No. 2: Decentralized federated search

This alternative is based on a set of decentralised independent search engine gateways that would have to be implemented on top of each document source, including the CELLAR. The interoperability of the search engine gateways would need to be guaranteed. This would allow direct searching in existing document sources, i.e. the new integrated access web application would pass a series of queries derived from end-user's request to a particular search engine gateway (or gateways), processing their responses and then presenting the results as a unique harmonised result list.

In this alternative all document data (content and context) would remain in document sources while the integrated access solution would only serve as an indexing and presentation layer. This could be regarded as one of the strengths of this alternative, as no additional effort would be needed as regards updates of the data.

However, this solution would require significant technical improvements of the document sources, so that they would be able to process end-users' requests. The overall outcome of this solution would depend on the APIs of the external search engine gateways and their performance. Also, this solution would provide only limited support for mobile users. As each document source is unique, it would be necessary to change IT infrastructure of some document sources to achieve the ability to implement the search engine gateway.

- Alternative No. 3: Centralised federated search

This alternative envisages the implementation of a new central search engine that indexes all document sources. The integrated access application would pass a query corresponding to end-user's request to the central search engine which would process the result and push it to end-user's device. All content would remain in the document sources and the integrated access solution would be used as the presentation layer.

The strength of this solution is that its performance is independent of document sources. However, the necessary condition would be a significant technical improvement of the document sources to enable indexation by the central search engine. Access to content would mean in this case access to content that is stored in a decentralised way by the different document sources and consequently the overall stability and reliability of this solution would be dependent on the stability and reliability of the different document sources.

- Alternative No. 4: Content harmonisation on the central level

This alternative is based on alternative No. 1 and provides some extensions and improvements to it. It envisages harmonising the entire content that is accessible through the integrated access solution, i.e. presenting it in a unified structured form (HTML5). For content that is not available in the unified structured format at the source, the PublicAccess.eu solution would support the necessary conversions.

This alternative consists of a robust integrated access database for storing all context information (metadata) and the unified content that is either obtained directly from the document sources or by conversion of legacy formats in the scope of the integrated access solution. All data obtained from the different document sources would be organised according to the common knowledge model of the integrated access solution. The database would be fed preferably from the CELLAR, or, in exceptional cases, directly from document sources.

Another key element of this solution would be the integrated access web application, which would pass a query derived from end-user's request to the integrated access database, process its response and push the results to end-user's device in a unified form. This application would provide a number of useful innovative features, such as maximal ergonomics and usability, intuitive search, personalisation and continuous adjustments of the solution (e.g. based on users' activities and in line with search engine requirements).

Final recommendation

Following the analysis of the four proposed alternatives, it appears that the alternative with satisfactory user experience, acceptable risk level and an excellent potential for growth is **alternative No. 1 - PublicAccess.eu solution built upon the CELLAR**. Moreover, some components of this alternative have already been tested in practice, since the CELLAR is being used as the repository for other access solutions such as EUR-Lex. However, its implementation would require a significant level of commitment from the relevant institutions, agencies and bodies, and it would be quite time-consuming. **Alternatives No. 2 and 3** also appear interesting, in particular due to their short implementation time, and for this reason it would be useful to test them by means of a proof of concept exercise in order to see whether they are capable of bringing the expected benefits.

2. Introduction

2.1. General introduction

All citizens of the EU and any natural or legal persons residing or having a registered office in a Member State have the right of access to the EU institutions documents protected by law. The right to public access to the EU documents is stipulated in Article 15(3) of the Treaty on the Functioning of the EU ('TFEU') and further specified in Regulation 1049/2001. Furthermore, it is upheld as a fundamental right in the EU's legal order, stated in Article 42 of the Charter of Fundamental Rights of the EU ('Charter'). The scope of the right to access encompasses both legislative and administrative procedures and accessibility is required to be as wide as possible.

The effective implementation of the public's right to access documents held by the EU institutions is one of several means by which the EU seeks to increase transparency and accountability towards its citizens. Currently, the majority of EU institutions have public document registers in place. However, the quality of these registers varies greatly, undermining their accessibility and usability by citizens, journalists, academics, civil society organisations, etc. Improvements are needed, for example, in how well the users can search within these registers so that already published documents can actually be found.

In this context, in 2016 the European Parliament launched a pilot project to facilitate and improve online access to a wider range of unclassified documents held by EU institutions, agencies and bodies and to increase the transparency of EU decision-making. The objectives of the pilot project were to identify and bring together the relevant unclassified documents which can be made available, as well as to structure them in a way that ensures interoperability and linking. The project is being implemented by the Publications Office.

The present study is the outcome of the analysis performed in the framework of the project. It looks into potential methods concerning how to bring together documents from different EU institutions, agencies and bodies, and to display them in a single place. Its aim is to ensure easy, seamless and searchable access for the benefit of various stakeholders and citizens interested in EU documents. The study explored and analysed the various options, impacts, drivers, constraints and related costs. Its goal is to identify and assess the most suitable and promising options for the future integrated access solution.

Each proposed option is accompanied by the analysis of the following considerations:

- Technical (metadata harmonisation, formats, aggregation/federation of data, etc.);
- Organisational (stakeholder cooperation models, time constraints, tasks and competencies required, etc.);
- Practical (attractiveness and ease of use from the user's point of view, adaptability, etc.);
- Financial (broad comparison of financial burden associated with proposed solution, possibilities of economies of scale etc.).

The study brings a comprehensive overview of the various scenarios to develop and operate the future integrated access solution in the most effective and efficient way, considering the architectural principles of the current architecture of EU document registries, best practices, and potential risks.

2.2. Structure of the study

There are common principles for all the main chapters of the study. Each chapter starts with an annotation of its content followed by a brief descriptive subchapter so that the reader is aware of the chapter's content from the beginning.

The remainder of this report has been organised as follows:

- **Section 3** provides a summary of the primary objectives and defines the scope of the study as well as the information sources used;
- **Section 4** provides an analysis of the current situation among given and selected European institutions and agencies relevant to an integrated access solution including the major findings from the investigation of the document registers of the EU institutions. It also describes the general methods used for the research, the aspects investigated and the level of investigation mainly from the end-user's point of view supplemented with the information obtained from direct communication with the institutions/agencies;
- **Section 5** gives the results of the practical exercise related to the subject of the study. It illustrates the approach taken by a random sample of common users (law students and government officers) to find information about the specific EU legislative process.
- **Section 6** describes possible solution alternatives for the future integrated access solution. This includes the methodology used, challenges and advantages and overall recommendations;
- **Section 7** provides a comprehensive glossary of the terms and a complete list of abbreviations used in the document.

The **Annexes** present the detailed data collected during the analysis and a description of additional project outputs, i.e. study database.

2.3. 4W – What, When, Who, Why

A lot of document sources were analysed in the Study. It is obvious that individual document sources are different, with their own metadata, vocabularies, structures, etc. Thus the study tries to shed light on more than just these differences. One of the main goals is to look for similarities shared by all the document sources. For this reason, all the problems of existing data were theoretically divided into four main categories, four essential and simple questions: What, When, Who, Why. These four questions are mentioned throughout the study. They serve as 'abstract connecting points' for two basic purposes:

- To evaluate all analysed document sources within a unified structure, which allows comparison of different document sources in the same way.
- More importantly, these four questions are in fact four fundamental issues to be solved by any future integrated access solution.

Why are they regarded as 'abstract connecting points'? Because these questions do not focus on the particular document source and technical principles behind its structure, they are focused on the user and their needs. **They are constructed as theoretical groups of items answering the user's demands.**

What do these questions investigate?

What

The questions 'what' relates to the document's content. Simply put, it asks '*What is the document about*'? The study analyses and evaluates individual document sources according to the existence of vocabularies such as Document Types, Topics, Subject Matters, etc. All these (and similar) vocabularies,

named differently in individual document sources, describe the content of the document. All of them say what the document is about. For example, a document type such as 'Press release' is so self-explanatory that any user looking in the group of documents only for press releases simply filters their search according to this document type and they get what they were looking for. Moreover, documents may be further classified with topics, so filtering the search according to the document type 'Press release' and a topic such as 'Agriculture' take the user straight to what they were looking for.

When

The question 'when' is connected to all time aspects of the document. Simply stated, it asks '*When was the document published?'*. However, it is not so easy to assign a specific date to a certain document, as there could be several time aspects connected to it (such as date of creation, date of adoption, date of signing, date of publication, etc.). Future integrated access solution would need to decide methodically what kind of document would be the most relevant. From an analytical point of view, document sources use several vocabularies concerning time aspects, such as Years, Months, Dates, etc. All these vocabularies are answers to the question 'When'.

Who

The question 'who' is linked with the author of the document. Simply said, it asks '*Who is responsible for the document?'* or even simpler '*Who is the creator of the document?'*. This is one of the most distinctive criteria to be used from any integrated access alternative point of view, as the whole project is based on gathering documents from several document sources – and several authors - together. Moreover, the question 'who' is not limited to the name of the EU Institution/Agency, but it goes deeper and may provide information on organisational units or even on real people behind the document (rapporteurs in committees, EU Commissioners, etc.). Several vocabularies such as Authors, Authorities, Rapporteurs, Corporate bodies, Countries, etc. come under this 'who' question and provide (more or less) clear answers.

Why

The question 'why' tries to uncover the reasons for the document's existence. Simply put, it asks '*Why does the document exist?',* or *'What is the purpose of the document?'* This question is not so perspicuous as the other three and it covers other important characteristics of the document, such as vocabulary regarding the procedure (belonging to a particular stage in a process elucidates a document's existence – for example, 'Publication of an act in the OJ as a procedural stage in the legislative process tells the reader about the very essence of the document, which was created as the final fair copy of adopted legislation). Other vocabularies such as Cases (appurtenance of the judicial document to a particular case) or References belong to this 'why' question.

2.4. Database accompanying the study

Apart from the text of the study itself, the structure of which is described in Annex 1, there is an additional interactive output of the analysis – an interactive PublicAccess.eu study database ('study database').

Since it had been necessary to collect and further analyse large amounts of diverse information, the dedicated study database as a repository for all information gathered within the analysis was initialized.

The study database also has a presentation environment, which presents, in an interactive form, both the document and other selected aspects of the Study. Consequently, many findings and connections are better expressed and more easily understood in this interactive form, rather than in the study document itself. The basic information both about the study database ontology and its environment can be found in the Annex 1 of the study. The links to the specific positions of the study database are handled via the footnotes.

3. Objectives and scope

3.1. Objectives of the study

At this moment, publicly available documents from EU institutions, agencies and bodies are scattered across a large number of repositories, and they are made available to citizens via multiple public registers and/or websites. Each entity maintains its own repository and website, making it impossible for the citizen to perform a single search in order to access a comprehensive range of EU documents on a given theme. A direct consequence of this multiplicity of repositories and websites is that the documents are not described in a uniform way, as the vocabularies and metadata used to describe them are not always compatible. Furthermore, the documents are stored in a variety of formats (DOC, DOCX, PDF, etc.), which create an additional layer of difficulty for ensuring integrated access.

The main objective of the study is to provide a wider insight into how different types of documents from different EU institutions, agencies and bodies could be brought together in a single place for user-friendly and seamless access.

3.2. Scope of the study

The study is primarily focused on all documents currently available via the public registers and websites of the EU institutions listed below with the exception of publications, i.e. documents which carry international identifiers (ISBN, ISSN, etc.). The text of the web pages themselves was not considered to be a document for the purposes of the study.

- European Parliament (Register of documents, Legislative Observatory, InterParliamentary EU Information System)
- Council of the European Union (Register of documents)
- European Commission (Register of Commission documents, Register of Commission expert groups, Comitology register)
- Court of Justice of the European Union (Register of Case Law – InfoCuria)
- European Central Bank (website)
- European Court of Auditors (website)
- Committee of the Regions (Register of documents)
- European Economic and Social Committee (Register of documents)
- European Ombudsman (Register of Cases, Register of Resources)
- The Publications Office of the European Union (EUR-Lex and Tenders Electronic Daily)
- The Education, Audiovisual and Culture Executive Agency (website)
- European Research Council Executive Agency (website)
- The Body of European Regulators for Electronic Communications (Document Register)
- European Union Intellectual Property Office (formerly ‘The Office for Harmonization in the Internal Market’) (Case Law Register)

3.3. Information sources used

The study's methodological approach was primarily based on extensive desk research and a literature review of all available documents relevant to the study's scope and through feedback from the Publications Office on the draft deliverables.

In addition, specific technical aspects of the current document registers were clarified during the stakeholders' consultations through e-mails. The stakeholders consulted, through their representatives, included the Publications Office and other EU institutions namely:

- European Parliament
- European Commission
- European Court of Auditors
- European Economic and Social Committee
- The Body of European Regulators for Electronic Communications
- European Research Council Executive Agency

Current situation of document publication in selected EU institutions and agencies

4. Current situation of document publication in selected EU institutions and agencies

The goal of this chapter is to establish the grounds for a definition of a possible future integrated access solution. It summarizes the results of the analysis of the current situation of document publication within the following EU institutions and agencies:

- EU institutions
  - European Parliament
  - Council of the European Union
  - European Commission
  - Court of Justice of the European Union
  - European Central Bank
  - European Court of Auditors
  - European Economic and Social Committee
  - Committee of the Regions
  - European Ombudsman
- Publications Office of the European Union
- EU regulatory agencies
  - The Body of European Regulators for Electronic Communications
  - European Union Intellectual Property Office
- EU executive agencies
  - The Education, Audiovisual and Culture Executive Agency
  - European Research Council Executive Agency

This chapter analyses the publication methods of **publicly available documents** on the websites of these EU institutions/agencies. Such documents are usually placed in dedicated parts of websites labelled as document registers, document libraries etc. However, there are situations where some important documents are published outside of a dedicated document register and are spread across the website. In such cases, the analysis also covers the parts of websites containing such documents.

This chapter consists of four subchapters:

- Chapter 4.1. describes the principles and grounds on which the analysis of the EU institutions and agencies was carried out and defines the terms and abbreviations used hereafter in the study.
- Chapters 4.2. and 4.3 contain an analysis of the document registers of the given EU institutions and agencies.
- Chapter 4.4. contains the conclusion from the analysis.

The possible future integrated access solution would reflect the most important outcomes of the analysis.

4.1. Method for the analysis of the document sources

A description of how each institution/agency publishes public documents is described in its own chapter. This includes a list and short description of those sources, as well as the place where the EU institution/agency publishes its public documents, i.e.:

- The document registers of the given EU institution/agency;
- Other sources (website sections) where the documents are published.

For each source, it is declared whether the source is further analysed and to what extent. A separate chapter is devoted to each source analysed. There were two methods used to conduct an analysis of the source:

1. **Thorough analysis** of the specific document register with a large number of published documents with the following outline structure:
   - General information
     - Access links
     - Time range covered
     - Number of documents published
     - A brief taxonomical overview of investigation results
   - Document types used
   - Metadata as relationships between documents and vocabularies
     - Originator of the document (e.g. Authority, Responsible unit, Author, Country etc.)
     - Time specifications (e.g. Year, Parliamentary term etc.)
     - Characteristics of what the document is about (e.g. Topic, Subject matter)
     - Other vocabularies used (e.g. Languages, Countries etc.)
   - Metadata as document attributes
     - Dates of creation, Dates of publishing, etc.
   - Relations between documents
     - Internal relations
     - External relations
   - End-user search possibilities with accent on
     - Search forms
     - List of results
     - Document details
     - General evaluation of the Search possibilities
   - Figure of a sample document extracted from the source
2. **General analysis** with a less strict structure: the output differs according to how the analysis of that particular source of published documents was carried out. These sources typically contain documents in considerably fewer numbers.

Both types of analyses conclude by summarising whether the particular document source could be re-used in a future integrated access solution. The conclusion of the analysis of the particular document source typically evaluates the presence of important metadata groups and opportunities possibilities for machine readability.

For the purposes of the detailed analysis (especially when comparing sources between themselves and identifying overlaps between the registries) a dedicated study database was created. The study database contains all the content of the analysis and the selected vocabularies extracted from particular registers. All information is presented in an interactive, structured form. In the situations when study database contains more content than is described in the study, the study's footnotes present interactive links to the relevant sites in the study database.

4.2. Analysis of EU institutions' document sources

The EU institutions that are in the scope of this study publish documents for the public on their websites using various mechanisms (the general term 'document sources' is used in the following text).

Typically, the document sources are placed in a dedicated section of the institutions' websites. These sections have specific functionality represented by an autonomous search. The term 'document register' is used for such sections. Some institutions have more than one document register.

Some EU institutions do not store all their documents in the document register but publish them across their websites.

The goal of the following analysis is to investigate all the ways in which the documents are published by each institution, and to break down the selected document registers or the special sections of the EU institution's website.

Each document source is handled in a separate subchapter following the structure specified in the chapter Method for the analysis of the document sources (see 4.1).

4.2.1. European Parliament

Composed of 751 directly-elected Members of the European Parliament ('MEPs') from 28 countries, the European Parliament ('EP') represents EU citizens. It acts as a co-legislator with the Council of the European Union on nearly all EU law and holds the other EU institutions to account.

Three EP document sources were analysed:

1. European Parliament's Register of Documents
2. Legislative observatory
3. IPEX

There are more documents in the website's special sections[1] which were analysed in the study database only.[2]

4.2.1.1. European Parliament's Register of Documents

4.2.1.1.1. General information

The European Parliament's Register of Documents ('RD-EP') has references to documents produced or received by the EP.

The **thorough analysis** of RD-EP was carried out.

4.2.1.1.1.1. Public access

The RD-EP is accessible at this general URL address:
http://www.europarl.europa.eu/RegistreWeb/search/simpleSearchHome.htm

[1] European Parliament – News
European Parliament – Think Tank
European Parliament – Committees
European Parliament - Plenary
European Parliament – Members of Parliament (MEPs)
European Parliament - Delegations

[2] The European Parliament website's special sections in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1098926.

A list of all documents is available at this URL address:
http://www.europarl.europa.eu/RegistreWeb/search/simple.htm?currentPage=1&datepickerEnd=31%2F12%2F2016

The public has direct access to the vast majority of these documents in their electronic form. Electronic access is free and no special authentication is needed.

Some public documents are not instantly available and this unavailability is clearly stated. However, the user can ask for such documents via the 'Request a document' form.

4.2.1.1.1.2. Time range covered

RD-EP began operating on 3 December 2001. However, some older documents are also available in RD-EP, dating back to 1991.

4.2.1.1.1.3. Overall volume of published documents

The volume of published documents up to the end of 2015 is as follows:

- ca 585 000 documents in the English language
  - ca 50 000 documents added each year
- ca 7 900 000 documents in all languages (including English)

4.2.1.1.1.4. Brief investigation of the RD-EP

Figure 1 shows the result of the taxonomic comparative analysis of the RD-EP[3].

[3] Brief investigation of the RD-EP in the study database:
http://atom.ts-publicaccess.eu/form/item?ItemId=1033761.

Figure 1: Overview investigation of the RD-EP

4.2.1.1.2. RD-EP Document types

The document types in the RD-EP[4] are stored in a hierarchically organised vocabulary (taxonomy of its Document types) whose entries represent the nature of documents in the RD-EP in depth. The

[4] RD-EP Taxonomy of the Document types at the source: http://www.europarl.europa.eu/RegistreWeb/search/typedoc.htm.

investigation revealed that the taxonomy of the document types in the RD-EP has a total of 206 nodes. For more detailed results of the composition of this taxonomy see below:

- 46 grouping entities were not used for document relationships[5]
- 161 of the real document types were used for document relationships
    - 37 document types were related to more than 1 000 documents (they are used more often than the others)
    - 13 document types were never used
    - 11 document types were linked from the RD-EP (either to the register of MEPs, see the chapter 4.2.1.1.3.1 below, or to the registers of the other institutions)
    - 59 document types were not used since 2015 (from the 8th Parliamentary term)
    - 78 document types were used in 2015
    - Based on the investigation from the points above it is evident that:
        - Each Parliamentary term will bring new projects thereby creating new document types.
        - Some document types are used less frequently than in the previous Parliamentary terms.

The taxonomy of the Document types is very well organised.

The relationship between a document and entries in the taxonomy of Document types has a 1:1 cardinality. This means that each document must be of exactly one document type.

4.2.1.1.3. Metadata as relationships between documents and vocabularies

4.2.1.1.3.1. Document originator vocabularies

Document originator vocabularies are used to specify the origin and responsibility for the documents in the RD-EP. Two separate vocabularies serve this purpose:

1. Authority
2. Author

4.2.1.1.3.1.1. Vocabulary of Authorities

The vocabulary of 'Authorities'[6] is a list of organisational structures of the EP. Each entry potentially represents a creator of the document.

This vocabulary contains 235 authorities. As a result of the investigation of the authorities this lead us to the possibility of reorganising them into 23 groups. One authority may belong to more than one group[7] and this reveals a number of inconsistencies and duplications.

Document types taxonomy in the RD-EP in the study database: http://atom.ts-publicaccess.eu/form/tree?TreeId=100027 (look for I011 (RD-EP) Document types).

[5] Grouping entities in the RD-EP not used for relationships to documents extracted in the study database: http://atom.ts-publicaccess.eu/form/class?ClassId=100030.

[6] Vocabulary of Authorities in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1030315.

[7] 23 groups of Authorities in the study database: http://atom.ts-publicaccess.eu/form/group?GroupId=1033778.

The relationship between a document and entries in the vocabulary of Authorities has a 1:0...n cardinality. This means that the document can have zero to an unlimited number (n) of Authorities as originators.

4.2.1.1.3.1.2. Vocabulary of Authors

The vocabulary of 'Authors' is a list of individual MEPs.

The investigation revealed that the total number of MEPs for the 5th - 8th Parliamentary terms (for the period of 2001 – 2019 covered by the RD-EP) is 3 640. [8]

Member details are publicly available for the 751 MEPs who are active in the current 8th Parliamentary term (between 2015 and 2019), in the MEP list.[9]

The MEP list includes the outputs of activities from individual MEPs grouped by the specific types of these outputs. These types are not identical with the document types described in the previous chapter 4.2.1.1.2. If the output activity of a particular MEP leads to a result in the form of a document, it is also listed in the RD-EP.

The relationship between a document and entries in the vocabulary of Authors has 1:0…n cardinality. This means that the document can have zero or more MEPs in the role of originators.

This vocabulary of Authors (MEPs) is used again in the analysis of the document source Legislative Observatory.

4.2.1.1.3.2. Vocabulary of Topics

The vocabulary 'Topic' includes a specification of what the document is about.

The EuroVoc thesaurus[10] is used for this vocabulary. EuroVoc is tracked separately in the hierarchy of the study database[11] for analytical purposes as it would play an important role in any alternative of the future integrated access solution.

The RD-EP is not integrated with EuroVoc in other way than through the facet filter in the lists of results. Moreover, only the Top 20 EuroVoc nodes are included in the facet. The EuroVoc domains, micro thesauruses and concepts are combined into a single flat list within the facet filter. The EuroVoc indexing principles used in the RD-EP are different to the Indexing Policy used on EUR-Lex.

The relationship between the document and the EuroVoc entries has 1:0…n cardinality. This means that the document can be indexed by zero or more EuroVoc entries.

4.2.1.1.3.3. Time range vocabularies

Time specifications are handled in the RD-EP by 2 separate vocabularies:

1. Year
2. Parliamentary term

---

[8] Controlled vocabulary of Authors - members of the European Parliament since 2001 at the source (MEP list): http://www.europarl.europa.eu/meps/en/directory.html?filter=all&leg=0.

[9] Details on the 751 members of the European parliament in the 8th Parliamentary term in the study database: http://atom.ts-publicaccess.eu/?b=id1033760.

[10] EuroVoc official website: http://eurovoc.europa.eu.

[11] EuroVoc transformed into hierarchy in the study database: http://atom.ts-publicaccess.eu/form/tree?TreeId=100096.

4.2.1.1.3.3.1. Vocabulary of Years

The purpose of the vocabulary of 'Years'[12] is clear from first sight. It brings the basic orientation in time to the user searching for the document in the RD-EP.

Using a vocabulary of Years as an extract from the document dates was investigated. The method for extracting years from dates depends on document types. Occasionally, it is extracted from the document date, on other occasions from the date of the event.

The relationship between a document and entries in the vocabulary of Years has 1:1 cardinality. This means that each document must have exactly one entry in the vocabulary of Years associated.

4.2.1.1.3.3.2. Vocabulary of Parliamentary terms

The vocabulary of 'Parliamentary terms'[13] determines the EP's election period and from this point of view, it brings another useful kind of orientation in time while searching for the document.

The relationship between a document and entries in the vocabulary of Parliamentary terms has 1:1 cardinality. This means that each document must have exactly one associated entry in the vocabulary of Parliamentary terms.

4.2.1.1.3.4. Vocabulary of Languages

Another vocabulary used intensively in the RD-EP is the vocabulary of 'Languages'[14]. It is used to identify the language of the document and inform about the other language versions.

The documents are translated into the other languages based on their importance or on the importance of their types.

The relationship between a document and entries in the vocabulary of Languages has 1:1 cardinality. This means that each document must have exactly one entry in the vocabulary of Languages associated.

4.2.1.1.4. Metadata as document attributes

The following document attributes are mandatory for each document in the RD-EP:

- Reference
- Dates
  - Document date
  - Reception date
  - Publishing date
  - Event date
- Format

The following document attributes are not mandatory:

- Numbers/References/Codes
  - Procedure number (identification of the Legislative procedure)
  - COM reference (European Commission number)
  - SEC reference from the European Commission
  - Source reference

[12] Vocabulary of Years in the RD-EP in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1061224.

[13] Vocabulary of Parliamentary terms in the RD-EP in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1031010.

[14] Vocabulary of Document languages in the RD-EP in the study database: http://atom.ts-publicaccess.eu/form/class?ClassId=100057.

- PE number
- GEDA reference
- 'Feuille de route' number
- Meeting
- Dossier
- Document number during session
- Adopted text reference
- PE number of the amendment
- ISBN

Their usage is based on Document type but exact dependencies are not investigable from the end user interface of the RD-EP.

4.2.1.1.5. Relations between documents

4.2.1.1.5.1. Internal relations

The internal relations between the documents are established by 'their soft form'. This means the advanced search form is pre-filled by selected document attributes (mainly PE number, Cn reference, Procedure number, COM reference etc.), see 4.2.1.1.4, and the list of results then shows all of the documents with the same document attribute set. Obviously, the internal structures of documents include additional meta-information.

4.2.1.1.5.2. External relations

No external relations were found.

4.2.1.1.6. End-user search possibilities

This section analyses the search options that are available to the end-user in the RD-EP, and it covers three main areas in line with section 4.1, i.e. the search form, the list of results, and the document detail along with a brief conclusion.

4.2.1.1.6.1. Search form

By default, the search is carried out through a query word or phrase in the simple search form equipped with an autocomplete functionality.

A simple search form is expandable to an advanced search form, which offers additional filtering according to following criteria:

- Register reference
- Other references
- Author
- Document type
- Date from/Date to

The basics of the usage of this additional criteria are described in the help article of RD-EP.[15]
Usage of more than one criteria is handled as a chain linked by the logical AND operator, which means the more criteria used, the fewer results produced.

4.2.1.1.6.2. List of results

The list of results consists of links to documents in the RD-EP, where each document entry is characterized by the following metadata:

- Document title (serves also as a link to open the document detail)

[15] Help article 'How to search' in the RD-EP: http://www.europarl.europa.eu/RegData/registreweb/help/1040004-1_EN.pdf.

- Date of the document
- Type of the document
- Register reference

The list of results is further equipped with additional options for narrowing the volume of documents by filtering through the facet filters for whose vocabularies are specified in chapters 4.2.1.1.2 - 4.2.1.1.3.4. Facets used are gathered in a special box from which they can be easily removed.

The document titles are not composed in a unified form. Sometimes they are written in full wording whilst at other times they are shortened, coded and not very self-explanatory.

4.2.1.1.6.3. Document detail

After clicking on the document title in the list of results, the user receives following additional details about the document:

- Link to open document (typically in PDF format) in the desired language version
- Extract from the document (machine generated)
- Authorities responsible for the document
- Year and Parliamentary term in which document was created
- Document date
- Date of entry
- Date of event
- Other documents attributed (used in internal relations see 4.2.1.1.5.1)

The document title is not displayed in the document detail. This diminishes the user friendliness.

4.2.1.1.6.4. General evaluation of the search functionality

The search functionality works very well.

The Advanced search is probably primarily targeted to EP internal or very skilled users because the usage of, for example, Register reference and Other references criteria requires detailed knowledge of their meaning and structure. The general user would find it difficult choosing the optimal search criteria.

The usage of facet filters is intended for the end-users. However, there is a certain limitation that only approximately the Top 20 items of the documents in the list may be used in the facet filters. This limitation does not allow the full utilization of the facet filters.

4.2.1.1.7. Sample document

To see documents from different document sources through the same lens, a randomly selected document from RD-EP was re-created in the unified structure of the study database. This could help to design the possible future integrated access solution or help to understand the treatment of metadata from different document registers in one location. The Sample document from RD-EP is shown in Figure 2 and is directly accessible in the study database.[16]

[16] Sample document from the RD-EP in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1001028.

Sample document **European Parliament (Register of documents)**

BASIC INFORMATION

- URL: http://www.europarl.europa.eu/RegistreWeb/search/resultDetail.htm?reference=AGRI_AD(2015)552060&fragDocu=FULL?epbox
- Register

Target
I011 European Parliament - (Register of documents)

COMMON TYPES OF METADATA

- Date of the document: 23.7.2015

COMMON TYPES OF VOCABULARIES

- Year

Target
2015

- Originator

Target
COMMITTEE ON AGRICULTURE AND RURAL DEVELOPMENT

- Type(s)

Target
1.4.10 Opinions - Committee documents

- Language(s)

Target
Bulgarian
Croatian
Czech
Danish
Dutch
Estonian
Finnish
Greek
Italian
Latvian
Lithuanian
Maltese
Portuguese
Romanian
Slovenian
Spanish
Swedish

PECULIARITY IN THE EUROPEAN PARLIAMENT REGISTER OF DOCUMENTS

- Date of entry: 28.8.2015
- EP other references:
  Document's PE number : 552.060
  Procedures in which document is involved : 2014/0256(COD)
  Commission COM document reference : COM(2014)0557
  Feuille de Route (translation and revision of documents) : 1069086
  Reference to file to which document belongs : AGRI/8/01657
- EuroVoc

Target
(D10) EUROPEAN COMMUNITIES
(D20) TRADE
(D24) FINANCE
(D28) SOCIAL QUESTIONS

- Parliamentary term

Target
8 (2015 - 2019)

- Author(s)

Target
Fredrick FEDERLEY
Giulia MOI
Miguel VIEGAS
Molly SCOTT CATO
Nicola CAPUTO
Stanislav POLČÁK

Figure 2: Sample document from the European Parliament's Register of Documents

4.2.1.1.8. Re-use of the RD-EP in view of integrated access solution

The RD-EP exhaustively covers relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT (is the document about)	Document type; Topics
WHEN (was the document published)	Years; Dates
WHO (is responsible for the document)	Authorities
WHY (purpose of the document)	Procedure; Other metadata

The main disadvantage for any integrated access solution is the absence of machine readability solution (API or at least parametrized RSS).

The search engine used for the RD-EP is of a high quality with a fast response speed. Efforts to create an API interface (and thus delivery of information in machine-readable form) would be justified by the increase in re-use options for any future integrated access solution.

4.2.1.2. Legislative Observatory

4.2.1.2.1. General information

'The Legislative Observatory is the database set up in 1994 as a tool for monitoring the EU institutional decision-making process, with a particular focus on the European Parliament's role. The Legislative Observatory publishes comprehensive records in English and French, known as 'procedure files'. The website contains records for all ongoing and completed procedures since the beginning of the fourth parliamentary term in July 1994.'[17]

At the beginning, it must be emphasized that the Legislative Observatory is a dynamic database application providing the process information about legislative processes – procedure files. These procedure files cannot be regarded as documents because their content develops and changes in time or has the potential for this. Therefore, the Legislative Observatory cannot be regarded as a document source in the view of the study.

The procedure files link to documents contained in the RD-EP (see 4.2.1.1).

For these reasons, the **general analysis** of the Legislative Observatory was carried out, not the **thorough** one. This means that metadata will not be investigated in detail and no sample document is attached.

4.2.1.2.1.1. Public access

The Legislative Observatory is accessible at http://www.europarl.europa.eu/oeil/home/home.do.

The search is accessible at http://www.europarl.europa.eu/oeil/search/search.do?searchTab=y.

The whole database is accessible for free and without the need for registration. However, it is only provided in two official languages of the EU (English and French).

As the Legislative Observatory is an integral part of the website of the EP, it shares multiple vocabularies with the RD-EP.

4.2.1.2.1.2. Time range covered

The first document available through the search dates back to 1972 – but it refers to just one document. From 1972 to 1987 only 11 documents are provided. The number of documents increases from 1994 onwards.

4.2.1.2.1.3. Overall volume of procedure files

The Legislative Observatory provides access to more than 14 000 procedure files.

The average annual increment of documents from 1994 to 2015 is ca 500 – 700 procedure files. An unusual increase in the number of procedure files occurred in 2015 when 1 145 procedure files were added.

4.2.1.2.1.4. Brief investigation of the Legislative Observatory

Figure 3 shows the result of the taxonomic comparative analysis of the Legislative Observatory.[18]

[17] What is the Legislative Observatory? Accessible at: http://www.europarl.europa.eu/oeil/info/info2.do.

[18] Brief investigation of the Legislative Observatory in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1091977.

Figure 3: Overview investigation of the Legislative Observatory

4.2.1.2.2. Metadata

The Legislative Observatory disposes of a rich set of metadata in the form of vocabularies, attributes and relationships e.g.

- Parliamentary term
- Procedure type
- Procedure status
- Rapporteur

- Committee
- Political group
- Key events
- Forecasts
- Plenary activities
- Parliament documents
- Council configuration
- Commission DG
- Information documents
- Other institutions and bodies
- Subject
- Geographical area
- Legal Basis
- Type of legislative act
- Part-session alerts

The metadata sets above describe the procedure files in an exhaustive manner.

4.2.1.2.3. End-user search possibilities

This section analyses the search options which are available to the end-user of the Legislative Observatory. It covers the same three main areas as section 4.1, i.e. the search form, the list of results, and the document detail with a brief conclusion.

4.2.1.2.3.1. Search form

The search form allows searching based on keywords that can be limited in three ways:

- Search for specific keyword/phrases
- Search for any of the keywords
- Search for all of the keywords

These options can be easily selected from the pre-set list of search options.

Alternatively, it is possible to search based on a reference. These references can be chosen from a predefined list. The available options are:

- Procedure reference – including a choice of the year and the number of the legislative process.
- Committee dossier – including choice of the acronym characterizing the specific committee accompanied by the Parliamentary term and by its number. However, acronyms characterizing the committees are not listed in any vocabulary and if the end-user does not know them, this search method for documents is ineffective.
- Parliament document – the Parliamentary document type may be specified further. In this case, the search form provides a list of possible values and so the user does not need to detect the code from other sources. The user may select:
    - Draft report (PE)
    - Report (A)
    - Motion for resolution, oral question (B)
    - Position, resolution, decision (T).

After selecting a particular document type, the user can select the number of the parliamentary term, the year and the number of the document.

- Council document – enables searching with reference to a document originating from the Council of the European Union (year, number and version may be specified here).
- Commission document – enables a search by references to a European Commission document. This option is supplemented by a predefined list of the Commission's documents:
    - COM document (COM)
    - JOIN document (JOIN)
    - SEC document (SEC)
    - SWD document (SWD)
    - Other Commission document (C)
    - Document for European Councils (CSE – historic)

After selecting a particular document type, the user can select the year and number of the document.

- Documents of other institutions and bodies – enables a search by references to a document originating from the other EU institutions. The search form includes this list of institutions:
    - High Representative
    - European Central Bank: opinion, guideline
    - European Court of Auditors: opinion, report, special report
    - Economic and Social Committee: opinion
    - Committee of the Regions: opinion
    - European Ombudsman: report, special report
    - European Data Protection Supervisor: opinion

After the selection of a particular institution, the user can select the year and number of the document.

- Legislative act – enables a search by references to a particular act. The search form provides a clear, predefined list from which the user can simply select the type of act they require, namely:
    - Directive (L)
    - Regulation (R)
    - Decision (D)
    - Budget (B)
    - Agreement (A)
    - Justice and Home Affairs act (F)
    - Common Foreign and Security Policy act (E)
    - Parliament/Council recommendation (H)
    - Declaration (C)
    - Interinstitutional agreement (Q)
    - Non-binding act (X)
    - Third pillar act (Y)

After the selection of a particular type of act, the user can select the year and the number. It should be emphasized that these acronyms respect the format of descriptor labeling of document types, which are used for the creation of a Celex number. In this way the user actually creates a Celex number of the searched act.

- Official Journal – enables a search by the document classification to a particular series of the OJ; these series can be chosen from a predefined list. Therefore, the user does not need to identify what series of OJ they need before they search. The options are:
    - L Series – Legislation
    - C Series – Information and Notices

- C Electronic Series – Information and Notices

After the selection of a particular institution, the user can select the year and number of the document.

- Reasoned opinion – enables a search specific reasoned opinion, when its acronym (PE) is already pre-set in the search form and the user only completes the number.

All options can be easily combined using the 'Add reference line', which is not limited in any way.

4.2.1.2.3.2. List of results

For each retrieved document the list of results provides document name and the number of the legislative process. After clicking on the 'more information' option, it also provides information on the relevant Committee or the responsible rapporteur.

Each particular result is complemented by the option to create a PDF overview of the document (process), open it in the form of an XML in a new window, use the RSS function or create a permanent link to the given document (process).

The list of results can be further narrowed by using the facet filters displayed on the right side of the screen. There is a large number of filters available and some individual filters are multi-level. Therefore, they enable the user to specify the selection not only by basic filter name but also subordinate categories.

The following filters are available:

- Parliamentary term
- Year
- Procedure type
- Procedure status
- Rapporteur
- Committee
- Political group
- Key events
- Forecasts
- Plenary Activities
- Parliament documents
- Council documents
- Council configuration
- Commission DG
- Information documents
- Other institutions and bodies
- Subject
- Geographical area
- Legal basis
- Type of Legislative act
- Part-session alerts

4.2.1.2.3.3. Document detail

The detail of the retrieved document is clearly divided into several basic sections described further in the following subchapters.

4.2.1.2.3.3.1. Basic information

This section provides the information about the type of legislative process (and its number) and it also shows the state (legislative stage) where the document is situated at that moment. There is also the name of the document and its classification according to the Subject and Geographical area to which the document relates.

4.2.1.2.3.3.2. Key players

As the name itself indicates, this section provides an overview of the responsible institutions involved in the procedure. The information is not limited to a specific institution as a whole, but also provides the responsible bodies (e.g. Committee, Rapporteur at the EP level or the DG for the European Commission level or composition of the European Council, which discussed the document).

4.2.1.2.3.3.3. Key events

This section provides a timeline of the most important events in the specific legislative process (e.g. the Commission‘s legislative proposal, the debate in the Council of the European Union, the debate in the EP, and the Final act published in the OJ, etc.). Some key events are also accompanied by a Summary or a link to a document, which was created during at that event (e.g. the link to the Commission's legislative proposal).

4.2.1.2.3.3.4. Technical information

This section provides basic factual information on procedures in the following structure:

- Procedure reference
- Procedure type
- Procedure subtype
- Legislative instrument
- Legal Basis
- Mandatory consultation with other institutions
- Stage reached in procedure
- Committee dossier

4.2.1.2.3.3.5. Documentation gateway

This section provides access to all key documents within the relevant legislative process. The list of documents is divided by institutions. It is possible to view the PDF version of a particular document directly from the list of or the user can display its version on EUR-Lex (if it exists).

4.2.1.2.3.3.6. Links to other sites

There are direct links to the other databases, for example a direct link to the system of Inter Parliamentary EU Information Exchange or to EUR-Lex. These links do not lead to a general introductory page or to an empty search form, but directly to the relevant document or process.

4.2.1.2.3.3.7. Final act

If the legislative process is already closed and the act was published in OJ, then the section ‘Final act’ provides quick information about the issue of OJ in which the document is located, alongside a link to the summary.

4.2.1.2.3.3.8. Delegated and/or Implementing acts

If the basic act is followed by certain secondary legislative acts, such as delegated (or implementing) acts, then the details section shows another part, for example ‘Delegated Acts’ (or ‘Implementing Acts’) which clearly lists them.

4.2.1.2.3.4. General evaluation of the search functionality

The Legislative Observatory generally provides a clear and simple search form for the users. The search for documents is easy and intuitive. The user is navigated by the system through the successive steps to the final findings in almost all search criteria. The user doesn't even need to know all possible criteria (for example, the types of legislative acts) because the search form provides them in the form of pre-defined lists. Nevertheless, there are still certain criteria that do not have such a list and can be unhelpful for inexperienced users (such as Committee reference). The list of results is accompanied by a truly vast array of Facet filters, so the user has many options to narrow their choices. In the same way, the details of the results shown are very rich in information. This clearly breaks them up into several sections, which can be quickly navigated through the interface.

4.2.1.2.4. Re-use of the Legislative observatory in view of integrated access solution

The Legislative Observatory exhaustively covers relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT (is the document about)	Document type; Topics
WHEN (was the document published)	Years; Dates
WHO (is responsible for the document)	Authorities
WHY (purpose of the document)	Procedure; Other metadata

The Legislative Observatory is also equipped with a very good API which enables extensive machine readability of data.

However, the Legislative Observatory is more database than the document register. It includes information and documents from other sources (such as RD-EP), so re-use of this information could be regarded merely as a duplication of existing documents. From this point of view, re-use of the Legislative Observatory for the future integrated access solution is unnecessary.

4.2.1.3. InterParliamentary EU information eXchange

4.2.1.3.1. General information

The InterParliamentary EU information eXchange ('IPEX') is a platform for the mutual exchange of information between the national Parliaments and the European Parliament concerning issues related to the EU, particularly in light of the provisions of the Treaty of Lisbon.[19]

Similar to the Legislative Observatory, the IPEX is a dynamic database application providing process information about national parliament scrutiny.

Although the detailed information is called a document here, these procedure files cannot be regarded as documents because their content develops and changes in time or has the potential for this.

For this reason, the analysis of the Legislative Observatory was carried out in a **general** way.

4.2.1.3.1.1. Public access

IPEX is available at http://www.ipex.eu/. Documents are accessible through the database of documents and its search form at http://www.ipex.eu/IPEXL-WEB/search.do.

4.2.1.3.1.2. Time range covered

The oldest documents in the database date back to 1997. However, the regular publishing of documents began in 2001 (as the database did not include any information between 1998 and 1999).

4.2.1.3.1.3. Overall volume of published documents

The empty search form shows almost 74 000 documents in total. There are 49 documents from 1997. At the beginning (i.e. from the year 2000) the number of documents increased very slowly and irregularly (18 documents in 2000, 4 documents in 2001, 15 documents in 2002). Gradually, the number of documents added rapidly increased (297 documents in 2005, 7 234 documents in 2010 and 6 216 documents in 2015).

4.2.1.3.1.4. Brief investigation

Figure 4 shows the results of the taxonomic comparative analysis of the Legislative Observatory.[20]

[19] General information about the IPEX accessible at: http://www.ipex.eu/IPEXL-WEB/about/aboutIpexl.do.

[20] Brief investigation of the IPEX in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1091975.

Figure 4: Overview investigation of IPEX

4.2.1.3.2. Metadata

IPEX uses many different types of metadata, some of the vocabularies that are used are identical with the other existing document registers (RD-EP, Legislative observatory, Register of the Commission documents, EuroVoc, etc.).

4.2.1.3.2.1. Reference code

The vocabulary 'Reference code'[21] contains 17 items. The codes consist of two to six characters. However, it is difficult for the end-user to understand their meaning. Some of the acronyms include familiar ones such as COM, SEC, SWD indicating the specific Commission document types. Other

[21] IPEX vocabulary of the Reference codes in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1134753 .

acronyms, for example COD, NLE etc., indicate the type of the legislative process. The other codes are identical with those used by the European Parliament (INI, ACC etc.). However, the meaning of all the codes may not be determined by the end-user reliably.

The relationship between a document and entries in the vocabulary Reference code has 1:1 cardinality. This means that each document must have exactly one reference code.

4.2.1.3.2.2. NP threshold reached for card

The vocabulary 'NP threshold' reached for card contains only three items:

- None
- Yellow
- Orange

The relationship between a document and entries in the vocabulary labelled as 'NP threshold reached for card' has 1:0…1 cardinality. This means that each document can be attached to either none or one entries.

4.2.1.3.2.3. National parliament (institution)

The vocabulary 'National parliament (institution)'[22] includes the list of all parliaments of the Member States and of the candidate Member States of the EU. When the Parliament has more chambers, then the chambers are included in the vocabulary separately. The vocabulary contains 39 items.

The relationship between the document and the vocabulary National parliaments (institution) has 1:0…n cardinality. This means that no national parliament is obligatory for documents (it depends on the state of national scrutiny).

4.2.1.3.2.4. State of scrutiny

The vocabulary 'State of scrutiny' contains only 4 items:

- All types
- Not started
- In progress
- Complete

Its title and the list of items apparently indicates the ability to filter the search results based on the current status of the approval process in the national parliaments.

The relationship between a document and entries in the vocabulary of State of scrutiny has 1:1…n cardinality. This means that each document must belong to at least one entry in the vocabulary of State scrutiny.

4.2.1.3.2.5. Search type

In this situation the Search type represents one of the following options:

- Scrutiny
- Document
- Dossier

[22] IPEX vocabulary of the National Parliaments in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1134673.

4.2.1.3.2.6. Celex category

The vocabulary 'Celex category'[23] may be figuratively categorized as a subject-matter type of dictionary. It is not clear to which Celex category this dictionary is related. Although Celex numbers (used mostly in connection with EUR-Lex) are assigned on the basis of various document categories (descriptors), the items listed in the vocabulary Celex category used by IPEX do not correspond with these descriptors.

The facet filter provides only the 50 most frequently used types of Celex category. The most frequently used Celex category is 'Budget', used 302 times in total. The Celex category 'Company Law', used 64 times, closes the list. Other Celex categories are not visible to the end-user since this vocabulary cannot be used as a search criterion directly in the search form, but only as a facet filter to restrict the list of results.

An item from this vocabulary (specific Celex type) is also noted in the document detail.

The relationship between a document and entries in the vocabulary Celex category has 1:1…n cardinality. This means that each document must belong to at least one entry in the vocabulary Celex category.

4.2.1.3.2.7. EuroVoc

The vocabulary 'EuroVoc' at the IPEX website is used as facet filter and provides only an overview of the 50 most frequently used descriptors. All of them corresponds to the official version of the thesaurus EuroVoc. Thus it may be assumed that all documents are being indexed with the official EuroVoc descriptors.

The most frequently used term is 'Council of Europe countries', used 1 013 times. The term 'Private International Law' is used 241 times and closes the list. More descriptors could not be made visible from the point of view of the end-user at first sight since this vocabulary cannot be used as a search criterion directly in the IPEX search form. Therefore, the total number of descriptors used remains for the end-user undetected.

The relationship between a document and entries in the EuroVoc vocabulary has 1:1…n cardinality. This means that each document must belong to at least one entry in the EuroVoc vocabulary.

4.2.1.3.2.8. Other metadata

Other metadata used for the document description or to search for the relevant documents are data containing only options yes/no.

This data include:

- Reasoned opinion
- Political dialogue
- Veto
- Important information to Exchange

The first three of these options can be combined with the date designated as 'transmitted by national parliament'.

[23] IPEX vocabulary of the Celex categories in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1134828.

4.2.1.3.3. Relations between documents

The documents contained in the IPEX database are linked to EUR-Lex, where the user can choose a specific language version.

In addition, the documents are associated with the Legislative Observatory database. The link opens the legislative process related to views of national parliaments' scrutinies.

No internal links between the documents within the IPEX portal were observed.

4.2.1.3.4. End-user search possibilities

This section analyses the search options that are available to the end-user in the IPEX, and it covers three main areas in line with section 4.1, i.e. the search form, the list of results, and the document detail with the brief conclusion.

4.2.1.3.4.1. Search form

The search form is available only in the 'advanced' mode, which is switched on automatically after entering the 'Documents' page. The search form allows showing the limited search criteria even in the 'advanced' mode. The search is based on the keywords and it can be limited to a specific language. This is also the only option which remains available in the limited search criteria mode.

The search form in the 'advanced' mode includes the following search criteria:

- Reference
  - Code
  - Year
  - Number
- Subject deadline (from – to)
- NP threshold reached for card
- Scrutiny information
- Reasoned opinion (yes/no)
- Political dialogue (yes/no)
- Veto (yes/no)
- Important information to Exchange (yes/no)

4.2.1.3.4.2. List of results

The list of results is sorted by name and dossier number, which consists of a reference code, year and number. Each dossier title also shows the Commission document (i.e. proposal) on which the national parliaments provide their opinion.

The data are further complemented by the title of the reference document (i.e. any of the 'proposal' documents).

The list of results can be further filtered by facet filters available on the left side of the screen. The facets provide a wide range of filtering, namely by:

- Search type
- Reference code
- Reference year
- Celex category
- EuroVoc
- Institution
- NP threshold reached for card
- Important information

- Vetoed
- Reasoned opinion
- Political Dialogue

The facets thus cover almost all search criteria available in the search form. The only exception is the date. Therefore, the list of results cannot be limited using the facet filters by the date the national parliaments send the document, for example, or by the deadline date. However, both these options are available as search criteria in the search form.

4.2.1.3.4.3. Document detail

The detail of the retrieved document shows the reference number of the legislative proposal and its title. At the same time, there are documents in DOC and PDF format designated as 'other documents'. These are the documents on which the national parliaments form their opinions.

The next section of the document detail labelled as 'source information' includes a duplicated reference to the Commission document, on which the national parliaments express their opinion. Moreover, it includes the date of its adoption and submission date, author, legal basis as well as its classification according to the Celex and EuroVoc dictionaries and descriptors.

The document detail provides a direct link to EUR-Lex (respectively up to 24 links to all the language versions sorted according to the official EU languages, not just one general link to EUR-Lex).

Another part is labelled by the dossier name, which includes the title and number of the reference document of the Commission again.

The last section is called 'Scrutiny status'. Here all the member countries and their parliaments (or relevant chambers of the national parliaments) are listed. The same list is then also presented for candidate member countries.

Each country has clear graphic information which captures the state of consultations at the national level. Each state is represented by a different graphic, which gives a clear and comprehensive overview of the status. The possible states are:

- No information available
- Scrutiny in progress
- Scrutiny completed
- A reasoned opinion has been sent
- Political dialogue
- A national parliament makes known its opposition to bridging clauses
- There is important information to exchange
- Subsidiarity procedure is in progress (but this particular symbol has not been in use since 22 June 2012)

The document detail also provides the link to the Legislative Observatory.

4.2.1.3.4.4. General evaluation of the Search functionality

The search is clear and understandable. The search provides plenty of options for experienced users who know the subject from the national parliaments on the draft EU acts and how to search and display the results. The search form, list of results and the document details are clear and sufficient, with a variety of graphical indicators improving the clarity and overall user experience.

However, it is evident that the IPEX portal is designed specifically for advanced users because it does not provide definite, easily accessible assistance on specific search criteria (for example 'NP threshold

reached for card'), names of dossiers, used reference codes, and so on. Therefore, a lot of information may be difficult to access or even incomprehensible from the perspective of the common user.

4.2.1.3.5. Re-use of the IPEX for integrated access solution

The IPEX exhaustively covers relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT (is the document about)	Document type; Topics
WHEN (was the document published)	Years; Dates
WHO (is responsible for the document)	Authorities
WHY (purpose of the document)	Procedure; Other metadata

Unfortunately, the IPEX is not equipped with public API or any other machine readability possibility.

Moreover, it is more a database than a document register, where documents from other sources are duplicated, so the re-use of the IPEX for the future integrated access solution is unnecessary for most of the content. The IPEX still provides some interesting information (on national scrutinies), but the re-use from a technical point of view is hard or even impossible and further development needs to be done.

4.2.2. European Council, Council of the European Union

The European Council brings together the Heads of State or Government of the EU Member States. It makes decisions on broad political priorities and important projects, but does not wield legislative power. It is one of the EU institutions.

The Council of the European Union ('CEU'), generally known as the Council (previously the Council of Ministers), represents EU Member State governments. Together with the EP, the CEU adopts legislation proposed by the European Commission. It is also one of the EU institutions.

The following document sources on the website of the European Council and the CEU were analysed:

1. Public register of Council documents
2. Central Archives Search Engine of the Council of the European Union
3. Council database of agreements and conventions

4.2.2.1. Public register of Council documents

4.2.2.1.1. General information

The Public register of Council documents ('RD-CEU') contains references to documents produced or received by the General Secretariat of the CEU.

It is a rich and valuable document source so the **thorough analysis** was carried out.

4.2.2.1.1.1. Public access

The RD-CEU is accessible at this general URL address:
http://www.consilium.europa.eu/register/en/content/int/?lang=en&typ=ADV

Electronic access is free; no special authentication is needed. Approximately one third of the documents[24] are not automatically available and this unavailability is clearly stated. Nevertheless, the user can ask for such documents via 'Request a document' form. The user can create an account which is used for sending personalised information via e-mail.

The user can use the following predefined views:

- The most important documents from several selected document types
- The latest documents

It is not possible to use the list of all documents that would enable the user to work with the whole database of public documents.

4.2.2.1.1.2. Time range covered

The RD-CEU began operating in 1999, the first document dates back to 4 January 1999.

4.2.2.1.1.3. Overall volume of documents

The volume of documents until the end of 2015: ca 350 000. The yearly increment is ca 8 - 10%.

The overall volume of documents in all language versions is ca 2 500 000.

[24] CEU annual report on access to documents – 2014: http://www.consilium.europa.eu/en/documents-publications/publications/2015/council-annual-report-access-documents-2014/.

4.2.2.1.1.4. Brief Investigation of the RD-CEU

Figure 3 shows the result of comparative analysis of the RD-CEU[25].

Figure 5: Overview investigation of the RD-CEU

4.2.2.1.2. Document types

The document type is determined by the relationship between one specific item in the list of the document types and the document itself.

The vocabulary of 'Document types' consists of a list of 28 types[26].

[25] Brief investigation of the RD-CEU in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1033771 .

[26] RD-CEU Document types list at the source: http://www.consilium.europa.eu/register/en/content/out/?typ=DOCHEAD&i=ADV&RESULTSET=1. RD-CEU Document types list in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1030122.

4.2.2.1.3. Metadata as relationships between documents and vocabularies

4.2.2.1.3.1. Authority – Originator + Addressee

The document property 'Originator' in the RD-CEU represents the Authority which created the document. The value of the document property Originator is available only in the document profile. It is not possible to use it for searching. The list of values for the property Originator has not been published anywhere.

The property Originator is the type of vocabulary. However, this vocabulary is not published and there are no tools in the document registry to work with it. The relationship Originator ←→ Authority is not set for all the documents.

There is the same situation for the analogical property named Addressee, which indicates the target subjects for which the document is intended.

4.2.2.1.3.2. Subject matter

The uncontrolled vocabulary 'Subject matter'[27] serves for the description of the documents' topics.

It consists of a mixed list of 497 items covering various areas of operation of the CEU:

- Countries, regions, continents
- Organisations, members
- Relations between countries, organisations etc.
- Types of documents
- Several policies
- Important documents
- Economical activities, security problematics, citizen services etc.

There is either none, one or more Subject matters assigned to one document.

Working with the Subject matter from the end-user's perspective is difficult because after selecting the desired Subject matter only its code and not its name is inserted in the search form.

A vocabulary of the same name is also used in EUR-Lex, where it has 265 items. It is therefore a different vocabulary, and in fact only overlaps by ca 10% (for an exact match of an item's name).

4.2.2.1.3.3. Language

The vocabulary Language is used to identify the language in which the document is available through the relation between the document and the entries in the vocabulary Language. Thus, the vocabulary Language allows the search results to be limited through a facet filter to documents in the desired language alone.

4.2.2.1.4. Metadata as document attributes

4.2.2.1.4.1. Document Number

Specification of the internal Council number/code of the document.

4.2.2.1.4.2. Interinstitutional File

Identification of the Legislative procedure to which the document is attached is represented by the year, number and the code of the procedure type.[28]

[27] RD-CEU Subject matter list in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1031907.

[28] MDR website: Interinstitutional procedures vocabulary: http://publications.europa.eu/mdr/resource/authority/procedure/html/legislativeprocedures-eng.html.

4.2.2.1.4.3. Document Date

Indication of the date when the document was created.

4.2.2.1.4.4. Date of Meeting

Indication of the Date of Meeting, whose output was the document.

4.2.2.1.5. Relations between documents

No internal nor external relations were found.

4.2.2.1.6. End-user search possibilities

This section analyses the search options that are available to the end-user in the RD-CEU and it covers the same three main areas as the section 4.1, i.e. the search form, the list of results, and the document detail with a brief conclusion.

4.2.2.1.6.1. Search form

The advanced search is easily available from the Documents & Publications submenu. A simple search form does not exist.

The user may use the fields organised into the following groups:

- Document Number, Interinstitutional File, Words in subject, Words in text
- Selection of vocabularies Subject matter, Document type
- Date interval for the Document date
- Additional filter for Public documents only, Document language

The use of more than one criteria is handled as a chain combined with logical AND operator. That means that the more criteria are used, the fewer results are produced.

4.2.2.1.6.2. List of results

The List of results consists of links to documents in the RD-CEU, where each document entry is equipped with the following information/meta-information:

- Document number (Link to open the document detail)
- Document title
- Date of the document
- Link to open the PDF of the document (PDF icon in color is clickable while PDF icon refers to unavailability and offers link to 'Request a document' page)
- Available language versions of the document

4.2.2.1.6.3. Document detail

After clicking on the document title in the list of results the user receives the following additional details about the document:

- Document number
- Title
- Link for opening the PDF copy of the document
- Interinstitutional File
- Subject matter classification
- Document type
- Originator
- Addressee
- Document date

- Related meetings
- Document language

4.2.2.1.6.4. General evaluation of the Search tool

The search options are not described in detail. The list of results includes 500 results by default, which can be increased to 9 999 results by a little workaround but no more beyond that. Thus, the user does not know exactly how many results are available to his search queries.

Advanced functionality in the form of further filtering of results through facet filters or recommending the values that could be entered into the search fields (i.e. autocomplete functionality) is not implemented.

Similarly to the search function in the RD-CEU, the advanced search form is designed more for the CEU internal or skilled users than for the general public end-users. It is necessary to be well acquainted with the structure of the RD-CEU and its search possibilities in order to find the desired information.

4.2.2.1.7. Sample document

To see documents from different document sources through the same lens, a randomly selected document from RD-CEU was re-created in the unified structure of the study database. This could help to design the future integrated access solution or help to understand the treatment of metadata from different document registers at one location. The sample document from RD-CEU is shown in Figure 6 and is directly accessible in the study database.[29]

[29] Sample document from the RD-CEU in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1033437.


Sample document **Council of the EU (Public register)**

BASIC INFORMATION

Title	COMMISSION REGULATION (EU) .../... of XXX amending Annex II to Regulation (EC) No 1333/2008 of the European Parliament and of the Council as regards the title of the food category 12.3 Vinegars Decision not to oppose adoption
URL	http://www.consilium.europa.eu/register/en/content/out/?&typ=ENTRY&i=ADV&DOC_ID=ST-15474-2015-INIT
Register	Target: I021 Council - (Register of documents)

COMMON TYPES OF METADATA

Number of the document	ST 15474 2015 INIT
Date of the document	22.12.2015
Date of the meeting	13.1.2016 15.1.2016

COMMON TYPES OF VOCABULARIES

Year	Target: 2015
Originator	Target: General Secretariat of the Council
Topic(s)	Target: FOOD LEGISLATION; GENERAL AGRICULTURAL POLICY; HEALTH
Type(s)	Target: "I/A" ITEM NOTE
Language(s)	Target: English

PECULIARITY IN THE REGISTER OF DOCUMENTS OF THE COUNCIL OF THE EU

Addressee(s)	Target: Permanent Representatives Committee/Council

Figure 6: Sample document from the RD-CEU

4.2.2.1.8. Re-use of the RD-CEU in view of integrated access solution

The RD-CEU covers most of the relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT (is the document about)	Document type; Topics
WHEN (was the document published)	Years; Dates
WHO (is responsible for the document)	Not available
WHY (purpose of the document)	Procedure; Other metadata

The main disadvantage of RD-CEU re-usability in any integrated access solution is the absence of machine readability solution (API or at least parametrized RSS). Nevertheless, some kind of machine readability could be most easily established by sending machine-generated queries to the search engine in the URL (GET Method e.g. sequence of queries with incremental dates of the documents day-by-day) to retrieve the information.

4.2.2.2. Central Archives of the Council of the European Union

The purpose of the Central Archives Search Engine of the Council of the European Union document source ('CASE') is the provision of archive documents produced or received by the CEU in the exercise of its functions.

Documents are digitized from microfiche facsimile of the originals into PDFs. It is stated that the documents are only available in French, however, it was found that there are also documents in German.

The CASE is a very specific document source with limited possibilities and for this reason, a **general** analysis was carried out.

4.2.2.2.1. Public access and availability

Only files and documents from more than 30 years ago are available to the general public. More recent files and documents can be requested by the document request form.

The CASE is accessible at this general URL address:
http://www.consilium.europa.eu/en/documents-publications/archives/.

From the general search URL http://www.consilium.europa.eu/en/documents-publications/archives/search/, the lists of files and documents contained in CASE in PDF format and organised by inventories and years were retrieved. Each list of the document is several hundred pages long and contains basic metadata on each document.

These lists of documents are useful for localizing the documents needed by the search functionality which is then very limited.

4.2.2.2.2. Overall volume of published documents

The publicly available content in the CASE is composed of documents and files cataloged in 4 inventories and contains 26 606 files and documents in downloadable PDF format.

4.2.2.2.3. Time range covered

The CASE contains documents from the period 1953 - 1975.

4.2.2.2.4. Brief investigation of the CASE

Figure 7 shows the result of the comparative analysis of the CASE.[30]

[30] Brief investigation of the CASE in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1098986.

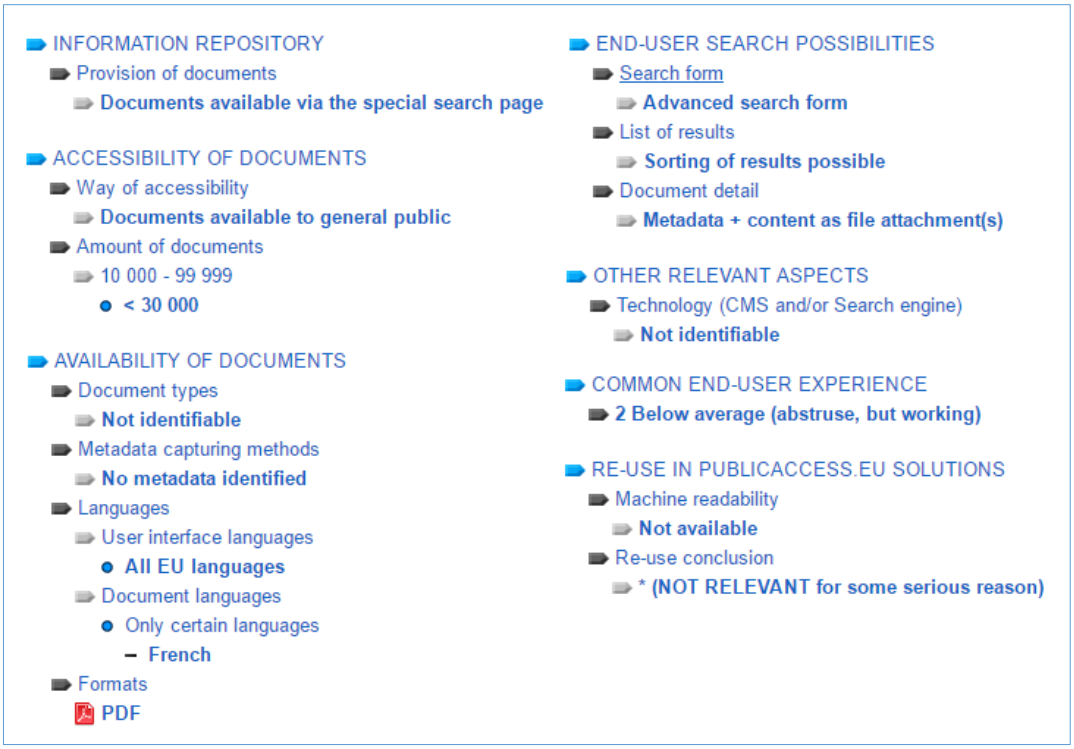


Figure 7: Overview investigation of the CASE

4.2.2.2.5. Re-use of the CASE in view of integrated access solution

The CASE does not cover any of relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT (is the document about)	Not available
WHEN (was the document published)	Not available
WHO (is responsible for the document)	Not available
WHY (purpose of the document)	Not available

As stated in the beginning of this analysis, the CASE is a very specific document source. From its content point of view, at least, it is very hard to imagine its use in any integrated access solution.

Moreover, there are no possibilities of machine readability.

The re-use of documents or information contained in the CASE is not possible without significant changes in the CASE, both from the technical and the content perspective.

4.2.2.3. Council database of agreements and conventions

Council database of agreements and conventions ('CEU-DAC') contains information on EU agreements and conventions. Besides the title of the document, it provides basic meta-information: date and place of signature, date of entry into force, ratification details and a link to the textual content of the document as published in the OJ.

In fact, this database is a very simple overview catalogue of documents, whose full content and metadata are already available in EUR-Lex. For this reason, the **general** analysis was carried out.

4.2.2.3.1. Public access

The CEU-DAC is accessible at this general URL address:
http://www.consilium.europa.eu/en/documents-publications/agreements-conventions/.

'Empty search' functionality is available, a list of all documents is available at this URL:
http://www.consilium.europa.eu/en/documents-publications/agreements-conventions/search-results/?dl=FR&title=&from=0&to=0.

Each entry in the list of documents could be expanded to the more detailed description of the agreement or convention with basic metadata. The table with an overview on basic accepting information is the most useful aspect of this document source.

4.2.2.3.2. Overall volume of published documents

The number of documents published in the CEU-DAC is ca 1 800.

4.2.2.3.3. Time range covered

The CEU-DAC contains documents signed since 1958.

4.2.2.3.4. Brief investigation of the Council database of agreements and conventions

Figure 8 shows the results of the comparative analysis of the CEU-DAC. [31]

Figure 8: Overview investigation of the Council database of agreements and conventions

[31] Brief investigation of the CEU-DAC in the study database:
http://atom.ts-publicaccess.eu/form/item?ItemId=1098988.

4.2.2.3.5. Re-use of the Council database of agreements and conventions in view of integrated access solution

The CEU-DAC only partially covers relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT (is the document about)	Not available
WHEN (was the document published)	Dates
WHO (is responsible for the document)	Authorities
WHY (purpose of the document)	Not available

Moreover, both the content and the metadata are fully provided via EUR-Lex. So the re-use of the CEU-DAC for the future integrated access solution is unnecessary, and its content should rather be obtained from EUR-Lex, where it is already technically implemented.

4.2.3. European Commission

The European Commission ('EC') is the EU's executive body and represents the interests of the EU as a whole. It proposes new EU legislation and ensures its correct application. It is one of the EU institutions.

The following document sources on the website of the EC were analysed:

1. Register of Commission documents
2. Comitology register
3. Register of Commission expert groups

4.2.3.1. Register of Commission documents

4.2.3.1.1. General information

The Register of Commission documents ('RD-EC') contains a number of documents with a focus on legislative documents created or prepared by the EC (COM, SEC, SWD, C, JOIN, OJ, PV).

A **thorough analysis** of RD-EC was carried out.

4.2.3.1.1.1. Public access

The RD-EC is accessible at this general URL address:
https://ec.europa.eu/transparency/regdoc/?fuseaction=search

A list of all documents is available after submitting an empty search form or at this URL address: https://ec.europa.eu/transparency/regdoc/?fuseaction=list&sortOrder=DESC

Electronic access is free and no special authentication is needed. Some of the documents are not immediately available and this unavailability is clearly stated. Nevertheless, the user can ask for such documents via the 'Request a document' form.

4.2.3.1.1.2. Time range covered

RD-EC began operating in 2001, the first document is dated 12 January 2001.

4.2.3.1.1.3. Overall volume of documents

The volume of documents at the end of January 2016:

- Final versions of the documents: ca 60 000
- All versions of the documents: ca 230 000

4.2.3.1.1.4. Brief investigation of the document register of the RD-EC

Figure 9 shows the result of comparative analysis of the RD-EC.[32]

[32] Brief investigation of the RD-EC in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000017.

Figure 9: Overview investigation of the RD-EC

4.2.3.1.2. Document types

The Document type is determined by the relationship between one specific item in the vocabulary of Document types and the document itself.

Two independent vocabularies are used for Document types.

4.2.3.1.2.1. Commission Reference types[33]

This classes the documents according to the phase of the process in which they were created and it contains the following items:

- COM - proposed legislation and other EC communications, legislative proposals, communications, reports, etc. to the CEU and/or the other institutions, and their preparatory papers
- C - documents relating to official instruments for which the EC has sole responsibility
- SWD - Commission staff working document

[33] RD-EC vocabulary of Commission reference types in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1033259.

- JOIN - EC and High Representative Joint Acts
- OJ - agendas of EC meetings
- PV - minutes of EC meetings
- SEC - documents which cannot be classified in any of the other series

Table 1 below shows the number of documents according to the EC reference type:

	Final version	All versions
C	40 207	155 613
COM	12 583	30 572
JOIN	158	353
OJ	151	1 935
PV	649	1 614
SEC	4 086	38 935
SWD	1 624	1 624
Total	59 458	230 646

Table 1: The number of documents according to the EC Reference type

4.2.3.1.2.2. Document types[34]

The vocabulary 'Document types' includes the following types of documents:

- Day note of delegation procedures
- Day note of empowerment procedures
- Day note of written procedures
- Delegated acts
- Legislative proposal
- Legislative proposal opened for feedback
- Reply to National Parliament

The interdependency of both vocabularies is shown in the following Table 2. It is evident that either:

- the Document types vocabulary includes only selected values or,
- not all documents are indexed based on this vocabulary

Because of this restriction, the vocabulary of Document types is not further considered in the analysis.

	All versions	Legislative proposal	Legislative proposal opened for feedback	Delegated acts	Reply to National Parliament	Day note of delegation procedures	Day note of written procedures	Day note of empowerment procedures
C	155 613	10	0	790	3 042	0	0	0
COM	30 572	5 523	22	0	0	0	0	0
JOIN	353	221	0	0	0	0	0	0
OJ	1 935	0	0	0	0	0	0	0
PV	1 614	0	0	0	0	0	0	0
SEC	38 935	75	0	0	0	1 283	1 351	1 247
SWD	1 624	0	0	0	0	0	0	0
	230 646	5 829	22	790	3 042	1 283	1 351	1 247

Table 2: The interdependency of Reference type and Document type vocabularies

[34] Vocabulary of Document types of the RD-EC in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1033290.

4.2.3.1.3. Metadata as relationships between documents and vocabularies

The RD-EC does not offer the ability to work with any thematic classification of documents (e.g. Topic or Subject). For this reason, such a vocabulary is not included in the description of vocabularies below.

The vocabularies available from the end-user point of view are described below.

4.2.3.1.3.1. Authority – Department responsible[35]

The Authority is represented by the relationship between the document and the item from the vocabulary 'Department responsible'. The vocabulary Department responsible is made up of a list of 61 items, which are:

- Analogous to the organisational structure of the EC – Directorates-General ('DGs')
- Include other EU organisational units
- Under EC responsibility

4.2.3.1.3.2. Year

The vocabulary 'Year'[36] is an extract of the document dates.

The vocabulary Year (in a form of the search form filter) allows the user to be in the right time span.

4.2.3.1.3.3. Language

The vocabulary 'Language' is used to identify the language in which the document is available. It does this by the relationship between the document and the entries in the vocabulary Language.

The vocabulary Language is used as an additional filter to search in the document titles in the advanced search form.

4.2.3.1.4. Metadata as document attributes

The following metadata is used as document attributes:

- Commission reference, e.g. PV(2016)2151/F1
- Document date, e.g. 28/01/2016
- Number, e.g. 2151
- Version, e.g. F1
- EUR-Lex reference, e.g. Celex:52015SC0706
- Dossier, e.g DL/2016/815; PH/2016/671; PH/2016/687
- Procedure, e.g. APP/2010/48; CNS/2016/10; CNS/2010/67

4.2.3.1.5. Relations between documents

Important document entries in the list of results are linked to EUR-Lex by external relations.

4.2.3.1.6. End-user search possibilities

This section analyses the search options that are available to the end-user in the RD-EC, and it covers three main areas in line with section 4.1, i.e. the search form, the list of results, and the document detail along with a brief conclusion.

The link to the RD-EC is not clearly visible on the main page of the EC website and the user who is not aware of its existence will hardly be able to find it.

---

[35] RD-EC vocabulary of Authorities – Department responsible in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1033313.

[36] RD-EC controlled vocabulary Year in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1061267.

4.2.3.1.6.1. Search form

The Advanced search is directly and clearly accessible from the left menu of the RD-EC. This is a classic approach to an advanced search, where the user can use several search criteria by filling the keywords into the search fields generally organised into the following groups:

1. Commission reference (Type, Year, Number)
2. Document date, Department responsible
3. Words in title (+ language filter)

Usage of more than one criterion is handled as a chain linked by the logical AND operator, which means that the more criteria are used, the fewer results are produced.

The full text search of documents, which is not interlinked with a primary Search form, is available in a separate tab. There are additional options to filter by date, language, and document format. In addition to the full text search in the RD-EC, it is possible to switch to full text search across the whole europa.eu domain.

4.2.3.1.6.2. List of results

After clicking the Search button without applying any filters, the user receives a list of all the documents that apply. This may be narrowed according to their requirements by repeating the search with some of the filters applied.

The List of results consists of links to documents in the RD-CEU, where each document entry is equipped with the following information/meta-information:

- Reference number
- Department responsible for drafting the document.
- Document title in the languages in which it is available
- Date of the document
- Accessibility of the document (link to the document or link to the document request form)

Monitoring through RSS channels can be set for each list of results. It can be also downloaded in XLS format (although only the first 1 000 documents in the results list).

4.2.3.1.6.3. Document detail

Beyond the information in the list of results the document detail in the form of summary information is not available. The document can be opened in the list of results directly, or it can be requested through the Request a document function.

4.2.3.1.6.4. General evaluation of the Search tool

The search options are not described in detail. However, this is not an issue since the functionality of the search is simple and intuitive for any common user.

The extended functionality in the form of additional filtering of results through the facet filters or suggesting values that can be entered into the search fields is not implemented. A disadvantage is that it is not possible to combine the full text search with a metadata search (of vocabularies, attributes, relations).

There is no API (which could be utilized to find structured information) visible or accessible from the end-user's point of view. The only possible way to find this is the use of RSS.

4.2.3.1.7. Sample document

To see documents from different document sources through the same lens, a randomly selected document from RD-EC was re-created in the unified structure of the study database. This could help to design the future integrated access solution or help to understand the treatment of metadata from different document registers at one location. The sample document from RD-EC is shown in Figure 10 and is directly accessible in the study database.[37]

[37] Sample document from the RD-EC in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1001028.

Sample document **European Commission (Register of Commission documents)**

BASIC INFORMATION

- Title: Proposal for a COUNCIL DECISION on the position to be taken on behalf of the European Union, in respect of the decisions to be adopted by Eurocontrol's Permanent Commission, on the roles and tasks of Eurocontrol and on centralised services
- URL: http://ec.europa.eu/transparency/regdoc/?fuseaction=list&n=10&adv=0&coteId=1&year=2015&number=805&version=F&dateFrom=&dateTo=&serviceId=&documentType=&title=&titleLanguage=&titleSearch=EXACT&sortBy=NUMBER&sortOrder=DESC
- Register

Target
I031 European Commission - Register of Commission documents

COMMON TYPES OF METADATA

- Number of the document: 805
- Date of the document: 23.11.2015

COMMON TYPES OF VOCABULARIES

- Year

Target
2015

- Originator

Target
DG Mobility and Transport

- Type(s)

Target
COM - Proposed legislation and other Commission communications to the Council and/or the otl

- Language(s)

Target
Bulgarian
Croatian
Czech
Danish
Dutch
English
Estonian
Finnish
French
German
Greek
Hungarian
Italian
Latvian
Lithuanian
Maltese
Polish
Portuguese
Romanian
Slovak
Slovenian
Spanish
Swedish

PECULIARITY IN THE REGISTER OF COMMISSION DOCUMENTS

- Final version (T/F): ✓ True

Figure 10: Sample document from the RD-EC

4.2.3.1.8. Re-use of the RD-EC in view of integrated access solution

The RD-EC covers most of the relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT **(is the document about)**	Document type Topics not available
WHEN **(was the document published)**	Years; Dates
WHO **(is responsible for the document)**	Authorities
WHY **(purpose of the document)**	Procedure; Other metadata

The internal structure (relationships to vocabularies) is very simple. The only major deficiency is the absence of any thematic information of the documents.

The main disadvantage of RD-EC re-usability in any integrated access solution is the absence of machine readability. However, customizable RSS channels could be used. Moreover, the regularly updated list of final Commission documents on the European Open Data Portal[38] ('ODP') could also be useful. It is also worth considering whether some of Commission documents (contained in the RD-EC) could not be more easily extracted from EUR-Lex for this project, as they are already processed in a way that ensures their re-use. However, since not all documents from RD-EC are subsequently published in EUR-Lex, the RD-EC would need to be improved from a technical viewpoint to provide full coverage of its documents.

[38] The list of Commission documents on the ODP: https://open-data.europa.eu/en/data/dataset/sg-regdoc.

4.2.3.2. Comitology register

4.2.3.2.1. General information

The Comitology register contains documents relating to the work of committees made up of representatives from EU countries. Their purpose is to assist the EC in implementing EU legislation. There is no overlap between the Comitology register and the RD-EC.

The Comitology register contains Documents organised in Dossiers.

A **thorough analysis** of the Comitology register was carried out.

4.2.3.2.1.1. Public access

The Comitology register can be accessed at this general URL address:
http://ec.europa.eu/transparency/regcomitology/index.cfm?do=Search.Search

The list of all documents is available after pressing the Search button.

Electronic access is free, and no special authentication is needed. Some of the documents are only available on direct request. The user can ask for such documents through a special 'Request a document' form.

4.2.3.2.1.2. Time range covered

The Comitology register began operating in 2008; the first document is dated 1 April 2008.

4.2.3.2.1.3. Overall volume

At the end of January 2016 the Comitology register contained:

- circa 48 000 documents
- circa 12 000 dossiers

The yearly increment is ca 7 000 documents.

4.2.3.2.1.4. Brief investigation of the document register of the Comitology register

Figure 11 shows the result of the comparative analysis of the Comitology register.[39]

[39] Brief Investigation of the Comitology register in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000021.

Figure 11: Overview investigation of the Comitology register

4.2.3.2.2. Dossier Type

Dossiers are further categorized as:

- Committee Meeting
- Written
- Unspecified

4.2.3.2.3. Types of documents

The Document type is held as a relationship between a document and its entry in the vocabulary of Document types.

Two independent vocabularies describing types of document are used.

4.2.3.2.3.1. Document types

The Vocabulary 'Document types'[40] includes the internal categorization of the Commission and it includes the following items:

- Agenda
- Draft implementing measure/act
- Other document
- Summary record
- Urgency letter
- Voting sheet

Table 3 below shows the number of documents based on the Commission reference:

Type	Total
Agenda	7 387
Draft implementing measure/act	19 374
Other document	2 569
Summary record	6 247
Urgency letter	508
Voting sheet	11 917
Total	48 002

Table 3: The number of documents based on the Commission reference

4.2.3.2.3.2. Procedures

The vocabulary 'Procedures'[41] represents the categorization of documents based on the EU Regulation concerning comitology.[42]

- Advisory (art. 3)
- Management (art. 4)
- Regulatory (art. 5)
- Regulatory with Scrutiny (art. 5a par. 1-5)
- Regulatory with Scrutiny (art. 5a par. 6 urgent)
- Safeguard (art. 6)
- Examination Procedure (art. 5, 6 and 7)
- Examination Procedure, Urgent (art. 5 and 8)
- Advisory Procedure (art. 4)
- Advisory Procedure, Urgent (art. 4 and 8)

The interdependency of both vocabularies is derived from the search, and it is shown in the following Table 4. During one exercise it was found that the search engine does not work consistently where, although there were many document types placed in Procedures, the search returned no results.

[40] Comitology register vocabulary of Document types in the study database http://atom.ts-publicaccess.eu/form/item?ItemId=1033883.

[41] Comitology register vocabulary of Procedures in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1033885.

[42] Regulation (EU) No 182/2011 on the EUR-Lex: http://eur-lex.europa.eu/eli/reg/2011/182/oj.

Type	art. 3	art. 4	art. 5	art. 5a p. 1-5	art. 5a par. 6	art. 6	art. 5, 6 and 7	art. 5 and 8	art. 4	art. 4 and 8
Agenda	-	-	-	-	-	-	-	-	-	-
Draft implementing measure/act	305	1 728	1 214	3 089	31	0	12 232	3	761	11
Other document	-	-	-	-	-	-	-	-	-	-
Summary record	-	-	-	-	-	-	-	-	-	-
Urgency letter	-	-	-	-	-	-	-	-	-	-
Voting sheet	-	-	-	-	-	-	-	-	-	-

Table 4: The number of documents based on the Procedure

4.2.3.2.4. Metadata as relationships between documents and vocabularies

The Comitology register does not have the ability to work with any thematic classification of documents (e.g. Topic or Subject). Therefore, such vocabulary is not gathered in the list of vocabularies below.

The vocabularies available for the end-user are described in the chapter below.

4.2.3.2.4.1. Authority – Department responsible

The vocabulary 'Department responsible'[43] is a list of 30 items, of which 26 are DGs and the remaining 4 are other organisational units of the EC.

Although all 30 items are also included in the vocabulary Department responsible in the RD-EC[44], it is a different vocabulary from the Comitology register system point of view.

Based on the end-user investigation, every document and every dossier has a relationship with just one item of this vocabulary Department responsible.

4.2.3.2.4.2. Authority – Committee

The vocabulary 'Committee'[45] is a list of 327 items which are currently functioning (respectively 451 items currently functioning and discontinued). It represents the specific purpose working committee which belongs to the one Department responsible (see 4.2.3.2.4.1 for the vocabulary Authority - Department responsible) and where the representatives of EU Member States participate.

Each document and each dossier are the product of the work of the same committee, which is captured in the Comitology register in the relationship between the document/dossier and the item from the vocabulary Committee.

4.2.3.2.4.3. Basic Act

The vocabulary 'Basic Act'[46] is a list of about 700 items: these are the acts that empower the EC to create a specific committee (of vocabulary Authority – Committee see 4.2.3.2.4.2 and for Department responsible see 4.2.3.1.3.1).

---

[43] Comitology register vocabulary of Department responsible in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1033919.

[44] RD-EC vocabulary of Department responsible in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1033313.

[45] Comitology register vocabulary of Committees in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1033931.

[46] Comitology register vocabulary of Basic Acts in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1033933.

Each committee has a relationship to one or more Basic Acts.

As the vocabulary contains many duplicates, it is an uncontrolled vocabulary and new entries in this vocabulary are introduced either at initiation or during activities of a particular committee.

4.2.3.2.4.4. Year

The vocabulary 'Year'[47] is an extract of the document dates.

The vocabulary Year (in the form of the search form filter) allows the user to find documents in the right timeframe.

4.2.3.2.5. Metadata as document attributes

Metadata varies depending on the type of document.

Common metadata used for all document types:

- Code
  - A unique code to identify the document
- Date
  - The date of the Committee meeting

Metadata specific for **Draft implementing act** document type:

- Co-decision
  - Whether the act passed by co-decision or not
- Status
  - The current status of the document within the procedure, as well as the date when the status was reached (and, if relevant, the full history of statuses and dates)

Metadata specific for **Voting sheet** document type:

- Procedure
  - The procedure used
- Opinion
  - The result of the vote
- Draft implementing act/measure
  - The act the vote refers to

4.2.3.2.6. Relations between documents

4.2.3.2.6.1. Internal relations

One internal relationship between documents was discovered: documents with document type 'Voting sheet' refer to relevant documents with document type 'Draft implementing measure/act'.

4.2.3.2.6.2. External relations

There is no relationship to the document registries outside the Comitology register nor to EUR-Lex. However, it could be useful at least for the items in the Basic Act vocabulary.

[47] Comitology register vocabulary of Years in the study database: http://atom.ts-publicaccess.eu/form/class?ClassId=100070.

4.2.3.2.7. End-user search possibilities

This section analyses the search options which are available to the end-user in the Comitology Register. It covers three main areas in line with section 4.1, i.e. the search form, the list of results, and the document detail with a brief conclusion.

The Comitology register is not clearly visible on the main website of the EC.

The Comitology register includes two search forms:

1. Search form for documents/dossiers
2. Search form for committees

The results produced by the committee search form do not include the documents of committees themselves (only the documents with the committees' Rules of Procedure are available). Thus, it is not analysed further.

4.2.3.2.7.1. Document/dossier advanced search form

The Advanced search is directly and clearly accessible from the menu on the left hand side of the Comitology register website.

This is a classic approach to an advanced search, where the user can use several search criteria by filling the keywords into the search fields organised into the following groups:

1. Dossier filters (Year, Number, Date, Type)
2. Authority filters (Service responsible, Committee, Basic Act)
3. Document filters (Number, Type, Procedure, Title, Language)

Usage of more than one criterion is handled as a chain linked with the logical AND operator, which means that the more criteria used, the fewer results produced.

Searching in the content of the documents is not available.

4.2.3.2.7.2. List of results

After clicking the Search button without applying any filters, the user will find a list of documents that may be narrowed according to their requirements by repeating the search with some of the filters applied.

The search results consist of the list of documents which fulfils the conditions of specified values.

The list of result consists of links to various registry entries, where each entry has the following information/meta-information:

- Code of the document and its superordinate dossier
- Document title in English
- Date of the document
- European Commission department responsible for the Committee and its support
- Link to the document (or documents in case of Dossier)

4.2.3.2.7.3. Document/Dossier detail

Document/Dossier detail can be shown by clicking the corresponding link.

The result is a dossier detail profile view, which contains the meta-information as an entry in the list of results.

The result is a detailed document profile view, which is structured as follows:

Common fields:

- Code: A unique code to identify the document
- Type of document (from the vocabulary of Document types, see 4.2.3.2.3.1)
- Committee (from the vocabulary Committees, see 4.2.3.2.4.2)
- Date of the Committee meeting
- DG (from the vocabulary Department responsible, see 4.2.3.2.4.1)

Metadata specific for Draft implementing act:

- Basic legal act (from the vocabulary Basic acts, see 4.2.3.2.4.3)
- Co-decision: Whether the act passed by co-decision or not
- Procedure (from the vocabulary Procedures, see 4.2.3.2.3.2)
- Link to relevant Draft implementing act/measure
- Status of the document within the procedure

Metadata specific for 'Voting sheet':

- Procedure (from the vocabulary Procedures, see 4.2.3.2.3.2)
- Opinion: The result of the vote
- Link to relevant Draft implementing act/measure

4.2.3.2.7.4. General evaluation of the Search tool

The search options are not described in detail. However, this is not an issue since the functionality of the search is simple and intuitive enough for any common user.

The extended functionality of additional filtering of results through facet filters or suggesting values that can be entered into the search fields is not implemented. In order to make the search more precise, the user must always return to the advanced search form.

The API, which could be utilized to find structured information, is not available.

4.2.3.2.8. Sample document

To see documents from different document sources through the same lens, a randomly selected document from the EC Comitology register was re-created in the unified structure of the study database. This could help to design the future integrated access solution or help to understand the treatment of metadata from different document registers at one location. The sample document from the EC Comitology register is shown in Figure 12 and is directly accessible in the study database.[48]

[48] Sample document from the Comitology register in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1044020.

Figure 12: Sample document from the Comitology register

4.2.3.2.9. Re-use of the Comitology register in view of integrated access solution

The Comitology register covers most of the relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT **(is the document about)**	Document type Topics (as a relationship between Act and Document)
WHEN **(was the document published)**	Years Dates
WHO **(is responsible for the document)**	Authorities
WHY **(purpose of the document)**	Procedure

But the unavailability of machine readability in the Comitology register represents a problem, and this needs to be resolved for any re-use options for the future integrated access solution.

4.2.3.3. Register of Commission expert groups

4.2.3.3.1. General information

Firstly, the Register of Commission expert groups maps the activities of consultative entities that help the EC with the preparation of legislative proposals, delegated acts and with the implementation of existing EU legislation, programs and policies.

Secondly, this register in some cases contains documents worked out by expert groups in different stages of their activities while in other cases it links to where these documents are published. However, there are also expert groups where activity documents were not found.

This, in fact, means that the register of Commission expert groups cannot be regarded as a document register in the usual sense of the term.

Nevertheless, a **thorough analysis** of the Register of Commission expert groups was carried out.

4.2.3.3.1.1. Public access

The Register of Commission expert groups is accessible at this general URL address:
http://ec.europa.eu/transparency/regexpert/index.cfm?do=search.search&searchType=advanced

All documents listing is available after pressing the Search button.

Electronic access is completely free, and no special authentication or login is needed.

4.2.3.3.1.2. Time range covered

The Register of Commission expert groups began operating in 2005. The first entries are dated 10 January 2005.

4.2.3.3.1.3. Overall volume

By the end of January 2016, the Register of Commission expert groups contained ca 6 500 documents directly published in this register.

4.2.3.3.1.4. Brief investigation of the document register of the Register of Commission expert groups

Figure 13 shows the result of a comparative analysis of the Register of Commission expert groups.[49]

[49] Brief investigation of the Register of Commission expert groups in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000019.

Figure 13: Overview investigation of the Register of Commission expert groups

4.2.3.3.2. Document types

The Register of Commission expert group does not use any vocabulary for the description of document types.

The types of documents may be derived from the description and titles of the document. The following types were found:

- Invitation
- Agenda
- Minutes
- Summary
- Presentation
- Study
- Analysis

4.2.3.3.3. Metadata as relationships between documents and vocabularies

As declared at the beginning of this chapter, the Register of Commission expert groups is an application for the description of activities of the expert groups rather than a document register. It uses many vocabularies. The following subchapters include their brief description because these vocabularies have an indirect relationship with the documents.

4.2.3.3.3.1. Authority – Expert groups

The vocabulary 'Committee' [50] is a list of 841 items – expert groups, of which 820 are of 'active' status, the remaining 21 being of 'on hold' status. One duplication of name was found (High-Level Forum for a Better Functioning Food Supply Chain) but both instances have a different code, one of them being of 'active' status while the other 'on hold'.

The titles and other information about the expert group's activities are in English or French without clear distinction or a filtering option.

Some of the expert groups have a descriptive name, the others have general names, e.g. 'Casting'.

Some of the expert groups include subgroups, which deal with a specified subset of activities.

The members of expert groups include individual experts, organisations and authorities of EU Member States (ministries).

4.2.3.3.3.2. Authority – Department responsible

The expert groups fall within the primary scope of one Department responsible and in some cases also to the secondary competence of the other associated Department responsible.

The vocabulary Department responsible[51] is a list of 32 items. It includes:

- 26 Commission DGs
- The remaining 6: other organisational units of EC where
  - 4 are also used for the Comitology register;
  - 2 are assigned only to the Register of Commission expert groups.

It is evident that this vocabulary differs from the vocabulary 'Department responsible' in the RD-EC[52].

4.2.3.3.3.3. Topic – Policy Areas

The vocabulary 'Policy Areas'[53] is a list of 59 items describing the areas of Expert group's responsibility through the relationship to the items in the vocabulary Expert groups. Typically, there is one relationship between the Expert group and the Policy Areas. However, in some cases, one Expert group has more Policy Areas.

The names of approximately 40% of the items begins with the word 'Other', and the Policy Area itself is stated in parentheses, which is not very practical from a usability point of view.

Similar vocabulary is found in the RD-EC, ca 20% of items are shared for both vocabularies.

4.2.3.3.3.4. Topic - Tasks

The vocabulary 'Tasks'[54] represents the tasks of the Expert groups.

It includes 64 items, where

---

[50] Register of Commission expert groups vocabulary of Expert groups in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1039425.

[51] Register of Commission expert groups vocabulary of Department responsible in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1041107.

[52] RD-EC vocabulary of Authorities – Department responsible in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1033313.

[53] Register of Commission expert groups vocabulary of Policy areas in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1041996.

[54] Register of Commission expert groups vocabulary of Tasks: http://atom.ts-publicaccess.eu/form/item?ItemId=1041997.

- 5 items represent the common definition of tasks for more than one Expert group;
- 59 items represent the definition of tasks for just one Expert group, where
    - 41 items begin with the word 'Other' and the task specification itself follows in the parentheses;
    - 18 items include tasks defined in the bulleted list.

As a result of this investigation, it is evident that the vocabulary Tasks is uncontrolled.

4.2.3.3.3.5. Expert group profile

The vocabulary 'Expert group profile'[55] includes five pieces of information characterizing the given Expert group, where each meta-information may have one of two predefined values.

This vocabulary is best explained through Figure 14.

Figure 14: Expert groups by their profile

One of the interesting findings is the fact that none of the Expert groups is of 'Closed' status. This is despite the last update of information on the 25 Expert groups being in 2011.

4.2.3.3.3.6. Authority – Members of expert groups

This vocabulary represents the complex list of members of the Expert groups. It includes detailed information about both institutional members and individual members, i.e. individuals.

The Register of Commission expert groups does not work with this information in any other way than by displaying the Expert group details.

4.2.3.3.4. Metadata as document attributes

4.2.3.3.4.1. Time range – Start of activity

Each Expert group has meta-information about the beginning of its activity in DD/MM/YYYY format.

4.2.3.3.4.2. Time range – Last update

Each Expert group has meta-information about the last update of information in the Register of Commission expert groups in DD/MM/YYYY format.

4.2.3.3.5. Relations between documents

Neither internal nor external relations between documents were investigated.

[55] Register of Commission expert groups vocabulary of Expert groups profile: http://atom.ts-publicaccess.eu/form/group?GroupId=1044004.

However, it is necessary to say once again that the Register of Commission expert groups is not a register of documents but it consists of the information about the expert groups. From this point of view, there are several links between the details of expert group descriptions displayed by the user interface. There are also links to various other websites in cases where the documents created or produced by the expert groups are not published in the Register of Commission expert groups but another place.

4.2.3.3.6. End-user search possibilities

This section analyses the search options that are available to the end-user in the Register of Commission expert groups. It covers three main areas in line with section 4.1, i.e. the search form, the list of results, and the document detail with a brief conclusion.

The Register of Commission expert groups is not explicitly visible on the main website of the EC.

The Register of Commission expert groups includes 2 search forms:

1. Quick search
2. Advanced search

Their functionality differs and so were analysed separately.

4.2.3.3.6.1. Quick search

4.2.3.3.6.1.1. Search form

There is a simple text field where the user can enter the query to search for:

- Complete information about the Expert group (using the 'LIKE' operator);
- The text of the attached documents (using full text search).

4.2.3.3.6.1.2. List of results

The list of results shows the code and the name of the Expert group and then all information from the Expert group item, which includes the searched word or expression.

The keywords used for the search are marked in bold in the retrieved results. If there are no bold words in the list of results, they are included in the content of the attached documents.

The list of results does not provide any additional filtering options such as facet filters.

4.2.3.3.6.1.3. Result detail

Clicking on the Expert group code opens the structured view of the Expert group record composed of the views on different groups of information included in the above-mentioned vocabularies, including structured information about the members.

4.2.3.3.6.1.4. General evaluation of the Quick search tool

The search options are not described in detail.

However, it is evident that the Register of Commission expert groups is an application targeted to a specific group of users, i.e. the members of the Expert groups.

The response time of the Register of Commission expert groups during the generation of the list of results was quite slow.

4.2.3.3.6.2. Advanced search

4.2.3.3.6.2.1. Search form

The Advanced search form is the structured search form, which enables the filtering of the search results according to all vocabularies except the vocabulary Authority – Members of Expert Groups.

4.2.3.3.6.2.2. List of results

The list of results shows the code and the name of the Expert group item.

The list of results does not provide any additional filtering options such as facet filters.

4.2.3.3.6.2.3. Result detail

Clicking on the Expert group code opens the structured view of the Expert group record composed of the views of different groups of information included in the above-mentioned vocabularies, including structured information about its members.

4.2.3.3.6.2.4. General evaluation of the Advanced search tool

The search options are not described in detail. But this is not a major issue because a moderately experienced user can use the advanced search functions fairly intuitively, although the appearance and functionality is considered quite 'outdated' from the point of view of current web design trends.

It is also evident that the Register of Commission expert groups is an application targeted to a specific group of users, i.e. the members of the Expert groups.

Unlike the Quick search, the Advanced search response time is fast.

4.2.3.3.7. Sample document

To see documents from different document sources through the same lens, a randomly selected document from Register of Commission expert groups was re-created in the unified structure of the study database. This could help to design the future integrated access solution or help to understand the treatment of metadata from different document registers in one location. The sample document from the Register of Commission expert groups is shown in Figure 15 and is directly accessible in the study database.[56]

As already mentioned in the beginning of this chapter, the Register of Commission expert groups is not a document register in the usual sense, because all the information is hidden in the vocabularies and metadata of authority that created it, for example in the *Advisory Committee for the Coordination of Social Security Systems (see Figure 15)*.

[56] Sample document from the Register of Commission expert groups in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1044018.

Figure 15: Sample document from the Register of Commission expert groups

4.2.3.3.8. Re-use of the Register of Commission expert groups in view of integrated access solution

The Register of Commission expert groups exhaustively covers relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT (is the document about)	Document type; Topics
WHEN (was the document published)	Years; Dates
WHO (is responsible for the document)	Authorities
WHY (purpose of the document)	Other metadata

The Register of Commission expert groups provides an excellent reusability for any integrated access solution – it provides a dump, updated daily, of the complete database in XML format with links to the documents.

However, it is necessary to reiterate that the Register of Commission expert groups is not a document register in the true sense of the word because its main purpose is to map the activities of the expert groups. Only the minority of working groups uses this register to publish their activities.

The Register of Commission expert group could be regarded more as a process-related database than as a document register. The need for these documents to be accessible in the future integrated access solution is questionable and needs to be decided at the management level, even though technically, the re-use of these documents is feasible.

4.2.4. Court of Justice of the European Union

The Court of Justice of the European Union ('CJEU'), established in 1952, interprets EU law and ensures it is applied uniformly in all the Member States. It also settles legal disputes between EU governments, individuals, companies or organisations and EU institutions. It is one of the EU institutions.

The general website of the CJEU is accessible at the link http://curia.europa.eu. The core information system of the website is called **InfoCuria** (see 4.2.4.1.). It provides quick and easy access to the Case Law produced by the CJEU (with relevant documents arranged according to their pertinence to a particular case).

However, several other documents and even registers can be accessed via the website. The main goal of the study with the CJEU is an in-depth analysis of the InfoCuria document register, but other registers/website sections were also analysed, such as the:

- Judicial calendar
- Access to administrative documents
- Press releases
- Various documents

4.2.4.1. InfoCuria

4.2.4.1.1. General information

InfoCuria is the most important registry provided by the CJEU. Its importance is underlined by the fact that the simple search form is displayed as the main element on the homepage of the CJEU's website and the 'Search for a case' option is the first thing users can see after choosing their language version of the portal.

A **thorough analysis** of the InfoCuria was carried out.

4.2.4.1.1.1. Public access

The simple search form is available directly from the homepage of the CJEU at http://curia.europa.eu. The advanced search form is easily accessible via the button  or directly accessible at http://curia.europa.eu/juris/.

The whole register of InfoCuria is accessible for free, and no special authentication is needed.

4.2.4.1.1.2. Time range covered

InfoCuria provides access to documents grouped by cases. Particular documents are searchable as well, but the premise that each document belongs to one particular case is retained.

InfoCuria provides access to cases back to 1953 (where 2 cases are available). From 1953 onwards, the number of available cases consistently rises.

4.2.4.1.1.3. Overall volume of documents

InfoCuria provides access to 34 689 cases. Each case consists of several documents (application, order, judgment, etc.). At the moment, InfoCuria provides access to 87 997 documents.[57]

[57] Information regarding number of cases as well as the number of documents is relevant as of 15 February 2016.

The number of cases grew very slowly in the first years of the CJEU's existence, but it has been rapidly increasing in the last two decades. There are 2 cases available from 1953, 5 cases from 1954, 9 cases from 1955 and 68 cases from 1960, but 418 cases from 1990 and 1671 cases from 2015.

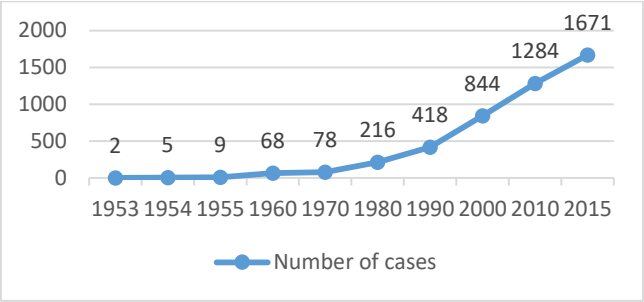


Figure 16: Annual increase of cases in InfoCuria

4.2.4.1.1.4. Brief investigation

Figure 17 shows the result of the comparative analysis of the InfoCuria[58].

- INFORMATION REPOSITORY
  - Provision of documents
    - **Documents available via the special search page**
  - Complementary features/applications
    - **More special databases**
- ACCESSIBILITY OF DOCUMENTS
  - Way of accessibility
    - **Documents available to general public**
  - Machine readability methods
    - **RSS**
  - Amount of documents
    - 10 000 - 99 999
      - **< 100 000**
- AVAILABILITY OF DOCUMENTS
  - Document types
    - More document types available
      - **< 30**
    - **Hierarchical organisation**
  - Metadata capturing methods
    - Vocabulary groups
      - **Document types**
      - **Authorities**
      - **Topic**
      - **Years (or other periods)**
      - **Languages**
      - **Process information**
    - Attribute groups
      - **Dates**
      - **Various codes**
      - **Various numberings**
      - **Other (e.g. descriptions, formats, etc.)**
    - Relations
      - Internal relations
        - **Related content within the website**
      - External relations (between registers)
        - **EUR-Lex**
  - Languages
    - User interface languages
      - **All EU languages**
    - Document languages
      - **Only certain languages**
- Formats
  - **PDF**
  - **webpage**
- END-USER SEARCH POSSIBILITIES
  - **Empty search possible**
  - **Document content indexed**
  - Search form
    - **Simple search form**
    - **Advanced search form**
  - List of results
    - Facet filtering of the results available
      - **by document type**
      - **by authority**
      - **by language**
    - Sorting of results possible
      - **by date**
      - **by number/code**
  - Document detail
    - **Content + metadata in a (set of) webpage(s)**
- OTHER RELEVANT ASPECTS
  - Technology (CMS and/or Search engine)
    - **Not identifiable**
- COMMON END-USER EXPERIENCE
  - **4 Above average (close to perfection)**
- RE-USE IN PUBLICACCESS.EU SOLUTIONS
  - Vocabularies groups
    - **Document types**
    - **Originators**
    - **Topics**
    - **Years**
  - Attribute groups
    - **Dates**
    - **Process stage description**
  - Machine readability
    - **Not available**
  - Re-use conclusion
    - ****** (API or RSS: NO, Metadata set: FULL)**

Figure 17: Overview investigation of the InfoCuria

[58] Brief investigation of InfoCuria in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000023.

4.2.4.1.2. Title composition

The titles of the cases vary depending on the state of the proceeding.

The title of a case pending (case in progress) is based on the case number followed by the parties to the dispute. The case number is constructed from the acronym of the author (C for Court of Justice, T for General Court, F for Civil Service Tribunal) followed by the serial number and year. An example of such a title: C-128/16 P - Commission v Spain.

The title of a case closed is much more complex and is based on the type of the final decision (judgment, order, …), the composition of the court, date of delivery, names of the parties to the dispute, type of procedure and case number. An example of such a title:

Judgment of the Court (Grand Chamber) of 15 February 2016.
J. N. v Staatssecretaris van Veiligheid en Justitie.
Request for a preliminary ruling from the Raad van State.
Case C-601/15 PPU

4.2.4.1.3. InfoCuria Document types

Documents available via InfoCuria are grouped into 3 main levels:

- Documents published in the European Court of Reports
  - Eight different types of documents belong to this group (e. g. judgments, orders, opinions, decisions, etc.[59])
  - As it can clearly be seen from the name of the group, these documents were published or will be published in the European Court Reports (or European Court Reports – Staff Cases).
  - Documents published in the European Court Reports are available in all the official languages of the EU on the day of their delivery.
- Documents not published in the European Court Reports
  - Three types of documents belong to this group – judgments, orders and decisions (review procedure). However, document type 'order' needs to be further specified by the sublevel of nodes. There are 30 different types of 'orders' searchable using InfoCuria (e. g. Legal Aid, Declinature, Intervention, Substitution of parties, etc.[60]
  - These documents were delivered or made since 1 May 2004 and not published in the European Court Reports.
  - Full text search is available, however, only in the language of the case and the language of deliberation.
- Notices published in the OJ
  - These documents are notices published in the OJ since 1 January 2002; the notices regard new cases or closed cases[61] (notice regarding the judgment, notice regarding the order, etc.)

[59] Documents published in ECR in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1044041.
[60] Documents not published in ECR in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1044043.
[61] InfoCuria notices published in OJ in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1044045.

- As these documents are published in the OJ, they are also available via EUR-Lex (see more on EUR-Lex: Types of documents in EUR-Lex[62], sector 6 - descriptors CN, CA, CB, CU, CG, TN, TA, TB, FN, FA, FB or more information regarding sector in EUR-Lex in the subchapter 4.2.10.2.1).

While searching for documents, the user has the option to look for all types of documents as well as for one or more selected types of document. A list of all documents is also available so that the preferred document may be easily selected from that list.

It is also possible to filter the search in the view of time (a defined period from – to, the last 8 days, last month, last year, last 5 years or specific date).

4.2.4.1.4. Category of references

The cases (and documents covered by cases) could be connected to specific acts of the EU, law or other case law. For these reasons, InfoCuria allows searching for a particular reference, where the most frequent references are available in the predefined list of values. This vocabulary is based on different categories of references, but not all of the available categories are covered by the vocabulary.

The vocabulary Category of references[63] covers references to the:

- Treaty
- Regulation
- Case law
- Directive
- Decision
- Other

The first three are further branched into sublevels while the last three are contained in the vocabulary as top levels only.

The vocabulary Category of references then covers the most frequently used or the most important values from Treaties, Regulations and Case law.

Concerning Treaties, the vocabulary contains 23 entries (such as TFEU (Lisbon), etc.). The search according to this vocabulary can be extended by the option to search by subdivisions: article/paragraph/subparagraph/letter.

As for Regulations, the vocabulary contains 10 entries (such as Staff Regulations of Officials, Rules of procedure, etc.). Here as well is the possibility to search according to subdivisions (annex/article/paragraph, subparagraph, letter).

Concerning case law, the vocabulary is further structured according to court (Court of Justice, General Court, Civil Service Tribunal); moreover, every court has its own types of case law as predefined values (e. g. judgment, order, decision, etc.). The search can be further specified by the number of the order and year.

Directives, Decisions and 'Other' options are not further specified, but a general search is possible:

---

[62] Types of documents and corresponding sectors on EUR-Lex: http://eur-lex.europa.eu/content/tools/TableOfSectors/types_of_documents_in_eurlex.html.

[63] InfoCuria vocabulary of Category of references in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1051963.

- Directive
    - Available to search according to the number/year and subdivisions/annex/article/paragraph/subparagraph/letter.
- Decision
    - Available to search according to the number/year and subdivisions annex/article/paragraph/subparagraph/letter.
- Other
    - This option is based on Celex numbers, where one or more Celex numbers can be entered into the search form as free text.

4.2.4.1.5. Metadata as relationships between documents and vocabularies

4.2.4.1.5.1. Document originator vocabularies

Document originator vocabularies are used in the InfoCuria to specify the origin and responsibility for the documents. Three separate vocabularies serve this purpose:

1. Formation of Court
2. Judge-Rapporteur
3. Advocate general
4. Country (as a source of a question referred for a preliminary ruling)

4.2.4.1.5.1.1. Authority - Formation of Court

The CJEU while deciding individual cases may sit as a full court, in a Grand Chamber of 15 Judges or in a Chamber of three or five Judges. The vocabulary 'Formation of Court'[64] available in the InfoCuria represents these different formations. The vocabulary is divided into subsections in view of the 3 courts (Court of Justice, General Court, Civil Service Tribunal); each of these subsections has predefined values corresponding to a particular court.

There are 17 different formations available for The Court of Justice (e. g. The Full Court, The President, The Grand Chamber, The First Chamber, The Second Chamber, etc.), 25 different formations for The General Court (e. g. The Full Court, The President, The Single Judge, The Grand Chamber, etc.) and 7 different formations for The Civil Service Tribunal (e. g. The Full Court, The President, The Single Judge, The First Chamber, etc.).

Overall, the user has the option to choose from 49 different formations of the CJEU. However, the information regarding the Formation of the Court is available in the InfoCuria only after the case is closed.

While searching for a particular formation, one or more formations can be selected, formations can only be selected from a predefined list available in the search and no free text option is available here.

The relationship between the case and the entry in the vocabulary Formation of Court has 1:1 cardinality. This means that the case must be of exactly one entry from this vocabulary.

4.2.4.1.5.1.2. Person – Judge-Rapporteur

The Judge-Rapporteur is the Judge in charge of drafting the judgment. The vocabulary 'Judge-Rapporteur'[65] in the InfoCuria consists of 146 names of Judges Rapporteurs. When searching for a

[64] InfoCuria vocabulary of Formation of Court in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1051397.

[65] InfoCuria vocabulary of Judge-Rapporteur in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1051494.

particular Judge-Rapporteur, one or more names can be selected; names can only be selected from a predefined list available in the search, no free text option is available here.

The Judge-Rapporteur is designated by the President of the CJEU soon after the document's initiating proceedings have been lodged. However, the information regarding Judge-Rapporteur is available in the InfoCuria only after the case is closed.

Information about the Judge is then presented in the section 'Procedural Analysis Information' of the 'Case information'.

The relationship between a case and the entry in the vocabulary of Judge-Rapporteur has 1:0…1 cardinality. This means that a case can be attached to either none or one Judge-Rapporteur.

4.2.4.1.5.1.3. Person – Advocate General

The Advocates General are appointed for six years, and their tasks are to assist the Court and present legal opinions in cases assigned to them. The CJEU is composed of 11 Advocates General. The vocabulary 'Advocate General'[66] in the InfoCuria which consists of names of 51 Advocates General.

While searching for a particular Advocate General, one or more names can be selected. The names can only be selected from a predefined list available in the search; no free text option is available here.

The information regarding the Advocate General is available in the InfoCuria register of the document only after the case is closed and presented in the section 'Procedural Analysis Information' of the 'Case information'.

The relationship between a case and the entry in the vocabulary of Advocate-General has 1:0…1 cardinality. This means that a case can be attached to either none or one Advocate-General.

4.2.4.1.5.1.4. Country (as the source of a question referring to a preliminary ruling)

One type of proceeding before the CJEU is the reference for preliminary rulings, where the national courts may (and in some situations must) refer to the CJEU and ask them to clarify a point concerning the interpretation of EU law. The country of the court's origin is searchable using InfoCuria.

The vocabulary 'Country' is only used as the source of a question referring to a preliminary ruling[67]. It contains 29 entries – 28 current Members States of the EU and 1 entry for 'Benelux'. However, searching for 'Benelux' returns no results, so this special entry in the search form is questionable.

The relationship between a case and the entry in the vocabulary of Country has 1:0…1 cardinality. This means that the case can be attached to either none or one Country.

4.2.4.1.5.2. Topic vocabularies

The InfoCuria uses 2 different types of vocabularies for the classification of document subject:

1. Systematic Classification Scheme
2. Subject matter

[66] InfoCuria vocabulary of Advocate General in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1051821.

[67] InfoCuria vocabulary of Country (as the source of a question) in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1051907.

4.2.4.1.5.2.1. Systematic Classification Scheme

The vocabulary 'Systematic Classification Scheme' is based on the classification scheme of the Digest of case law[68], which is a systematic collection of the summaries of judgments and orders published in the European Court Reports and the European Court Reports – Staff Cases. Each paragraph is represented by one or more classification codes. The vocabulary (as well as the Digest itself) is divided into 2 main parts as a result of the Treaty of Lisbon entering into force:

- Classification scheme after the Treaty of Lisbon (applied to case law since 2010);[69]
- Classification scheme before the Treaty of Lisbon (applied to case law from 1954 to 2009.)[70]

The Systematic Classification Scheme is constructed as an expansive group tree with up to 5 sublevels. Overall, the scheme after the Treaty of Lisbon consists of 1 048 codes and the scheme before the Treaty of Lisbon consists of 1 165 codes. High number of codes gives this vocabulary a nuanced and exhaustive coverage of topics.

While searching in InfoCuria, the predefined list of all codes is available in the tree diagram structure. Moreover, when activating/deactivating the option 'Include earlier/new scheme' the search engine can return codes from one classification scheme and from both classification schemes (before/after Lisbon) along with the corresponding codes from both schemes (which are to be automatically searched.)

The relationship between a case and the entry in the Systematic Classification Scheme vocabulary has 1:1…n cardinality. This means that each document must belong to at least one entry in the vocabulary.

4.2.4.1.5.2.2. Subject matter

The vocabulary 'Subject matter'[71] is based on the subdivisions of the Treaties; it corresponds to the legal basis covered by the application or request at hand. After the case is closed, the code corresponds to the legal basis of the judgement (or the order, decision or opinion).

The vocabulary contains 182 codes and is quite general. It is specific for a few areas of EU activities, for instance in the areas of agriculture and customs duties.

This vocabulary is also used for the classification of documents on EUR-Lex. However, the vocabulary of Subject matters on EUR-Lex contains 265 codes, so it is an enhanced version of the vocabulary used in InfoCuria.

The relationship between a case and the entry in the Subject matter vocabulary has 1:1…n cardinality. This means that each document must belong to at least one entry in the vocabulary.

4.2.4.1.5.3. Additional vocabularies

Additional vocabularies used in InfoCuria:

---

[68] The Digest of case law is accessible as separate online register (without search possibilities). However, only French version of the Digest is available (Répertoire de Jurisprudence) at http://curia.europa.eu/common/recdoc/repertoire_jurisp/bull_home/index.htm).

[69] InfoCuria Classification scheme after the Treaty of Lisbon in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1049182.

[70] InfoCuria Classification scheme before the Treaty of Lisbon in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1049183.

[71] InfoCuria vocabulary of Subject matter in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1044303.

1. Procedure and result
2. Case status
3. Language

4.2.4.1.5.3.1. Procedure and result

The vocabulary 'Procedure and result'[72] is divided into 2 main parts:

- **Type of procedure** with 27 entries (e. g. Opinion procedure, Appeals, Review, Arbitration clause, etc.) and
- **Type of result** with 11 entries (e. g. adjourned, interlocutory judgment, application granted, dismissal on substantive grounds, etc.).

4.2.4.1.5.3.2. Case status

The 'Case status'[73] vocabulary contains just 2 entries:

- **Cases closed** – which covers only closed cases.
- **Cases pending** – which covers on-going cases, stayed cases and cases in the process of being discontinued or withdrawn).

While searching, also the possibility to search for both ('all') options is available.

4.2.4.1.5.3.3. Language

The vocabulary 'Language'[74] contains the 24 official languages of the EU and when searching with it, it could apply to the language of the case or the language of the Opinion.

4.2.4.1.6. Metadata as document attributes

- Case number
- Name of the parties
- European Case Law Identifier ('ECLI')
- Date of delivery
- Date of the Opinion
- Date of the hearing
- Date of the lodging of the application initiating proceedings

4.2.4.1.7. Relations between documents

4.2.4.1.7.1. Internal relations

While looking from the point of view of cases, InfoCuria provides internal links within the system to relevant documents from previous stages of the decision process. For example, if the procedure is already closed and the final judgment is available, InfoCuria provides links to the application, the opinion of the Advocate-General and the final judgment.

4.2.4.1.7.2. External relations

Documents contained in the InfoCuria are interlinked with EUR-Lex. There are two types of documents that link to EUR-Lex.

---

[72] InfoCuria vocabulary of Procedure and result in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1044668.

[73] InfoCuria vocabulary of Case status in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1051984.

[74] InfoCuria vocabulary of Languages in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1051938.

Firstly, if the document was published in the OJ, a link to EUR-Lex is provided. InfoCuria provides a list of all the languages in which the EUR-Lex document is available, so the user can open their preferred language version directly from the InfoCuria.

Secondly, there are links to the legislation cited in the proceeding. For example, if the proceeding deals with the CHARTER, InfoCuria provides a direct link to the text of the CHARTER in EUR-Lex.

4.2.4.1.8. End-user search possibilities

This section analyses the search options that are available to the end-user in the InfoCuria, and it covers three main areas in line with section 4.1, i.e. the search form, the list of results, and the document detail with a brief conclusion.

4.2.4.1.8.1. Search form

4.2.4.1.8.1.1. Simple search form

The simple search form is accessible directly from the homepage of the CJEU's website and is its dominant feature.

The user has the option to choose the Case Law according to the three main courts constituting the CJEU (Court of Justice, General Court, Civil Service Tribunal).

Another option is to enter the case number. A predefined template for the number is displayed as an aid to help the user.

The user may search for the case using the text field dedicated to the names of the parties or limit their search by using a time range (from – to).

As a supplement to the simple search form, there is an option to access the most recent judgements and opinions sorted according to the 3 courts.

4.2.4.1.8.1.2. Advanced search form

The advanced search form is easily accessible from the simple search form, but it has a permanent address so that it can be directly accessed (http://curia.europa.eu/juris).

This enhanced search form provides complex and rich search criteria; each criterion is accompanied by a help box. Help is also accessible as a single document in the General information section on the left hand side[75].

Basic criteria from the simple search form are also presented (court, case number, date), and several other criteria are added.

Some criteria can only be searched using predefined vocabularies:

- **Documents** (see 4.2.4.1.3)
- **Subject-matter** (see 4.2.4.1.5.2.2)
- **Procedure and result** (see 4.2.4.1.5.3.1)
- **Systematic classification scheme** (see 4.2.4.1.5.2.1)
- **Formation of the Court** (see 4.2.4.1.5.1.1)
- **Judge-Rapporteur** (see 4.2.4.1.5.1.2)
- **Advocate General** (see 4.2.4.1.5.1.3)
- **Country** (see 4.2.4.1.5.1.4)
- **Authentic language**.

[75] Help is accessible at http://curia.europa.eu/common/juris/en/aideGlobale.pdf# and is available in all official languages.

Other criteria are based on text input:

- Case number
- Name of the parties
- ECLI
- Text

Some criteria are 'combined', which means some values can be selected from predefined options while other values are added as free text:

- **Period or Date** (all types, date of delivery, date of the Opinion, date of the hearing, date of the lodging of the application initiating proceedings + free option to enter date values)
- **References to case law of legislation**

The last type of criteria are predefined options which could be ticked off in the search mask:

- **Case status** (all cases, cases closed, cases pending)
- **Court** (all, Court of Justice, General Court, Civil Service Tribunal)

Search criteria can be combined, but at least one search criterion has to be chosen. The empty search could not be performed with the enhanced search form, although it is possible to complete it with the simple search form.

4.2.4.1.8.2. List of results

The system provides an option to define a list of results before the search is done. The user has the possibility to set 3 different options on how to sort the results from the search:

- The automatic option is the default setting. It groups the list of results by cases unless the criteria 'Documents or 'Text' were used for the search).
- The option 'List of cases' also displays a list of results grouped by cases, but regardless of the criteria used for the search.
- The option 'List of documents' displays a list of documents (and does not depend on the criteria used for the search).

There is another option to define the list of results before launching the search - taking into account descending or ascending orders of cases and dates.

The **list of results by case** shows cases marked by case number and title, followed by the main metadata such as type of proceeding, names of the parties, reference to the Report of Cases and links to the text (link within the system InfoCuria and link to the EUR-Lex – if the document was published therein).

Each case should be further examined by clicking on Case information (Case in progress), which provides details on the document.

The **list of results sorted by documents** provides a comprehensive table of information: the case number, document type (including ECLI number), date of the document, names of the parties, subject-matter, a link to document within InfoCuria system and a link to EUR-Lex.

Criteria used for searching are still clearly visible at the top of the results list page.

4.2.4.1.8.3. Document detail

The detail of the case provides very complex and exhaustive information regarding that particular case divided into 2 parts: the Documents in the case (with lists of all documents belonging to that case) and the Legal Analysis of the decision or the case.

The part labelled 'Legal analysis' provides a lot of information, such as the subject-matter, systematic classification scheme, citations of Case Law or legislation, several dates, references, etc. There is also the Procedural Analysis Information section which informs the user of several additional pieces of legal

information (such as the procedure and the result, the formation of the Court, Judge-Rapporteur, Advocate General, language of the case, etc.).

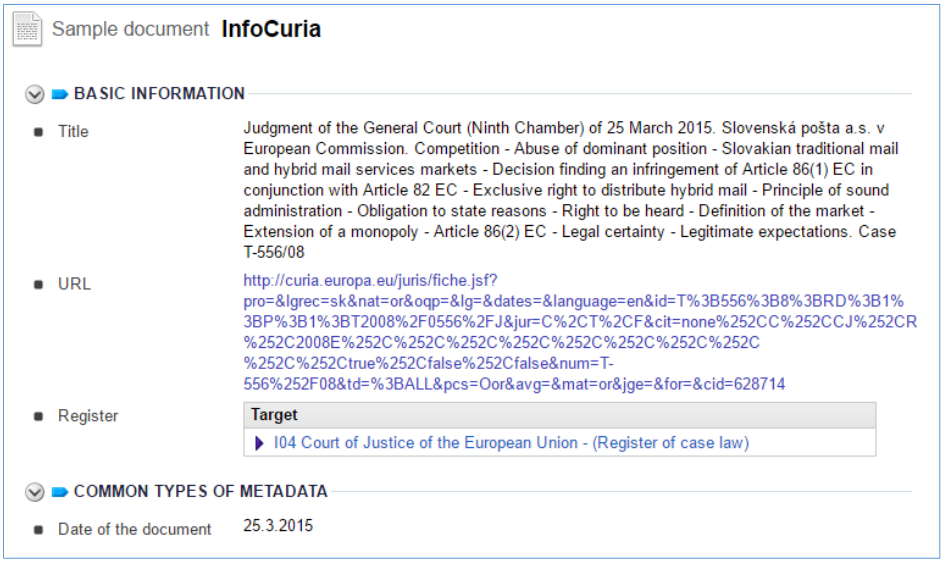
Moving through the sections of the detail of the case is simplified by the menu on the left hand side (which can be completely hidden if needed).

4.2.4.1.8.4. General evaluation of the Search functionality

The search functionality in the InfoCuria document register is very complex and easy to use - with a lot of assistance available to the user in the form of help guides. The user has a wide range of possibilities regarding how to search for what is he looking. Even if the user does not know what to write in the search fields, they are navigated through the predefined lists of values for several criteria.

4.2.4.1.9. Sample case

To see documents from different document sources through a common lens, a randomly selected case from InfoCuria was re-created in the unified structure of the study database. This could help to design the future integrated access solution or help to understand the treatment of metadata from different document registers in one location. The sample case from InfoCuria is shown in Figure 18 and is directly accessible in the study database.[76]



[76] Sample document from the InfoCuria in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1062284.

COMMON TYPES OF VOCABULARIES

Year	Target: 2015
Topic(s)	Target: 4.08.02.00 -General; 4.08.02.02.01 -Product or products constituting a market; 4.08.03.02.08 -Decision of the Commission; 4.08.04 -Public undertakings; Dominant position
Type(s)	Target: 2.1 Judgements
Language(s)	Target: Bulgarian; Croatian; Czech; Danish; Dutch; English; Estonian; Finnish; French; German; Greek; Hungarian; Italian; Latvian; Lithuanian; Maltese; Polish; Portuguese; Romanian; Slovak; Slovenian; Spanish; Swedish

PECULIARITY IN THE INFOCURIA

Documents in case	Judgment (OJ); 24/04/2015 Judgment ECLI:EU:T:2015:18925/03/2015; 25/03/2015 Judgment (Information) ECLI:EU:T:2015:189; 25/03/2015 Application (OJ); 07/03/2009
Subject matter	Annulment of Commission Decision C (2008) 5912 final of 7 October 2008 declaring amendments by Slovakia to its legislation on postal services which extend the monopoly of the historical operator Slovenská Pošta in the provision of hybrid mail services to be contrary to Article 86(1) EC read in conjunction with Article 82 EC (Case COMP/39.562 – Slovakian Postal Law)
References	EC Treaty (Amsterdam), Article 82 EC Treaty (Amsterdam), Article 86 EC Treaty (Amsterdam), Article 253 Directive 97/67 Directive 2002/39 Directive 2008/6 Judgment: OJ C 155 from 11.05.2015, p.20 Application: OJ C 55 from 07.03.2009, p.35
Date lodging	17.12.2008
Parties	Slovenská pošta Commission
Academic writing	Idot, Laurence: Entreprise dotée de droits exclusifs ou spéciaux, Europe 2015 Mois Comm. n° 5 p.24 (FR)
Formation of the Court	Target: (General Court) Ninth Chamber
Judge-Rapporteur	Target: Popescu
Procedure and result	Target: Actions for annulment; Dismissal on substantive grounds

Figure 18: Sample document from InfoCuria register

4.2.4.1.10. Re-use of the InfoCuria in view of integrated access solution

The InfoCuria covers all of the relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT (is the document about)	Document type Topics
WHEN (was the document published)	Years; Dates
WHO (is responsible for the document)	Authorities
WHY (purpose of the document)	Cases; Other metadata

The internal structure (relationships to vocabularies) is very exhaustive.

The main disadvantage of InfoCuria for re-use in any integrated access solution is the absence of machine readability solution.

It is also worth considering whether InfoCuria documents could not be more easily extracted for any integrated access solution from EUR-Lex, as these judicial documents are republished in EUR-Lex in a well-structured form.

4.2.4.2. Judicial Calendar

The Judicial Calendar provides comprehensible access to specific hearings.

The **general analysis** of the Judicial Calendar was carried out.

4.2.4.2.1. Public access

The search form is accessible free of charge at http://curia.europa.eu/jcms/jcms/Jo2_17661.

4.2.4.2.2. Overall volume of hearings

The total number of hearings possible to be retrieved varies depending on the current situation. This means that the Judicial Calendar always shows the information from the last 4 - 6 weeks only: no historical data can be searched.

The search made on April 04, 2016 showed 161 hearings in all.

4.2.4.2.3. Meta-information

The Judicial Calendar uses the vocabulary 'Language of the case' with 24 languages.

Another vocabulary in use is the vocabulary Events, which contains 5 entries:

- Hearing (Opinion of the Court)
- Hearing
- Opinion
- Opinion of the Court
- Delivery of judgment.

However, there are not any helpful explanatory notes that can distinguish between these entries. The user without some specialized knowledge have quite some difficulty discerning between Hearing (Opinion of the Court), Hearing, Opinion and Opinion of the Court.

Another vocabulary which is in use is the vocabulary Court containing 3 branches of the Court:

- Court of Justice
- General Court
- Civil Service Tribunal.

The vocabulary 'Language of the Court', as well as the vocabulary regarding the court, is the same as the one used in InfoCuria.

4.2.4.2.4. End-user search possibilities

There is only one type of search form (no advanced search form is available). The search form is easy to understand.

The user has the option to search according to the

- Date (from – to)
- Events
- Case number
- Language of the case
- Names of the parties
- Court

The results list could be further sorted by date, court, and courtroom. For every hearing, the user will find information available regarding the date, basic information about the case itself (number of the hearing (=case number), names of parties, court, the formation of the court, subject matter), as well as the language of the case and the Courtroom. Moreover, it provides a direct link to the application which brought the case (both HTML and PDF versions are provided).

4.2.4.2.5. Re-use of the Judicial Calendar for the integrated access solution

Taking into account the special character of the Judicial Calendar, its re-use in any integrated access solution is questionable. But if this information has to be re-used, some technical improvements need to be performed.

4.2.4.3. General website search

The website of the Court provides a simple as well as an advanced search of its content. This search could be used for several other parts of the website (for example for Press releases), which do not have their own search forms.

The simple search form only provides free text search (Google-like), but the advanced version makes this general search an interesting option to use.

It allows the user to perform the full text search using

- All words
- At least one word
- The exact phrase

Furthermore, the search could be limited to search just in page titles or page contents or attached files.

It also provides an option to search within a particular predefined time span: documents which were amended since a certain number of days, months or even years; the time range (number of days, months, and years) could be freely entered by the user.

The search form also provides the predefined group tree of all sections of the web page, so the user can filter their search to particular documents.

Regarding Press releases mentioned above, the user can use this general search form, limit the search to Press releases only and then use a full text search or even limit their search to press releases from the past 10 days.

This general website search is available at http://curia.europa.eu/jcms/jcms/j_6/?jsp=jcore%2Fsearch.jsp.

4.2.4.4. Other documents

4.2.4.4.1. Press releases

The list of press releases is available at http://curia.europa.eu/jcms/jcms/Jo2_16799. It provides a chronological view of press releases of the CJEU.

Each press release is accompanied by the title (generally based on the judgment and the case number), languages available, the subject matter of the decision taken and some information about the content of the press release.

There are no special search options available. The option to show Recent press releases is available, as well as to group the press releases according to the year of its publication.

The list of press releases is available at http://curia.europa.eu/jcms/jcms/Jo2_16799.

4.2.4.4.2. Various documents

The web page of the CJEU provides special access to the list of documents classed 'various documents'. These are documents concerning the CJEU and the development of the court system of the EU. No search form is available here.

These documents are available at http://curia.europa.eu/jcms/jcms/P_64268/.

4.2.4.4.3. Administrative documents

All other documents concerning the administrative activity of the CJEU can be accessible through a request form, which is available on the website in an online form as well as a PDF form. The basic point to be made is that these documents do not cover documents concerning cases, as they are publicly available via InfoCuria.

The online request form is available at http://curia.europa.eu/jcms/jcms/P_95917/.

4.2.4.4.4. Re-use of the 'Other documents' document source in view of integrated access solution

These Other documents are of specific content and some of them are not even freely available (only upon request), so their re-use in any integrated access solution is questionable. If they have to be re-used, technical improvements need to be implemented to automatically grant access to them.

4.2.5. European Central Bank

The European Central Bank ('ECB') publishes approximately 11 000 documents (different language versions and attachments are not taken into account) through its website (https://www.ecb.europa.eu).

The total number of available documents, in PDF format, is significantly higher. The ECB website contains ca 198 000[77] PDF files (including all language versions, attachments, and document types, which are outside of the scope of the study).

Documents are not published in any central register of documents, but they are located in several different sections of the website.

A **thorough analysis** of the ECB document source was carried out.

4.2.5.1. General information

As a general rule, the website of the ECB shows for each document a document's date and its name only. The other existing attributes include:

- Related documents (for example press release, summary, annex or older versions of the same document that are no longer valid)
- Comments on the document (for example the issue number and year of the newsletter in which the article is published – e.g. Monthly Bulletin 2000, Issue 4, information on availability)
- Attachments to the document (for example a ZIP file with Microsoft Excel spreadsheet or PDF attachment)
- Document classification (activity type, other hierarchy types such as: 'Economic Bulletin Focus/Economic activity/Projections', JEL Classification[78])
- Author/Authors/Group of authors

4.2.5.1.1. Public access

The public has direct access to all documents in their electronic form. Electronic access is free and no special authentication is necessary.

4.2.5.1.2. Time range covered

The earliest published documents are from the 1970s.

4.2.5.1.3. Overall volume of documents

Approximately 11 000 documents are published on the website. The vast majority of the documents can be found in three sections:

- Research & Publications[79] (circa 4 200 documents),
- Media[80] (circa 4 500 documents)
- About[81] (ca 2 400 documents)

---

77 Investigated through the results of the search according to the file type: https://www.ecb.europa.eu/home/search/html/index.en.html?q=filetype%3Apdf.

78 More information about JEL classification: https://www.aeaweb.org/econlit/jelCodes.php?view=jel.

79 ECB website - Research & Publications section: https://www.ecb.europa.eu/pub/html/index.en.html.

80 ECB website - Media section: https://www.ecb.europa.eu/press/html/index.en.html.

81 ECB website - About section: https://www.ecb.europa.eu/ecb/html/index.en.html.

More detailed analysis of the frequency of document publication (the number of documents published each year) in these sections is outlined in the following chapters.

A smaller amount of documents was also published on other sections of the ECB website. For example, there are a range of documents including Meeting minutes published in the subsection 'Market contact groups'.[82] There are several documents assigned to each meeting (Agenda, Summary and sometimes several presentations).

4.2.5.1.3.1. Research & Publication section volumes

The following Table 5 shows the number of documents in the 'Research & Publication' section divided by the activity.

Activity	Count
ECB	61
Monetary policy	1 174
Statistics	482
Payments & markets	213
Financial stability	170
International and European co-operation	397
Banknotes & coins	38
Legal	24
Total	2 559

Table 5: The number of documents in the 'Research & Publication' section divided by the activity

The next Table 6 shows the number of documents in each subsection of the 'Research & Publication' main section.

Subsection	Count
Economic research	125
Paper series	2 077
Conferences & seminars[83]	89
Economic Bulletin	1 144
Annual Report	28
Convergence Report	13
Financial Stability Review	23
Macroeconomic projections	47

Table 6: The number of documents in each subsection of the 'Research & Publication'

The following diagram illustrates the annual increments of publication in the 'Research & Publication' section. An increase of approximately 400 – 450 new publications may be forecasted for the year 2016.

[82] Market contact groups: https://www.ecb.europa.eu/paym/groups/html/index.en.html.
[83] Documents in HTML format.

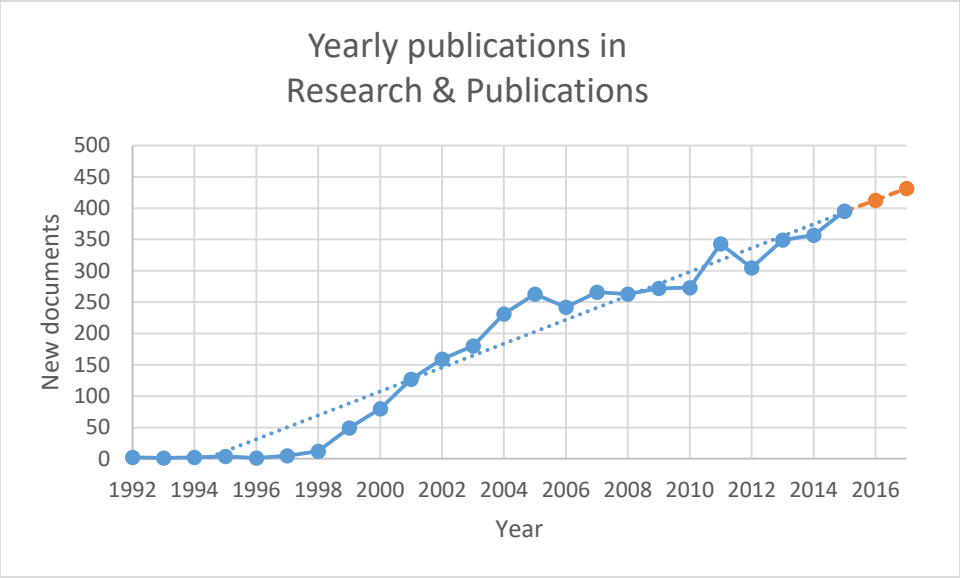


Figure 19: The annual increase of publications in the 'Research & Publication' section

4.2.5.1.3.2. Media section volumes

The following Table 7 shows the number of documents in the 'Press releases' subsection divided by the activity:

Activity	Count
ECB	430
Monetary policy	483
Statistics	183
Payments & markets	220
Financial stability	91
International and European co-operation	218
Banknotes & coins	104
Legal	28
Banking supervision	22
Others	92
Total	1 871

Table 7: The number of documents in the 'Press releases' subsection divided by the activity

The following Table 8 shows the number of documents in each subsection:

Subsection	Count
Press releases	1 871
Governing Council decisions	368
Press conferences	212
Monetary policy accounts	9
Speeches	1 895
Interviews	208

Table 8: The number of documents in each subsection

Table 8 shows that the 'Press Releases' and 'Speeches' subsections have the largest quantity of documents. The diagram below illustrates the annual increase in these two subsections. 'Governing Council Decisions' subsection is another more relevant part, which has regular increments of approximately 24 per year.

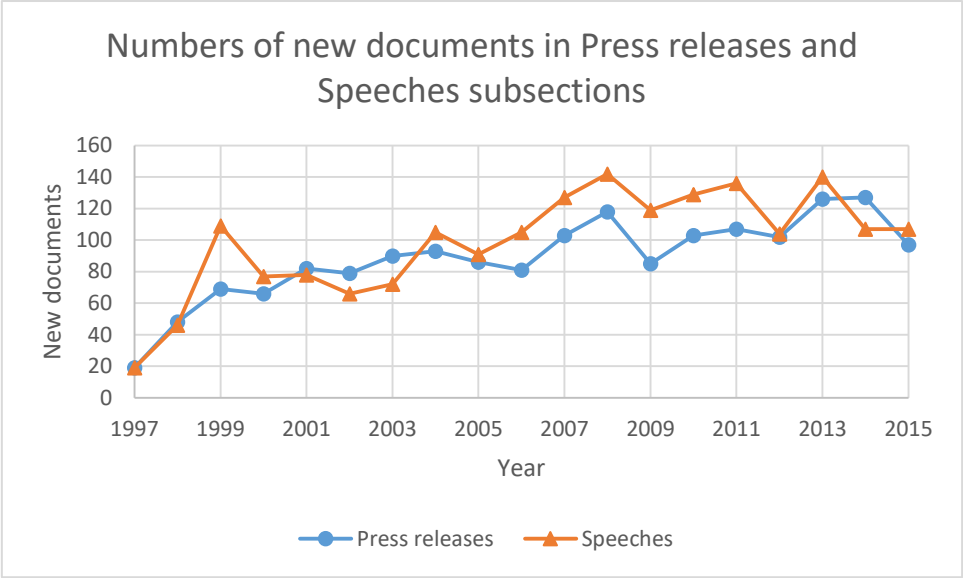


Figure 20: The annual increase of documents in the 'Press Releases' and 'Speeches'

4.2.5.1.3.3. About section volumes

In terms of the number of documents in the section 'About', the two most relevant subsections are: 'Legal framework' (1930 documents) and 'Procurement' (527 tenders). Most of the documents (90%) in the 'Legal framework' subsection are 'Opinions of the ECB.'

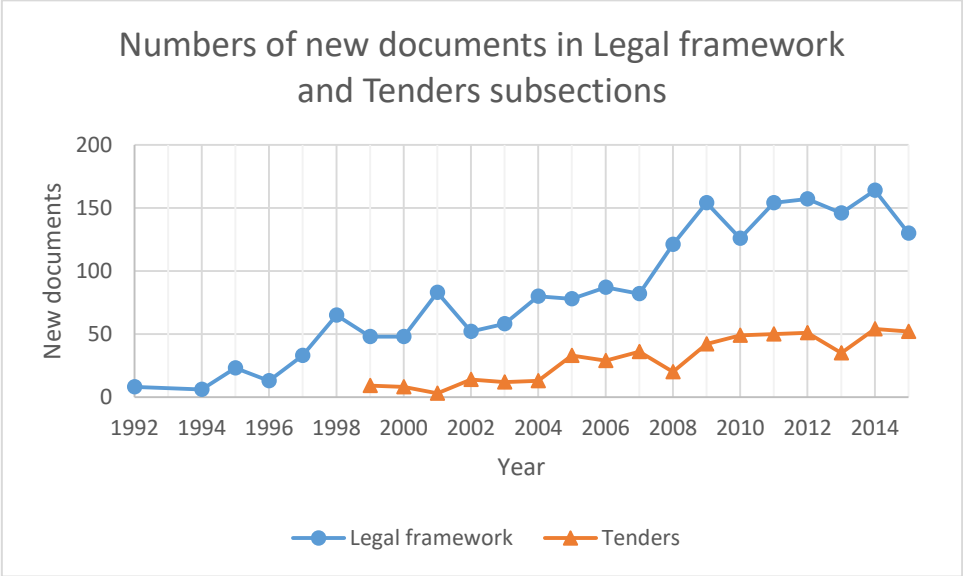


Figure 21: The number of new documents in the 'Legal framework' and 'Tenders' subsections

4.2.5.1.4. Brief investigation of ECB documents

Figure 22 shows the results of the comparative analysis of the ECB.[84]

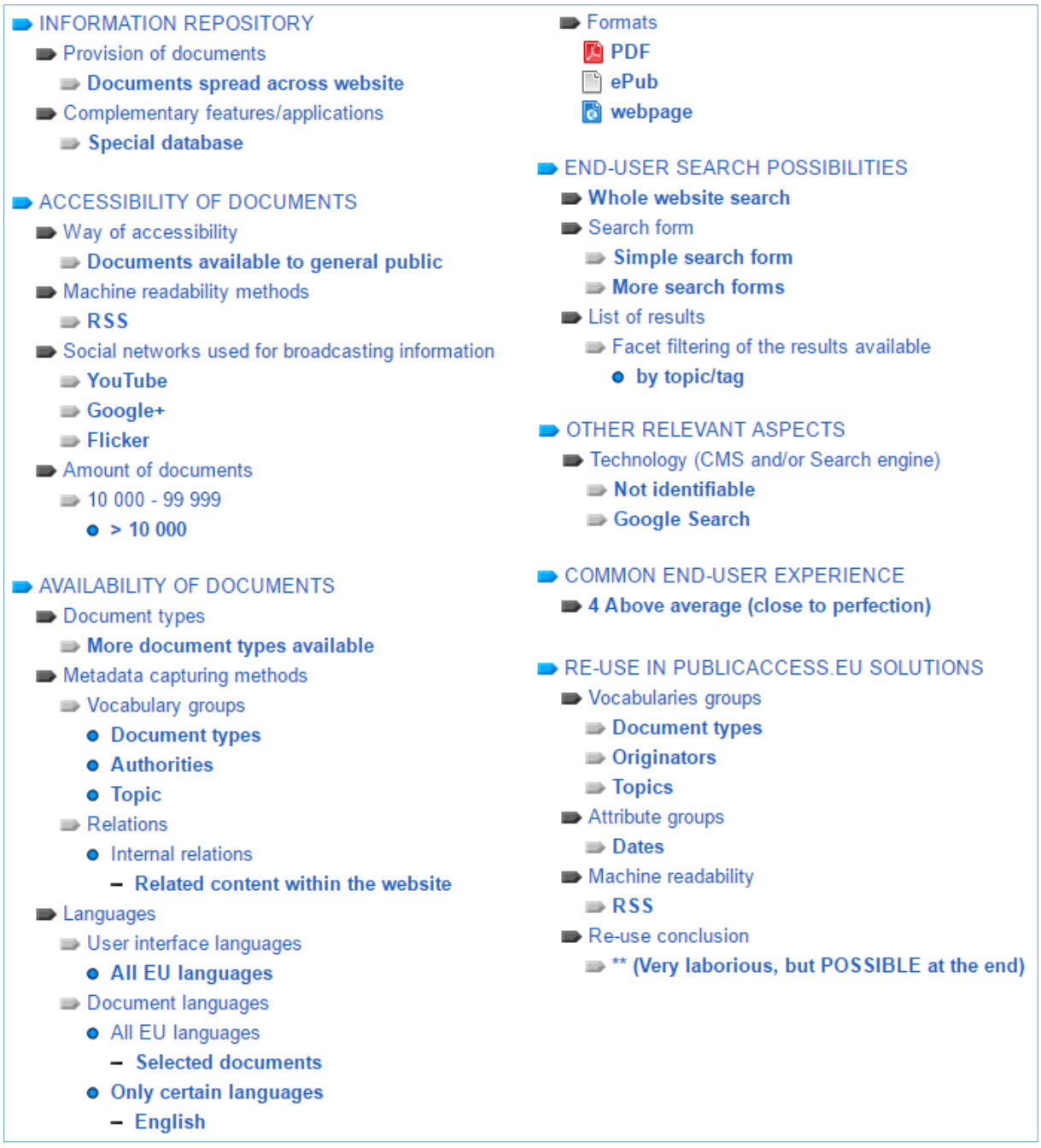


Figure 22: Overview investigation of the register of ECB documents

[84] Brief investigation of ECB website in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000025.

4.2.5.2. Document types

There is no visible document type classification at the ECB website.

The following main document types were found on the ECB website[85]:

- Article
- Opinion
- Bulletin
- Agenda
- Press release
- Interview
- Speech
- Annual report, Convergence report
- Stability review
- Projection
- Decision
- Guideline
- Opinion
- Recommendation
- Regulation

Some document types can be established indirectly from EUR-Lex.[86] The statistics found on this website indicate that there are currently about 1 000 published documents. From this amount, there are 175 documents of the type 'Opinion'. It was found that only about 10 % of the total number of documents of this type published on ECB website are also published on EUR-Lex.

4.2.5.3. Metadata as relationships between documents and vocabularies

4.2.5.3.1.1. Author

The vocabulary of 'Authors' is used for the documents in the 'Paper series'. Some authors are also organised into groups (for example the 'BACH Working Group'). If the metadata is assigned in this way, it is possible to view all the publications of a given author.

4.2.5.3.1.2. JEL Classification

This vocabulary is used to classify documents in the 'Paper series', but only in text form.

4.2.5.3.1.3. Documents category

Documents are often grouped into categories. For example, the articles in the Economic Bulletin are included in one of the thirteen categories[87] (for instance 'Monetary and financial' or 'Payment systems'). This categorization may be considered as document metadata as well.

---

[85] However, this list was derived from an investigation of the website and cannot be deemed a complete list of document types.

[86] All documents of the ECB published on EUR-Lex: http://eur-lex.europa.eu/search.html?DB_AUTHOR=cb&lang=en&SUBDOM_INIT=ALL_ALL&DTS_DOM=ALL&CASE_LAW_SUMMARY=false&type=advanced&qid=1456318183788.

[87] ECB website - Economic Bulletin Articles: https://www.ecb.europa.eu/pub/economic-bulletin/articles/html/index.en.html.

4.2.5.4. Metadata as document attributes

The attributes published on the ECB website include:

- Document title
- Document date (not displayed for all documents)
- Document notes (e.g. the location where the document was published)
- Topics of the document[88]
- Closing date ('Tenders' section[89])
- Procedure state ('Tenders' section[89])
- Notice identifier ('Tenders' section[89])

4.2.5.5. Relations between documents

The internal relations between the documents include for example:

- Document attachments (ZIP, CSV)[90]
- Links to previous (no longer valid) versions of the same document[91]
- Links to the document's press release
- Links to another web page related to the document's content
- Related documents (Annual report & Annual accounts[92], all documents and presentations given in the meetings[93])
- Summary of the document

4.2.5.6. End-user search possibilities

The ECB website does not provide an integrated global search of documents. A search box in the page header uses 'Google Custom Search'[94]; the results are arranged in five categories: Legal, Press, T2S, Publications and Statistics.

Specific subsections of the website containing documents typically offer their own search box. This allows document names and dates to be searched for in the given list of documents.

4.2.5.6.1.1. Search form

This section analyses the search options that are available to the end-user on the ECB website, and it covers three main areas in line with section 4.1, i.e. the search form, the list of results, and the document detail with a brief conclusion.

As already described, the document search is limited only to the name of the document and possibly to the textual representation of the document date. Full text search in the document texts is not

[88] ECB website – Research Bulletin: https://www.ecb.europa.eu/pub/economic-research/resbull/html/index.en.html.
Economic Bulletin Articles (topics): https://www.ecb.europa.eu/pub/economic-bulletin/articles/html/index.en.html.

[89] ECB website – Tenders section: https://www.ecb.europa.eu/ecb/jobsproc/tenders/html/index.en.html

[90] Examples can be found here: https://www.ecb.europa.eu/pub/research/statistics-papers/html/index.en.html

[91] ECB website – example of additional information with documents no longer in force: https://www.ecb.europa.eu/ecb/legal/1001/1010/html/act_11213_amend.en.html.

[92] ECB website – Annual Report section: https://www.ecb.europa.eu/pub/annual/html/index.en.html.

[93] ECB website – example of documents and meeting information regarding Foreign exchange contact group: https://www.ecb.europa.eu/paym/groups/fxcg/html/index.en.html.

[94] Google Custom Search Engine: https://cse.google.com/cse/?hl=en.

supported in the search box present above the lists of the documents. However, the global search also considers the text of the documents. In any case, it does not find all the documents based on their metadata.

4.2.5.6.1.2. List of results

Search results do not provide sufficient information. It is possible to filter the results according to the five categories mentioned above only when using 'Google Custom Search'.

4.2.5.6.1.3. Document detail

A 'Document Details' page does not exist.

4.2.5.6.1.4. General evaluation of the Search tool

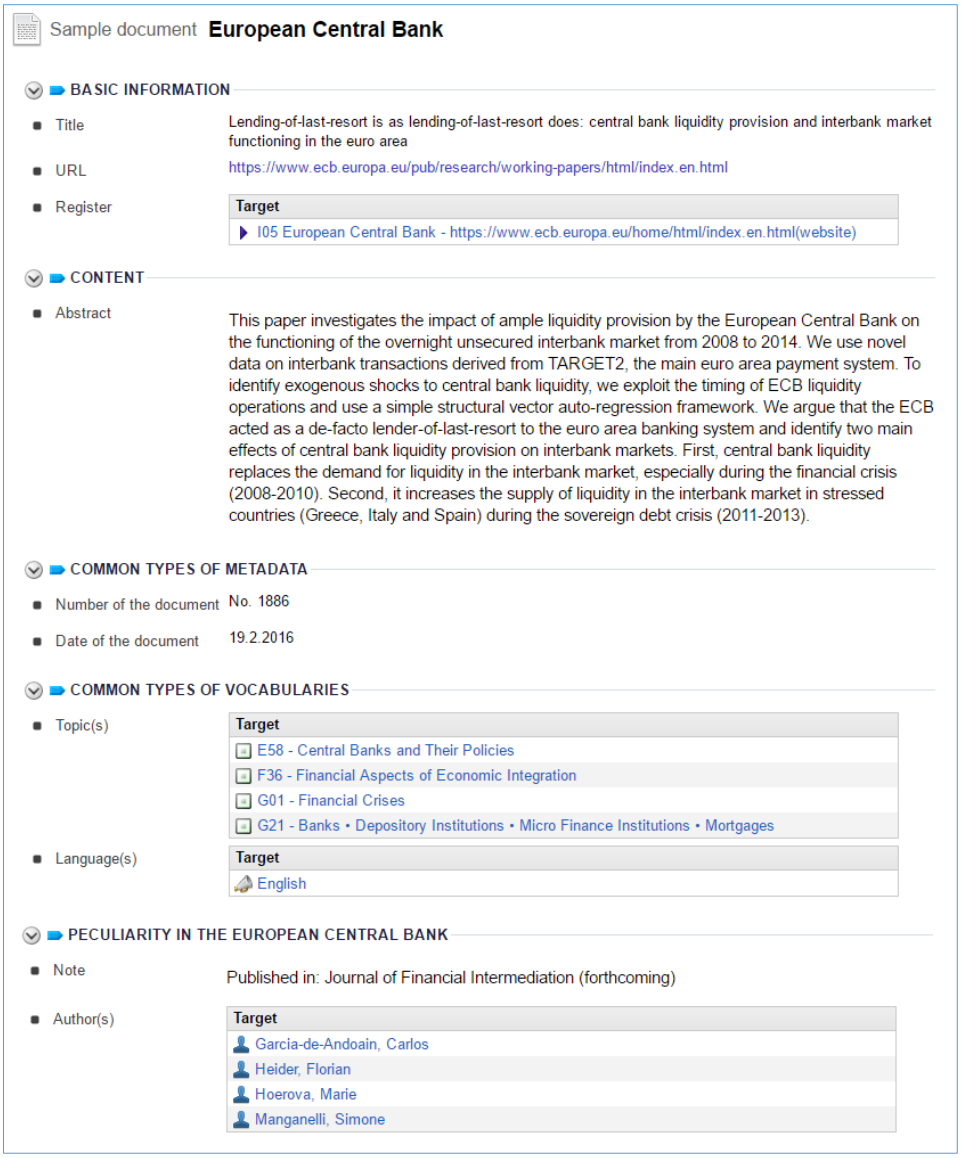
The documents published on the ECB's website are clearly and comprehensively divided according to the structure of the website. However, it is likely that extending the global search to use the metadata for filtering would increase usability.

4.2.5.7. Sample document

To see documents from different document sources through the same lens, a randomly selected document from ECB website was re-created in the unified structure of the study database. This could help to design the future integrated access solution or help to understand the treatment of metadata from different document registers in one location. The sample document from the ECB website is shown in Figure 23 and is directly accessible in the study database.[95]

The selected document lacks the document type. The document type may be deduced to be an 'article' or 'working paper,' but given the fact that this information is not explicitly mentioned on the website it is also not included in Figure 23.

[95] Sample document from the ECB in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1054030.



Sample document **European Central Bank**

BASIC INFORMATION

Title	Lending-of-last-resort is as lending-of-last-resort does: central bank liquidity provision and interbank market functioning in the euro area
URL	https://www.ecb.europa.eu/pub/research/working-papers/html/index.en.html
Register	Target: I05 European Central Bank - https://www.ecb.europa.eu/home/html/index.en.html(website)

CONTENT

Abstract	This paper investigates the impact of ample liquidity provision by the European Central Bank on the functioning of the overnight unsecured interbank market from 2008 to 2014. We use novel data on interbank transactions derived from TARGET2, the main euro area payment system. To identify exogenous shocks to central bank liquidity, we exploit the timing of ECB liquidity operations and use a simple structural vector auto-regression framework. We argue that the ECB acted as a de-facto lender-of-last-resort to the euro area banking system and identify two main effects of central bank liquidity provision on interbank markets. First, central bank liquidity replaces the demand for liquidity in the interbank market, especially during the financial crisis (2008-2010). Second, it increases the supply of liquidity in the interbank market in stressed countries (Greece, Italy and Spain) during the sovereign debt crisis (2011-2013).

COMMON TYPES OF METADATA

Number of the document	No. 1886
Date of the document	19.2.2016

COMMON TYPES OF VOCABULARIES

Topic(s)	Target: E58 - Central Banks and Their Policies; F36 - Financial Aspects of Economic Integration; G01 - Financial Crises; G21 - Banks • Depository Institutions • Micro Finance Institutions • Mortgages
Language(s)	Target: English

PECULIARITY IN THE EUROPEAN CENTRAL BANK

Note	Published in: Journal of Financial Intermediation (forthcoming)
Author(s)	Target: Garcia-de-Andoain, Carlos; Heider, Florian; Hoerova, Marie; Manganelli, Simone

Figure 23: Sample document from the register of ECB documents

4.2.5.8. Re-use of the Register of ECB documents in view of integrated access solution

The Register of ECB documents covers most of the relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT (is the document about)	Document type Topics
WHEN (was the document published)	Dates
WHO (is responsible for the document)	Authorities
WHY (purpose of the document)	Limited metadata

The main disadvantage of Register of ECB documents for re-usability in any integrated access solution is the absence of machine readability solution. However, customizable RSS channels could be used.

4.2.6. European Court of Auditors

The European Court of Auditors ('ECA') audits EU finances. Its role is to improve EU financial management and report on the use of public funds. It is one of the EU institutions.

4.2.6.1. General information

4.2.6.1.1. Public access

The ECA has a general website with documents spread across the whole website.

The ECA website is accessible at this general URL address: http://www.eca.europa.eu/en/Pages/ecadefault.aspx.

Documents resulting from the professional activities of the ECA can be found in the ECA Register of publications, available at http://www.eca.europa.eu/en/Pages/PublicationSearch.aspx (sections Our products/Search publications).

Some publications in this register are equipped with ISSN/ISBN code and are simultaneously published in the EU Bookshop as well.

Several other documents are also published on this website:

- ECA Rules of procedure[96]
- Ethics[97]
- Some bibliographies[98]
- A booklet about ECA history[99]
- Public procurement, Press releases, Job opportunities[100]
- Reports by the external auditor[101]

The whole content published on the ECA website is accessible to the general public without any limitation and no special authentication is needed.

A **thorough analysis** of the ECA Register of publications was carried out.

4.2.6.1.2. Time range covered

The ECA Register of publications covers the period since the establishment of ECA. Therefore, it contains documents published from 1977. However, up until the year 2000 the majority of documents are copies of documents published in the OJ.

---

96 Rules of procedure of ECA: http://www.eca.europa.eu/Lists/ECADocuments/RULES_PROCEDURE_2010/RULES_PROCEDURE_2010_EN.PDF.

97 ECA website – Ethics section: http://www.eca.europa.eu/en/Pages/Ethics.aspx.

98 ECA website – Library section: http://www.eca.europa.eu/en/Pages/LibraryArchives.aspx.

99 Booklet reflecting 35th Anniversary of ECA: http://www.eca.europa.eu/Lists/ECADocuments/REFLECTIONS_35TH_ANNIVERSARY/Reflections_35anniversary.pdf.

100 ECA website – Public procurement section: http://www.eca.europa.eu/en/Pages/PublicProcurement.aspx.

101 ECA website – Reports by external auditor section: http://www.eca.europa.eu/en/Pages/ExternalAuditor.aspx.

4.2.6.1.3. Overall volume of documents

The functional limitations of the ECA Register of publications make it difficult to determine the total number of documents. Some indications show that the number of documents published in the English language do not exceed 5 000 and the number of documents in all language versions do not exceed 50 000.

4.2.6.1.4. Brief investigation

Figure 24 shows the result of the comparative analysis of the ECA Register of publications.[102]

Figure 24: Overview of the investigation results of the ECA Register of publications

[102] Brief investigation of the ECA register of publications in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000027.

4.2.6.2. Document types

The document type is determined by the relationship between one specific item in the vocabulary of the document types and the document itself.

The following document types were extracted from the search form[103]:

- Annual report
- Opinion
- Special Report
- Specific Annual Report
- Activity Report
- Landscape review
- EU Audit in Brief
- Work programme
- Letter of the President

4.2.6.3. Metadata as relationships between documents and vocabularies

Analysis of the ECA Register of publications is based on the investigation of the search form possibilities and the answers given by ECA representatives to the general questions.

4.2.6.3.1. Topic vocabularies

4.2.6.3.1.1. Topic - Spending Area

The vocabulary 'Spending Area'[104] represents the specific area that the document covers. It is reinforced by the relationship Document ←→ Topic. This vocabulary is a list of 17 items.

4.2.6.3.2. Additional vocabularies

4.2.6.3.2.1. Year

The vocabulary 'Year' is a useful extract of the document dates.

The vocabulary Year allows the document to be positioned in time. It was found useful for the future integrated access solution and that is why it is additionally analysed in the study database.[105]

4.2.6.3.2.2. Language

The vocabulary 'Language' is used to identify the language in which the document is available through the relation between the document and entries in the vocabulary Language. It is used only in the document detail and not used in the search form, so no additional investigation was performed.

4.2.6.4. Metadata as document attributes

The following list of document attributes in the ECA Register of publications is based on an investigation of search possibilities and the answers given by ECA representatives to the general questions:

- Free text
- Document identifier
- Document date
- Date of publication

[103] Document types of the ECA register of publications in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1052001.

[104] Spending Area vocabulary of the ECA register of publications in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1052021.

[105] Controlled vocabulary Years of the ECA register of publications in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1052066.

4.2.6.5. Relations between documents

Neither internal nor external relationships between documents were investigated.

4.2.6.6. End-user search possibilities

This section analyses the search options that are available to the end-user in the ECA register of publications, and it covers three main areas in line with section 4.1, i.e. the search form, the list of results, and the document detail with a brief conclusion.

4.2.6.6.1. Search form

The Advanced search is accessible from the top menu of the ECA register of publications under the option 'Our products'/'Search publications'.

It includes a form with filtering capabilities based on four criteria:

1. Keywords
    - There is no full text search within documents available, but only filtering of the strings contained in one specific field (named Freetext), which is filled during the publication of the document.
2. Type
    - This filter is populated by items from the vocabulary Document type (see 4.2.6.2).
3. Spending Area
    - This filter is populated by the items from the vocabulary Spending Area (see 4.2.6.3.1.1).
4. Year
    - This filter is populated by the items from the vocabulary Year (see 4.2.6.3.2.1).

Usage of more than one criterion is handled as a chain linked by the logical AND operator, which means the more criteria are used, the fewer results are produced.

4.2.6.6.2. List of results

After clicking the Search button without applying any filters, the user gets a list of all documents that may be narrowed according to their requirements by repeating the search with some of the filters applied. The number of documents found is not shown, nor can it be derived from any information. The user can only browse from page to page.

The List of results consists of links to documents in the Document Register of the ECA, where each document entry is attributed to the following information/meta-information:

- Date of the document
- Document title in the language selected for the whole website
- Document type
- Link to the document (in PDF od EPUB format in selected documents format with document language selection possibility)

Response time during the generation of results is quite slow, which makes it difficult to perform any in-depth analysis.

4.2.6.6.3. Document details

Clicking the document title opens document details.

The document details contain the following information:

- Date of document
- Document title in the language selected for the whole website
- Annotation (several lines of text)
- Link to the document (in PDF or EPUB format with language choice)

4.2.6.6.4. General evaluation of the Search functionality

The Search tool functionality is clear for the common user except that what can be entered into the field 'Keywords' is not explained.

The disadvantages are:

- Absence of full text search in the content of documents;
- Slow response time, the generation of the list or results takes > 5 seconds and sometimes even multiples of this period.

4.2.6.7. Sample document

To see documents from different document sources through the same lens, a randomly selected document from the ECA Register of publications was re-created in the unified structure of the study database. This could help to design the future integrated access solution or help to understand the treatment of metadata from different document registers at one location. The sample document from the ECA Register of publications is shown in Figure 25 and is directly accessible in the study database.[106]

Figure 25: Sample document from the ECA Register of publications

[106] Sample document from the ECA Register of publications in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1052126.

4.2.6.8. Re-use of the ECA Register of publications in view of integrated access solution

The ECA Register of publications covers most of the relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT **(is the document about)**	Document type Topics not available
WHEN **(was the document published)**	Years Dates
WHO **(is responsible for the document)**	Not available
WHY **(purpose of the document)**	Limited metadata

The main disadvantage of ECA Register of publications re-usability in any integrated access solution is the absence of machine readability solution, but customizable RSS channels could be used.

4.2.7. Committee of the Regions

The Committee of the Regions ('CoR') is an EU consultative body with 350 members, representing local and regional authorities from 28 member countries. It must be consulted during EU decision-making in the fields of transport, employment policy, education, culture, public health, economic, social, territorial cohesion, environment, and energy.

4.2.7.1. CoR website general information

The CoR has a very complex website full of information on its organisation activities, members, events, etc. which is accessible at this general URL address:
http://cor.europa.eu/en/Pages/home.aspx.

The following important document sources were found on the CoR website and analysed separately:

1. CoR Documents Manager
2. CoR Members' Portal
3. CoR Studies/Brochures register
4. CoR General website search

The published documents are accessible to the general public without any restrictions. Document sources Nos. 1 and 2 include a non-public section, where login is needed and user accounts are not available to the general public. However, it is possible to request the unpublished documents.

The CoR website includes published documents in the above-mentioned sections of the website throughout its whole existence, i.e. the period since 1994.

4.2.7.2. Committee of the Regions Documents Manager

4.2.7.2.1. CoR-DM general information

CoR Documents Manager ('CoR-DM') is the main source of documents published on the CoR website.

CoR-DM consists of three sections:

1. Opinions
2. Public Documents (RED)
3. COM documents

A **thorough analysis** of the CoR Documents Manager document source was carried out.

4.2.7.2.1.1. Public access

The CoR-DM is accessible at this general URL address:
https://dm.cor.europa.eu/corDocumentSearch

4.2.7.2.1.2. Overall volume of published documents

The volume of published documents up to the end of February 2016 is shown in Table 10:

CoR-DM section	Documents in English	Documents in all languages
Opinions	1 505	23 390
Public Documents (RED)	15 570	225 314
COM documents	9 585	122 907

Table 9: Volume of published documents in CoR-DM

The section COM documents has a 1:1 overlap with the RD-EC. This means that all documents included in this section are also published in the RD-EC.

In the section Public Documents (RED) a slight overlap (163 documents) with the EESC register of documents was found.

All three sections have almost identical search functionality. Therefore, they are not further differentiated.

4.2.7.2.1.3. Brief general investigation of the CoR-DM

Figure 26 shows the result of the comparative analysis of the CoR-DM.[107]

Figure 26: Overview investigation of the CoR Documents Manager

[107] Brief investigation of the CoR-DM in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1052272.

4.2.7.2.2. CoR-DM Document types

The vocabulary of 'Document types' in the CoR-DM[108] contains a flat list of document types.

The investigation of CoR-DM Document types is based on the extraction of document types from the facet filter in the list of results. Given the fact that this filter includes only the top 20 items, CoR-DM Document Types list below cannot be considered as complete. Specifically, this means that there are 3 752 documents (about 1.6% of the total) other than the types listed below.

The following Document types were extracted:

- Amendment
- Amendment to the commission meeting
- Amendment to the plenary session
- Annex
- Committee opinion adopted in plenary session
- Consultative work, various
- Draft opinion
- Draft opinion of the Committee
- Invitation
- Invitation, proposal of programme negotiations
- Minutes
- Minutes from the meeting
- Bureau
- Point plenary session
- Press Release
- Rapporteur's amendment to the commission meeting
- Rapporteur's amendment to the plenary session
- Resolution
- Working document

The relationship between a document and entries in the vocabulary of CoR-DM Document types has 1:1 cardinality. This means that each document must be of exactly one document type.

4.2.7.2.3. Metadata as relationships between documents and vocabularies

4.2.7.2.3.1. Document originator vocabularies

Document originator vocabularies are used to specify the origin and responsibility for the documents in the CoR-DM. Two separate vocabularies serve this purpose:

1. Dossier
2. Rapporteur

4.2.7.2.3.1.1. Vocabulary of Dossiers

The vocabulary of 'Dossiers'[109] is represented by a list of selected CoR Commissions organised as a flat list where each Commission's mandate period is represented by one standalone entry.

The investigation into the facet filter items in the list of results shows that:

- Dossier is further specified by the attribute number.
- Commissions used in Dossiers are also included in the Document Search source (in the general website search) but already without specification of the mandate period.

The relationship between a document and entries in the vocabulary of Dossiers has 1:0…1 cardinality. This means that each document can be attached to either none or one Commissions (Dossier).

[108] CoR-DM vocabulary of Document types in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1052281.

[109] CoR-DM vocabulary of Dossiers in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1052781.

4.2.7.2.3.1.2. Vocabulary of Rapporteurs

The vocabulary of 'Rapporteurs'[110] is a list of Members of the CoR responsible for the documents. It is organised as a flat list of 196 entries.

The investigation shows that:

- Rapporteurs are stated here only by their family names.
- There is no reference to the database of CoR members.[111]

The relationship between a document and entries in the vocabulary of Rapporteurs has 1:0…1 cardinality. This means that the document can be attached to either none or one rapporteur.

4.2.7.2.3.2. Time range vocabularies

Time specifications are handled in the CoR-DM by three independent vocabularies:

1. Years
2. Plenary sessions
3. Months in Years

It may be concluded that the vocabulary of Dossiers (see 4.2.8.1.3.1.2) also specifies the time range in the information of the mandate period of the Commissions.

4.2.7.2.3.2.1. Vocabulary of Years

The vocabulary of 'Years'[112] brings the basic orientation in time to the user searching for the document in the CoR-DM.

The vocabulary is prepared as an extract from the document dates.

The relationship between a document and entries in the vocabulary of Years has 1:1 cardinality. This means that each document must have exactly one entry associated with it in the vocabulary of Years.

4.2.7.2.3.2.2. Vocabulary of Plenary sessions

The vocabulary of 'Plenary sessions'[113] defines the event in which the document was adopted.

It was discovered that:

- It is only used in the Opinions section of the Document manager.
- It is used from the 99th plenary session, which was held between 31 January 2013 and 1 February 2013.
- Entries in the vocabulary are composed of the number of the Plenary session.

The relationship between the opinion document and entries in the vocabulary of Plenary sessions has 1:0…1 cardinality. This means that each opinion document can belong to either none or one entry in the vocabulary of Plenary sessions.

---

[110] CoR-DM vocabulary of Rapporteurs (Members of CoR) in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1052340.

[111] List of CoR members at the source http://cor.europa.eu/en/search-center/Pages/members.aspx?Function=Member.

[112] CoR-DM vocabulary of Years in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1052882.

[113] CoR-DM vocabulary of Plenary sessions in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1052831.

4.2.7.2.3.2.3. Vocabulary of Months in Years

The vocabulary of 'Months in Years'[114] is used for two purposes:

1. It determines the month in the year in which the document was adopted.
2. It determines the month in which the meeting was held.

It was found that it is an extract from the relevant dates.

The relationship between a document and entries in the vocabulary of Plenary sessions has 1:0…1 cardinality. This means that each document can belong to either none or one entries in the vocabulary of Months in Years which is either:

- The Adoption date or,
- The Meeting date

4.2.7.2.3.3. Vocabulary of Languages

The last vocabulary intensively used in the CoR-DM is the vocabulary of 'Languages'.[115] It is used to identify both

- The original language of the document,
- The additional languages into which the document has been translated,

and for providing the document in the requested language version to the user.

The relationship between a document and entries in the vocabulary of Languages has 1:1…n cardinality. This means that each document must belong to at least one entry in the vocabulary of Languages.

4.2.7.2.4. Metadata as document attributes

The following document attributes were discovered, which are mandatory for each document in the CoR-DM:

- Dossier
  - Number in Dossier (see 4.2.8.1.3.1.2 for more information about Dossier in the CoR-DM)
- Document
  - Code (composed of the codes of vocabulary entries and the document attributes)
  - Number of document
  - Status of document
  - Production date
  - Meeting date
  - Document format (also used as a standalone facet in the lists of results)

4.2.7.2.5. Relations between documents

4.2.7.2.5.1. Internal relations

One internal relationship was found between the document and dossier. It is used only selectively and the usage depends on the Document type.

---

[114] CoR-DM vocabulary of Months in Years (used in the CoR-DM for the Adoption date and the Meeting date) in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1061358.

[115] CoR-DM vocabulary of Document languages the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1052361.

4.2.7.2.5.2. External relations

Some documents are related to the procedure number.

4.2.7.2.6. End-user search possibilities

This section analyses the search options that are available to the end-user in the CoR-DM and it covers three main areas in line with section 4.1, i.e. the search form, the list of results, and the document detail with a brief conclusion.

4.2.7.2.6.1. Search form

By default, the search is carried out through a query word or phrase in the simple search form. The content of documents is indexed and searchable.

In the 'Opinion' section the simple search form is accompanied by filters for:

- Plenary session
- Commission
- Rapporteur
- Document language

Usage of more than one criterion is handled as a chain combined with the logical AND operator, which means that the more criteria are used, the fewer results are produced.

4.2.7.2.6.2. List of results

There are three switchable views on the generated list of the result: default, table, compact. The most useful is the table view on the List of results. This consists of links to documents in the CoR-DM, where each document entry is characterized by a rich set of the following metadata:

- Dossier
- Document Number
- Document type
- Original language
- Additional languages
- Year of the document
- Status of the document
- Rapporteur
- Document date
- Meeting date
- Procedure

Sorting by the metadata is available, but it only works with a small number of results.

The List of results is further equipped with additional options for reducing the volume of documents by filtering through the facet filters for which the vocabularies specified in chapter 4.2.7.2.3 are used.

Although the document titles are not composed in a unified form, they are descriptive enough. One additional useful functionality is implemented: each specific list of results is also available as an RSS feed.

4.2.7.2.6.3. Document detail

There is no document details page. After clicking on the title of the document, it is directly downloaded.

4.2.7.2.6.4. General evaluation of the Search functionality

The search functionality works very well. However, the user interface is not targeted to end-users outside of the CoR and it is not easy to get familiar with it because it requires a lot of experience to use. User manuals for the Search function exist – but it requires login into the internal system.

4.2.7.2.7. Sample document

To see documents from different document sources through the same lens, a randomly selected document from CoR-DM was re-created in the unified structure of the study database. This could help to design the future integrated access solution or help to understand the treatment of metadata from different document registers in one location. The sample document from CoR-DM is shown in Figure 27 and is directly accessible in the study database[116].

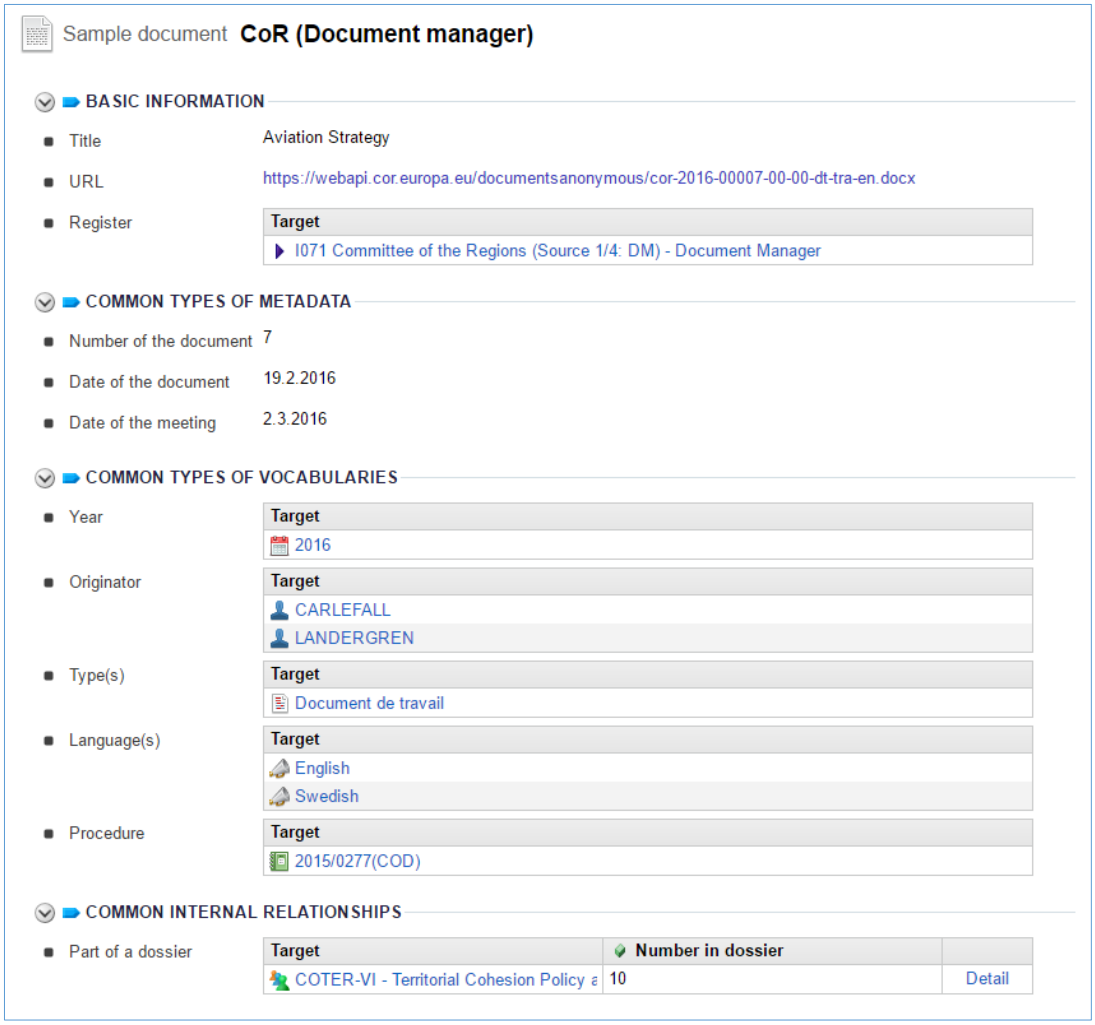


Figure 27: Sample document from the CoR Documents Manager

4.2.7.2.8. Re-use of the CoR-DM in view of integrated access solution

The CoR-DM covers most of the relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

[116] Sample document from the CoR-DM in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1061529.

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT (is the document about)	Document type Topics not available
WHEN (was the document published)	Dates
WHO (is responsible for the document)	Authorities
WHY (purpose of the document)	Procedure, Other metadata

The CoR-DM is equipped with many metadata. However, the vocabularies are not publicly available and they are not even described (e.g. what exactly the document types represent). The major deficiency is the absence of any thematic information about the documents.

CoR-DM does not provide API or any other web service for automatic retrieval of the content. It only provides RSS functionality, which could be considered as a source for retrieving documents for any possible integrated access solution.

4.2.7.3. Committee of the Regions Members' Portal

CoR Members' Portal - Transfer of Administrative Documents ('TOAD') is composed of:

- Information on CoR meetings and the provision of relevant meeting documents to CoR Members;
- Search functionality on CoR works currently in progress;
- Various links to other places of CoR website.

The subsequent analysis of the TOAD is brief (i.e. not fully according to the principles described in the chapter 4.1) due to the following reasons:

- TOAD is primarily targeting internal CoR users, mainly CoR members;
- There is a low volume of documents;
- There is an overlap (document duplicates) between TOAD and CoR-DM.[117]

Only the **general analysis** of the CoR Members' Portal was carried out.

4.2.7.3.1. Public access

The TOAD is accessible at this general URL address: https://toad.cor.europa.eu/. A user guide is available at the address http://www.toad.cor.europa.eu/CORHelp.aspx.

The search tool for opinions and resolutions is available at this URL address: https://toad.cor.europa.eu/CORWorkInProgress.aspx.

[117] Examples of the same document
in TOAD: https://webapi.cor.europa.eu/documentsanonymous/cor-2015-06328-00-00-pac-tra-en.docx
in CoR-DM: https://toad.cor.europa.eu/ViewDoc.aspx?doc=obsolete%5cEN%5cCOR-2015-06328-00-00-PAC-TRA_EN.docx&docid=3138080.

4.2.7.3.2. Overall volume of dossier entries

The volume of published dossier entries in the TOAD up to the end of February 2016 is about 1 200. Sometimes it contains one document in various languages, which means language versions are not taken into account in the total above.

4.2.7.3.3. TOAD meta-information

Only one vocabulary was investigated in the TOAD – the vocabulary of 'Commissions'[118], containing 41 entries. Some of them are also described with the mandate period specification. It is very similar to the one described in the chapter 4.2.7.2.3 and also used for the specification of Dossiers in the search form.

The Following meta-information were found to be in in the TOAD:

- Dossier title
- Dossier (folder) number
- Document title
- Rapporteur
- Opinion number
- COM number
- Procedure
- Adoption date

4.2.7.3.4. End-user search possibilities

The TOAD search is quite hard to use, at least for the common end-user, because of its poor performance (response time). The empty search (adoption date > 1 January 1994) of the 1 200 documents produced a list of results only after ca 4 minutes.

4.2.7.3.5. Re-use of the TOAD in view of integrated access solution

The overall concept and performance of the TOAD do not bring a lot of possibilities of its re-use in any integrated access solution. Significant technical changes would need to be carried out to re-use documents from TOAD.

4.2.7.4. Committee of the Regions Studies/Brochures register

This source allows the possibility of browsing through CoR Studies and Brochures as the main public CoR outputs – well-formatted documents in PDF format, some of them equipped with an ISBN code.

Only a **general analysis** of the CoR Studies/Brochures register was carried out.

4.2.7.4.1. Public access

The CoR Studies and Brochures register is accessible at this general URL address:
http://cor.europa.eu/en/documentation/Pages/index.aspx.

4.2.7.4.2. Overall volume of published documents

There are ca 100 documents published in this register, ca 50 studies and ca 50 brochures.

[118] TOAD vocabulary of Commissions (used also for Dossiers) in the study database:
http://atom.ts-publicaccess.eu/form/item?ItemId=1061546.

4.2.7.4.3. CoR Studies/Brochures register meta-information

Only one vocabulary was investigated in the TOAD – vocabulary of CoR Policies[119] (also called themes) containing 8 entries.

The following meta-information was discovered elsewhere in the TOAD:

- Document title
- Authority
- Keywords (this is in fact a vocabulary, but it is not published in any way)
- Available languages/translations
- Publication date

4.2.7.4.4. End-user search possibilities

The CoR Studies/Brochures does not provide any kind of search. Only filtering by policy and keyword is available in the register.

4.2.7.4.5. Re-use of the CoR Studies/Brochures register for future integrated access solution

The content of these documents suggests that these documents should be further used. However, from the technical point of view, there are no possibilities of re-use. Technical changes need to be implemented in order to provide access to these documents.

4.2.7.5. Committee of the Regions General website search

As already mentioned in the beginning of the analysis, the CoR website is a comprehensive source of various information and documents.

Only a **general analysis** of the CoR General website search was carried out.

4.2.7.5.1. Public access

The CoR General website search is accessible at this general URL address:
http://cor.europa.eu/en/search-center.

There are some special functionalities for:

- Opinions Search:
  http://cor.europa.eu/en/activities/opinions/pages/Opinions-Search.aspx, where
  - The results (Opinions) are also published in the OJ.
  - There is an overlap (document duplicates) with TOAD and CoR-DM.
- Documents Search (CoR-DS)
  http://cor.europa.eu/en/search-center/pages/search-documents.aspx

4.2.7.5.2. Overall volume of documents in the CoR-DS

The total number of documents has been found by submitting several inquiries into the above search tool, and it can be estimated at circa 20 000.

4.2.7.5.3. Document types

No types are explicitly specified. The following examples of document types were found empirically:

- Brochure
- Event

[119] CoR Studies/Brochures register vocabulary of Policy areas in the study database:
http://atom.ts-publicaccess.eu/form/item?ItemId=1052236.

- Speech
- Presentation
- Opinion
- Agenda, etc.

4.2.7.5.4. CoR-DS meta-information

Only one vocabulary was investigated – the vocabulary of 'Keywords'. This vocabulary is managed by the users, so it is uncontrolled. It is not published in any way, only as a filtering facet in the list of results, so it was not investigated further.

Metadata attributes available:

- Document title
- Machine generated abstract
- Author
- Date
- Format
- File size

4.2.7.5.5. Document formats

- PDF
- DOC/DOCX

4.2.7.5.6. End-user search possibilities evaluation

Only the simple search is available. After generating a list of results, there are facet filtering possibilities usable for narrowing the volume of the list of results by format, date, and keyword.

One disadvantage is the absence of a full text search of document content as well as the performance of the search engine. It takes more than five seconds to generate a list of results.

4.2.7.5.7. Re-use of the CoR-DS in view of integrated access solution

From the content point of view, these documents should be further used. However, from the technical point of view, there are no possibilities of re-use. Technical changes need to be implemented in order to provide access to these documents.

4.2.8. European Economic and Social Committee

The European Economic and Social Committee ('EESC') is a consultative body with 350 members from EU Member countries belonging to three groups: Employers, Workers and Various Interests.

The EESC opinions are forwarded to the CEU, the EC, and the EP. EESC thus has a key role to play in the Union's decision-making process.[120]

EESC has a very complex website full of information on its members, press releases, themes of interests, events, activities etc. which is accessible at this general URL address: http://www.eesc.europa.eu/.

An in-depth analysis of EESC Register of documents was carried out as part of the study.

4.2.8.1. EESC Register of documents

4.2.8.1.1. General information

The EESC Register of documents ('RD-EESC') serves as the main access point to

1. Opinions of the EESC;
2. Public documents produced by EESC;
3. COM documents (documents originating from the EC).

The RD-EESC has two areas depending on the access rights: the Public area and the Members' area. Only the Public area is subject to study.

It is necessary to add that that RD-EESC uses the same structure and the same techniques and technology as the CoR Documents Manager (CoR-DM).

The **thorough analysis** of the RD-EESC document source was carried out.

4.2.8.1.1.1. Public access

RD-EESC is accessible from several places of the EESC website and navigation to the search is intuitive.

All three main parts of RD-EESC have their URL's based on Document management system of the EESC (https://dm.eesc.europa.eu), varying with the prefix:

1. EESC Opinions: https://dm.eesc.europa.eu/EESCDocumentSearch/Pages/opinionssearch.aspx,
2. Public Documents: https://dm.eesc.europa.eu/EESCDocumentSearch/Pages/redsearch.aspx,
3. COM Documents: https://dm.eesc.europa.eu/EESCDocumentSearch/Pages/comsearch.aspx

All parts are accessible free of charge; the 'Sign in' option is available for registered users, but is not necessary for searches.

4.2.8.1.1.2. Time range covered

The oldest documents available date back to the year 1991. The RD-EESC does not provide a search according to the individual years. It filters the results in groups of years, for example 'Less than 1994', '1994 up to 2000' etc.

4.2.8.1.1.3. Overall volume of documents

The volume of published documents up to the end of February 2016 is shown in Table 10:

[120] More about the EESC: http://www.eesc.europa.eu/?i=portal.en.about-the-committee.

RD-EESC section	Documents in English	Documents in all languages
Opinions	5 540	80 609
Public Documents (RED)	37 271	446 203
COM documents	9 585	122 907

Table 10: Volume of published documents in RD-EESC

The section COM documents has an overlap of 1:1 with the RD-EC. This means that all documents included in this section are also published in the RD-EC.

In the section Public Documents (RED) there is a slight overlap (26 documents) with the CoR-DM register of documents.

All three sections have an almost identical search functionality and so will not be further differentiated.

4.2.8.1.1.4. Brief general investigation of the RD-EESC

Figure 28 shows the result of comparative analysis of the RD-EESC.[121]

Figure 28: Overview investigation of the RD-EESC

[121] Brief investigation of the RD-EESC in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000029.

4.2.8.1.2. Document types

It is best to divide the vocabulary[122] containing document types according to the type of the search mentioned above.

The EESC Opinions section works with 2 document types:

- Opinion
- Resolution

The Public Documents section works with 20 document types. 10 from them are also used in the CoR-DM.

- Amendment - Specialized section
- Amendment to the plenary session
- Annual Report
- Committee opinion adopted in plenary session
- Consultative work, various
- Draft information report
- Draft opinion
- Draft opinion of the Committee
- Information report
- Invitation, proposal of programme negotiations
- Minutes from the meeting
- Minutes of the meeting
- Opinion of the Specialized section
- Preliminary draft information report
- Preliminary draft opinion
- Press Release
- Report on the own-project opinion
- The report to the plenary session
- The report to the request for new opinion
- Working document

The COM Documents section works with 5 Document Types based on the way in which COM documents are generally categorised:

- COM
- SEC
- SWD
- C
- JOIN

4.2.8.1.3. Metadata as relationships between documents and vocabularies

4.2.8.1.3.1. Document originator vocabularies

Document originator vocabularies are used to specify the origin and responsibility for the documents in the EESCED. Two separate vocabularies serve this purpose:

1. Document source
2. Dossier
3. Rapporteur

4.2.8.1.3.1.1. Document Source

The vocabulary 'Document Source' describes the institution from which the documents originate. It has 3 entries:

- EESC
- CoR-EESC (only applicable in the section 'Public documents')
- CoR

[122] RD-EESC vocabulary of Document types in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1065398.

- COM (applicable only in the section 'European Commission Documents').

4.2.8.1.3.1.2. Vocabulary of Dossiers

The vocabulary of Dossiers[123] represents the EESC section in which the opinion has been adopted and is organised as a flat list.

The findings according to the facet filter items in the list of results show that:

- Dossier is further specified by the attribute number.
- 10 dossiers are active while another 10 are historical.

The relationship between a document and entries in the vocabulary of Dossiers has 1:0…1 cardinality. This means that the document can be attached to either none or one Section (Dossier).

4.2.8.1.3.1.3. Vocabulary of Rapporteurs

The vocabulary of Rapporteurs[124] is a list of Members of EESC designated as the rapporteur of the EESC opinion and organised as a flat list of 565 entries.

The investigation shows that:

- Rapporteurs are stated here only by their family names
- There is not any reference to the database of EESC members database[125]

The relationship between the document and the entries in the vocabulary of Rapporteurs has 1:0…1 cardinality. This means that the document can be attached to either none or one rapporteur.

4.2.8.1.3.2. Time range vocabularies

Time specifications are handled in the RD-EESC by three independent vocabularies:

1. Years
2. Plenary sessions
3. Months in Years

4.2.8.1.3.2.1. Vocabulary of Years

The vocabulary of Years brings the basic orientation in time to the user searching for the document in the RD-EESC.

It was established that vocabulary is prepared as an extract from the document dates.

The relationship between a document and entries in the vocabulary of Years has 1:1 cardinality. This means that each document must have exactly one entry associated with it in the vocabulary of Years.

4.2.8.1.3.2.2. Vocabulary of Plenary sessions

The vocabulary of Plenary sessions[126] determines the event in which the document was adopted.

---

[123] RD-EESC vocabulary of Dossiers in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1133003.

[124] RD-EESC vocabulary of Rapporteurs (Members of EESC) in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1133087.

[125] List of EESC members at the source: http://memberspage.eesc.europa.eu/.

[126] RD-EESC vocabulary of Plenary sessions in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1134213.

It was found that:

- It is used only in the Opinions section of the RD-EESC.
- Has been used since 2000.
- Entries in the vocabulary are composed of the number of the Plenary session.

The relationship between an opinion document and entries in the vocabulary of Plenary sessions has 1:0…1 cardinality. This means that each opinion document can belong to either none or one entry in the vocabulary of Plenary sessions.

4.2.8.1.3.2.3. Vocabulary of Months in Years

The vocabulary of Months in Years is used for two purposes:

1. It determines the month in the year in which the document was adopted.
2. It determines the month in which the meeting was held.

It was found that it is an extract from the relevant dates.

The relationship between a document and entries in the vocabulary of Plenary sessions has 1:0…1 cardinality. This means that each document can belong to either none or one entry in the vocabulary of Months in Years which is either:

- The Adoption date or,
- The Meeting date

4.2.8.1.3.3. Vocabulary of Languages

The last vocabulary to be used intensively in the RD-EESC is the vocabulary of Languages.[127] It is used to identify both:

- The original language of the document;
- The additional languages into which the document has been translated.

as well as to provide the document in the requested language version to the user.

The relationship between a document and entries in the vocabulary of Languages has 1:1…1 cardinality. This means that each document must belong to at least one entry in the vocabulary of Languages.

4.2.8.1.4. Metadata as document attributes

The following document attributes were investigated, which are mandatory for each document in the RD-EESC:

- Dossier
  - Number in Dossier
- Document
  - Code (composed from the codes of vocabulary entries and the document attributes)
  - Number of document
  - Status of document
  - Document format (used also as a standalone facet in the lists of results)
  - Date of document

[127] RD-EESC vocabulary of Document languages in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1134509.

- Meeting date
- Adoption date
- Production date
- Requesting service
- Confidentiality
- Version status
- Document version
- Document part
- Responsible administrator

4.2.8.1.5. Relations between documents

4.2.8.1.5.1. Internal relations

One internal relationship was found between document and dossier. It is used only selectively and the usage depends on the Document type.

4.2.8.1.5.2. External relations

Some documents are related to the procedure number. However, from the user point of view it is not clear if it is a relationship or an information on an attribute.

4.2.8.1.6. End-user search possibilities

This section analyses the search options that are available to the end-user in the RD-EESC, and it covers three main areas in line with section 4.1, i.e. the search form, the list of results, and the document detail with a brief conclusion.

4.2.8.1.6.1. Search form

The search form in the RD-EESC is not split into simple search and advanced search forms but provides one common search form with basic search criteria.

The dominant part of the search is free text search (Google-like), but auto-complete is not available.

The search is based on four additional options:

- **Plenary Session** – there is the possibility to base the search on the number of the plenary session, which can be selected from a predefined list of values; moreover, another free text search is available for searching among numbers (and titles) of plenary sessions.
- **Section** – the search can be based on the section such as External Relations, Single Market, Production and Consumption, etc. there is also a free text search available.
- **Rapporteur** – the search can be limited to particular rapporteur of the EESC (see 4.2.8.1.3.1.3); the list of rapporteurs is available as well as additional free text search.
- **Language** – the user can limit their search according to the language of the document.

4.2.8.1.6.2. List of results

Results are by **default** displayed as titles with several additional pieces of information which are visible directly after the search. Each result is accompanied by information on the rapporteur, original language, date of the document, date of the meeting, the administrator responsible, type of the document, date of adoption, date of 'production' and the internal EESC number.

The list of results can be modified by several options. The list can be switched from default view described above to the '**Table view**'. Table view provides a very comprehensive view in rows and columns and covers all the information from the default view. But there are additional options to the default view provided by this Table view:

- Download button – allows downloading the opinion in a DOC file;
- Share button – this button automatically generates a stable URL link to the document;
- Additional button – once again the very same Download function and Share function are available, but accompanied by 'Rank' information.

The 'Rank' of the document provides information about the relevance of the document to the search in the list of results, however, it is not very clear for the average user, as the Search manual is available only after log in. The 'Rank' is based on a 4-digit number and in the tooltip it shows the rank as a percentage of how well the document fits predefined subject matters (such as 'civil society', 'human rights', etc.). Unfortunately, the complex list of these subject matters is not available and it is not possible to search for them accordingly unless one uses their names in the full-text search. Without knowing the complex list this is an extremely difficult task.

The last viewing option is the '**Compact View**', which displays the Name of the document (based on EESC number), Actions (Share, Download, Rank), Title of the document, Date of procedure, Rank.

The results can be sorted by:

- Relevance (based on the Rank number in descending order)
- Date – newest vs. oldest

The list of results can be further filtered by facets on the left side, so the user can easily filter their search by:

- Document Source: only 'Any Document Source' or 'EESC' could be chosen
- Document Year: years are grouped into periods such as 'Less than 1994', '1994 up to 2000'. After clicking on one period it automatically changes to a more detailed time period. For example, clicking on 'Less than 1994' automatically changes other options and new option as 'Less than 1991', '1991 up to 1992' is visible. This way, The user can go deeper and deeper into time periods, but has no option to choose one particular year at the beginning.
- Document Language: the user can limit his search to particular language version of results
- Document Type: see 4.2.8.1.2
- Result Type: the user can limit their search according to the type of the document (Microsoft Word, WordPerfect, RichText).
- Dossier: the user can limit their search according to the dossier type. However, only codes of dossiers (as INT, NAT, TEN) are available without any explanation.
- Adoption Date
- Meeting Date
- Rapporteur
- Modified Date: this filter provides an option to filter the search according to the time period of last changes, where predefined values Any Modified Date, Past Month, Past Six Months, Past Year, and Earlier are available.

4.2.8.1.6.3. Document detail

There are no options to make just one document visible. This information is visible only from the list of results and is described in the previous chapter (4.2.8.1.6.2).

4.2.8.1.6.4. General evaluation of the Search functionality

The search functionality works well. However, the user interface is not targeted to end-users outside of the EESC and it is not easy to get familiar with it because it requires a lot of experience to use. A

user manual for the Search functionality exists, but it requires logging into the internal system available only for members.

4.2.8.1.7. Sample document

To see documents from different document sources through the same lens, a randomly selected document from RD-EESC was re-created in the unified structure of the study database. This could help to design the future integrated access solution or help to understand the treatment of metadata from different document registers at one location. The sample document from RD-EESC is shown in Figure 29 and is directly accessible in the study database.[128]

Figure 29: Sample document from the EESC Register of documents

[128] Sample document from the RD-EESC in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1134536.

4.2.8.1.8. Re-use of the RD-EESC in view of integrated access solution

The RD-EESC covers most of the relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT **(is the document about)**	Document type Topics not available
WHEN **(was the document published)**	Dates
WHO **(is responsible for the document)**	Authorities
WHY **(purpose of the document)**	Procedure, Other metadata

The RD-EESC is equipped with many metadata. However, the vocabularies are not publicly available and they are not described at all (e.g. what exactly the document types represent). A major deficiency is the absence of any thematic information about the documents.

The RD-EESC does not provide API or any other web service for automatic retrieval of the content; it only provides RSS functionality, which could be considered a source for documents retrieval for a possible integrated access solution.

4.2.9. European Ombudsman

The European Ombudsman ('EO') is an independent and impartial body that holds the EU administration to account. The EO investigates complaints about maladministration in EU institutions, bodies, offices, and agencies. Only the CJEU, acting in its judicial capacity, falls outside the EO's mandate. The EO may find maladministration if an institution fails to respect fundamental rights, legal rules or principles, or the principles of good administration.[129]

4.2.9.1. General information about the EO website

The EO office provides information on its website about its organisation, activities, cases, and documents. The EO website is accessible at this general URL address: www.ombudsman.europa.eu.

The following document sources were found on the EO website and analysed separately:

1. EO Register of Cases
2. EO Register of Resources

There are several places in the EO website where other documents, such as press releases, can be found.

The published documents are accessible by the general public without any special authentication. Documents in the Register of Resources are partially available only upon request.

4.2.9.2. EO Register of Cases

4.2.9.2.1. EO Register of Cases general information

The EO Register of Cases ('EO-RC') provides, primarily through its search interface, a complex access of the documents relevant to the cases processed throughout the whole history of the EO.

Documents are displayed in a user-friendly way in this register – documents are clearly readable in the web browser (only a PDF copy is attached). Each document has its own navigation tools which allow easy orientation within the document and documents are also linked together by semantic context, through which the entire genesis of the case may be observed.

A **thorough analysis** of the EO-RC document source was carried out.

4.2.9.2.1.1. Public access

The EO-RC is accessible at this general URL address:
http://www.ombudsman.europa.eu/cases/advancedsearch.faces.

4.2.9.2.1.2. Overall volume of published documents

The volume of documents published in the EO-RC up to the end of February 2016 is 5 237.

4.2.9.2.1.3. Time range covered

The EO-RC covers cases documents since 1995, from the beginning of the EO operation.

[129] Source of the Description: http://www.ombudsman.europa.eu/en/atyourservice/whocanhelpyou.faces#/page/3.

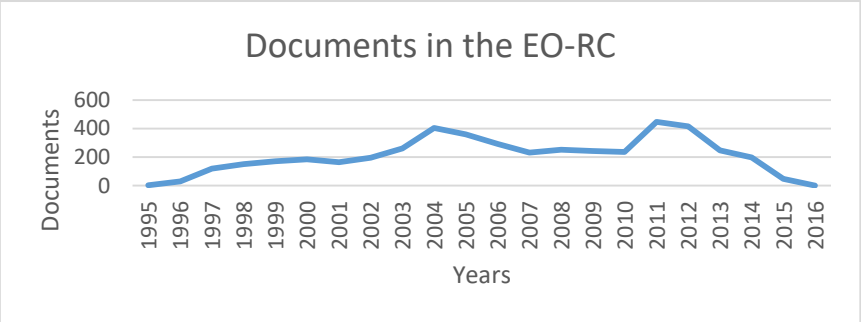


Figure 30: Volume of published documents published in the EO-RC by years

4.2.9.2.1.4. Brief investigation of the EO-RC

Figure 31 shows the results of the comparative analysis of the EO-RC.[130]

- INFORMATION REPOSITORY
  - Provision of documents
    - **Documents available via the special search page**
- ACCESSIBILITY OF DOCUMENTS
  - Way of accessibility
    - **Documents available to general public**
  - Machine readability methods
    - **RSS**
  - Social networks used for broadcasting information
    - **YouTube**
    - **Twitter**
    - **LinkedIn**
    - **Google+**
  - Amount of documents
    - 1 000 - 9 999
      - **< 6 000**
- AVAILABILITY OF DOCUMENTS
  - Document types
    - More document types available
      - **< 10**
  - Metadata capturing methods
    - Vocabulary groups
      - **Document types**
      - **Topic**
      - **Languages**
      - **Process information**
    - Attribute groups
      - **Various codes**
    - Relations
      - Internal relations
        - **Related content within the website**
- Languages
  - User interface languages
    - **All EU languages**
  - Document languages
    - **Only certain languages**
- Formats
  - **PDF**
  - **webpage**
- END-USER SEARCH POSSIBILITIES
  - **Empty search possible**
  - **Document content indexed**
  - Search form
    - **Advanced search form**
  - List of results
    - Facet filtering of the results available
      - **by document type**
      - **by date/year**
      - **by topic/tag**
      - **by language**
  - Document detail
    - **Content + metadata in a (set of) webpage(s)**
- OTHER RELEVANT ASPECTS
  - Technology (CMS and/or Search engine)
    - **Not identifiable**
- COMMON END-USER EXPERIENCE
  - **5 Excellent (complex, fast, intuitive)**

Figure 31: Overview investigation of the European Ombudsman Register of Cases

[130] Brief investigation of the EO-RC in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000033.

4.2.9.2.2. EO-RC Document types

The Document types in the EO-RC [131] are represented by a vocabulary containing a list of 8 Document types.

Document type	Count of documents
Cases opened	607
Closing summaries	647
Decisions	3 171
Recommendations	252
Special reports	19
Case descriptions	0
Correspondences	520
Friendly Solutions	21

Table 11: EO-RC Document types

Documents are bundled into cases which (in this sense) serve as dossiers for documents.

The EO-RC allows a special view where the documents can be seen by years and months.

The relationship between a document and entries in the vocabulary of EO-RC Document types has 1:1 cardinality. This means that the document must be of exactly one document type.

4.2.9.2.3. Metadata as relationships between documents and vocabularies

No specific Document originator vocabulary was found in the EO-RC.

4.2.9.2.3.1. Thematic vocabularies

There are more thematic vocabularies in the EO-RC which describe characteristics of the document from various points of view:

- Institutions concerned[132]: 68 entries representing EU institutions and agencies whose operations are or could be the object of complaints or information requests.
- The Keyword section, with three vocabularies:
  1. Field of Law[133]: 20 entries generally describing areas of EU operation (from Agriculture to Transport.)
  2. Types of maladministration alleged[134]: 21 entries (e.g. Duty of care or Right to be heard.)
  3. Subject matter[135]: 7 entries describing the subject matter of cases (e.g. Tendering or Contracts execution.)

[131] EO-RC vocabulary of Document types in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1064486.
[132] EO-RC vocabulary of Institutions concerned in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1064866.
[133] EO-RC vocabulary of Field of Law in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1064864.
[134] EO-RC vocabulary of Types of maladministration alleged in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1064874.
[135] EO-RC vocabulary of Subject matter in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1064870.

- Type of settlement[136]: 13 entries describing the result of the case (e.g. Dealt by a Court or No maladministration found), only used in the Document type Decision and Summary.
- Recommendation result[137]: 13 entries used in Document type Recommendation describing the result (e.g. if the Recommendation was accepted by the institution or not.)

The relationship between the documents and the thematic entries has 1:0…n cardinality. This means that no theme or topic is obligatory for documents.

4.2.9.2.3.2. Time range vocabulary – vocabulary of Years

The vocabulary of Years[138] allows the user to position their search in time when searching for a document in the EO-RC.

The relationship between the documents and entries in the vocabulary of Years should have 1:1 cardinality. Nevertheless, some 20% of the documents in the EO-RC do not have this relationship.

4.2.9.2.3.3. Vocabulary of Languages

The last vocabulary intensively used in the EO-RC found is the vocabulary of Languages.[139] It is used to:

- Identify the original language of the document;
- Display the other languages into which the document is translated;
- Provide the document in the language version requested by the user.

The relationship between the documents and entries in the vocabulary of Languages has 1:1…n cardinality. This means that each document must belong to at least one entry in the vocabulary of Languages.

4.2.9.2.4. Metadata as document attributes

The following document attributes were found. They are mandatory for each document in the EO-RC:

- Case Code
- Dates - information about the date a case was opened break down by the Document type

4.2.9.2.5. Relations between documents

4.2.9.2.5.1. Internal relations

There is quite a rich array of internal relations between documents.

Documents are bundled into cases and each case contains a set of documents.

Each document also contains a link to the other relevant documents within the case.

4.2.9.2.5.2. External relations

No external relations linking documents or cases to documents in another register were found.

---

[136] EO-RC vocabulary of Type of settlement in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1064872.

[137] EO-RC vocabulary of Recommendation result in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1064868.

[138] EO-RC vocabulary of Years in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1065373.

[139] EO-RC vocabulary of Document languages in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1064837.

4.2.9.2.6. End-user search possibilities

This section analyses the search options that are available to the end-user in the EO-RC. It covers three main areas in line with section 4.1, i.e. the search form, the list of results and the document detail along with a brief conclusion.

4.2.9.2.6.1. Search form

There is an advanced search form available for investigating cases and documents included in cases.

This advanced search form is exhaustive and allows filtering using all vocabularies specified above (see 4.2.9.2.3) and all metadata attributes (see 4.2.9.2.4):

- Filtering of text contents by the words or phrases
- Language
- Case number
- Document type (multi-select)
- Dates (from-to)
- Institutions concerned
- Field of Law, Type of maladministration alleged, Subject Matter
- Type of settlement
- Recommendation result

Use of more than one criterion is processed as a chain linked by the logical AND operator. This means that the more criteria are used, the fewer results are produced.

4.2.9.2.6.2. List of results

This consists of links to the documents organised by Document type and organised into groups with five results from each Document type. Each document is accompanied by its document date.

4.2.9.2.6.3. Document detail

The document detail is very user-friendly. The content of the document is directly displayed as well as the navigation over the chapters or sections. Some documents also have the possibility of downloading a PDF copy.

4.2.9.2.6.4. General evaluation of the Search functionality

The search functionality works very well. It is also very intuitive for the end-users.

4.2.9.2.7. Sample document

To see documents from different document sources through the same lens, a randomly selected document from EO-RC was re-created in the unified structure of the study database. This could help to design the future integrated access solution or help to understand the treatment of metadata from different document registers at one location. The sample document from the EO-RC is shown in Figure 32 and is directly accessible in the study database.[140]

Figure 32: Sample document from the European Ombudsman Register of Cases

[140] Sample document from the EO-RC in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1065257.

4.2.9.2.8. Re-use of the EO-RC in view of integrated access solution

The EO-RC covers most of the relevant metadata necessary for answering basic PublicAcess.eu Project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT **(is the document about)**	Document type Topics
WHEN **(was the document published)**	Dates
WHO **(is responsible for the document)**	Not available
WHY **(purpose of the document)**	Procedure, Other metadata

Although EO-RC works well, it is quite difficult to imagine that it could be used in any of the integrated access solution alternatives. There are two main reasons for this:

1. Absence of any machine readable format (API etc.)
2. The EO-RC database is specifically adapted to support EO processes and so it would be hard to generalise.

EO-RC documents could be regarded as important from a public interest point of view, however, serious technical difficulties are foreseen for its re-use in any integrated access solution.

4.2.9.3. European Ombudsman Register of Resources

4.2.9.3.1. European Ombudsman Register of Resources general information

The European Ombudsman Register of Resources ('EO-RR') provides, primarily through the search interface but also through some prepared views, a complex overview of the documents (such as correspondence or contracts) produced by the EO office in addition to the cases.

There are only a small number of documents in the EO-RR and for this reason the analysis was shortened.

4.2.9.3.1.1. Public access

The EO-RR is accessible at this general URL address:
http://www.ombudsman.europa.eu/en/resources/pr/publicregistersearchpage.faces.

Some of the documents are available directly from the register while others only on request.

A **general analysis** of the EO Register of Resources document source was carried out.

4.2.9.3.1.2. Overall volume of published documents

The volume of published documents published in the EO-RR up to the end of February 2016 is 667.

4.2.9.3.1.3. Time range covered

The EO-RR covers case documents produced in the period 2012 – 2015. There are no documents from 2016.

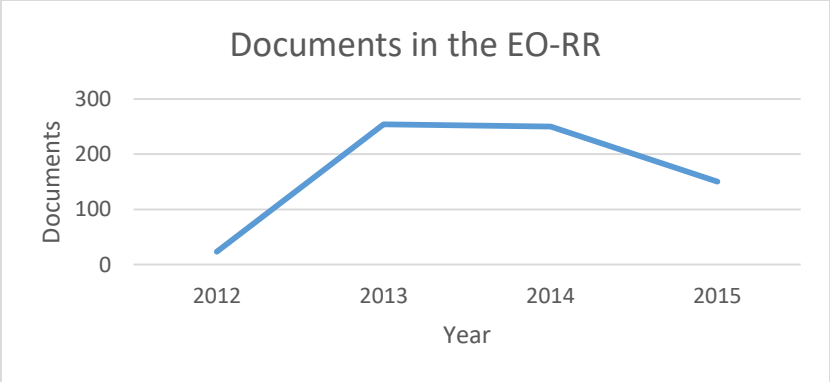


Figure 33: Volume of published documents published in the EO-RR by years

4.2.9.3.2. EO-RC Document types

The Document types in the EO-RR[141] are represented by a vocabulary containing a list of 24 Document types. Of the 667 documents the most predominant document types are: Letters (454 examples), Notes (53 examples) and Decisions (40 examples) while the other Document types are used very rarely.

The relationship between a document and entries in the vocabulary of EO-RR Document types has 1:1 cardinality. This means that the document must be of exactly one document type.

[141] EO-RR vocabulary of Document types in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1065273.

4.2.9.3.3. Metadata as relationships between documents and vocabularies

4.2.9.3.3.1. Document originator vocabulary – Business unit

The vocabulary of Business units[142] is used to specify the origin and responsibility for the documents in the EO-RR. There are 6 entries in this directory.

The relationship between a document and entries in the vocabulary of Business units has 1:1 cardinality. This means that this meta-information is obligatory in the EO-RR _Vocabularies

4.2.9.3.3.2. Thematic vocabulary - Classification

The vocabulary Classification[143] categorises the topics within the documents.

The investigation revealed that it contains a 4-level taxonomy of 324 entries. Most of them do not have any documents attached.

The relationship between the documents and the thematic entries has 1:1 cardinality. This means that Classification is an obligatory piece of meta-information for each document.

4.2.9.3.3.3. Time range vocabulary – vocabulary of Years

The vocabulary of Years[144] allows basic positioning in time to the user searching for a document in the EO-RC.

The relationship between a document and entries in the vocabulary of Years has 1:1 cardinality. Therefore, the relation is in the EO-RR obligatory.

4.2.9.3.4. Metadata as document attributes

The following document attributes were investigated in the EO-RR:

- Reference
- Content description metadata
  - Subject
  - Description
- Date specifications
  - Document date
  - Registration date
  - Publication date

4.2.9.3.5. Relations between documents

4.2.9.3.5.1. Internal relations

Related documents are interconnected.

4.2.9.3.5.2. External relations

No external relations linking documents or cases to documents in other registers were investigated.

[142] EO-RR vocabulary of Business units in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1065330.
[143] EO-RR taxonomy of Classification in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1061632.
[144] EO-RR vocabulary of Years in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1065223.

4.2.9.3.6. End-user search possibilities

This section analyses the search options that are available to the end-user in the EO-RR and it covers three main areas in line with section 4.1 - the search form, the list of results, and the document detail along with a brief conclusion.

4.2.9.3.6.1. Search form

The main method of searching is a simple search which allows document retrieval filtered by words in their Subject and/or description.

The simple search form is easily expanded to the advanced one, which provides following filtering possibilities:

- Reference
- Full text search in the content of the document
- Dates (from-to)

The use of more than one criterion is handled as a chain linked with the logical AND operator. This means that the more criteria are used, the fewer results are produced.

4.2.9.3.6.2. List of results

The list of results consists of links to the documents with additional meta-information. Each result is equipped with the following meta-information:

- Document date
- Reference number
- Classification
- Type
- Language

Results can be filtered by the following facet filters:

- Year
- Type
- Business Unit
- Medium

4.2.9.3.6.3. Document detail

Document detail provides a link to the document and its annexes (if any) in addition to the following metadata:

- Reference
- Subject
- Description
- Document type
- Document date
- Registration date
- Publication date
- Classification
- Responsible unit
- Medium
- Online disclosure

4.2.9.3.6.4. General evaluation of the Search functionality

The search functionality works very well. It is also very intuitive for the end-users.

The main problem is the small number of documents. This leads one to believe that only a fraction of EO documents are published on the EO-RR.

4.2.9.3.7. Re-use of the EO-RR in view of integrated solution

The EO-RR covers relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT (is the document about)	Document type; Topics
WHEN (was the document published)	Dates
WHO (is responsible for the document)	Not available
WHY (purpose of the document)	References; Other metadata

It is quite difficult to imagine that it could be used in any of the integrated access solution alternatives. There are two main reasons for this conclusion:

1. An absence of any machine readable format
2. An incomplete document set

4.2.10. EUR-Lex

4.2.10.1. General information

The ontology of EUR-Lex data uses the model 'Functional Requirements for Bibliographic Records' ('FRBR')[145]. FRBR defines several levels of abstraction for each publication (WEMI model):

- **W**ork
- **E**xpression
- **M**anifestation
- **I**tem

In the case of EUR-Lex data model, it can be stated that:

- Work defines the document;
- Expression defines the language in which the document is written;
- Manifestation specifies the format in which it is presented (PDF, XML[146], DOC, HTML, print);
- Item defines a particular file/stream (xy.doc, xy.html, xy.pdf).

Each level has its own specific metadata. Although both Expression and Manifestation are also significant from the perspective of the Study, the most important level in terms of metadata is Work, which constitutes the majority of all analysed metadata.

The data store from which the information displayed on EUR-Lex is drawn is the Common repository for content and metadata ('CELLAR'). The data stored in the CELLAR database can be accessed directly via a SPARQL[147] endpoint published at: http://publications.europa.eu/webapi/rdf/sparql or indirectly via a web service of EUR-Lex.[148]

The ontology of the CELLAR database is called the Common Data Model ('CDM') and it is accessible in the OWL form[149] via the SPARQL endpoint or at the Metadata registry ('MDR').[150]

Outside the WEMI model the CDM ontology includes several more abstractions:

- Agent (a resource with 'power to act': Institution, Country, Person, Organisation, etc.)
- Administrative unit
- Concept (used for classification, a 'concept' is usually encoded within a vocabulary, taxonomy or thesaurus)
- Language
- Dossier (used for grouping documents, for instance a legislative or non-legislative procedure)

[145] Functional Requirements for Bibliographic Records website: http://www.ifla.org/frbr-rg.
[146] XML format used by EUR-Lex is Formex. It allows fragmentation of the content of the document (see http://formex.publications.europa.eu/formex-4/manual/manual.htm#ARTICLE or http://formex.publications.europa.eu/formex-4/manual/manual.htm#PARAG element definitions).
[147] W3C Recommendation: SPARQL 1.1 Query Language: https://www.w3.org/TR/sparql11-query/.
[148] Web service of the EUR-Lex (available after login): http://eur-lex.europa.eu/protected/web-service-registration.html.
[149] W3C: Web Ontology Language: https://www.w3.org/2001/sw/wiki/OWL.
[150] MDR website homepage: http://publications.europa.eu/mdr/index.html.

Alongside the ontology definition, the MDR website also contains definitions for controlled vocabularies, taxonomies, and thesauri used by CELLAR (EUR-Lex). Publications of Named Authority Lists ('NAL's') and Atelier for Translation Tables in the Office ('ATTO') can be retrieved through the MDR website[151] and are also made available on the ODP.[152]

Besides documents, EUR-Lex also publishes 'Legislative procedures', which follows the life cycle of a legislative procedure from the moment it is launched until the final law is adopted. Legislative procedures have their own metadata. The other document sources (institutions) may refer to the same procedure within their register of documents (for instance through an attribute in the relevant document).

4.2.10.1.1. Interinstitutional Metadata Maintenance Committee

For the purposes of the study, it is important to be aware of the Interinstitutional Metadata Maintenance Committee ('IMMC'), whose objective is to ensure metadata governance at the interinstitutional level. One of the IMMC priorities is to ensure harmonization of the metadata used between the institutions. To avoid possible confusion, it is necessary to mention that only part of all the metadata used in EUR-Lex data belongs to the responsibility of the IMMC as only some metadata are used by all involved institutions (outside the Publications Office). The metadata in scope can be divided into two groups:

- Core metadata set which is mandatory for all the institutions;
- Extension metadata which can make some additional metadata mandatory for a specific institution or exchange.

The IMMC defines specific means for transmissions of exchanged metadata. Documentation for the Inter-institutional Transmission Protocol, Inter-institutional Publication Format and Inter-institutional Transmission Format is available on http://publications.europa.eu/mdr/core-metadata/index.html.

The Transmission Protocol uses e-TrustEx[153] for the actual file (metadata & text) exchange.

Exchange metadata is stored within XML files for which a specific XML format is used. Its XSD schema defines the following types of controlled vocabularies for the 'core metadata':

- Resource-type (type of the resource at the 'work' level)
- Corporate body (used for 'agent')
- Country (used for 'agent')
- Role (indicates the role of the 'agent')
- Event (specifies the type of a given event within a procedure)
- Interinstitutional procedure (type of the procedure, i.e. 'Ordinary legislative procedure')
- Language + Multilingual codes
- Treaty (can define a 'legal basis' for a procedure)
- Format (format of a resource, i.e. 'DOCX', 'HTML')

[151] MDR website – Authorities: http://publications.europa.eu/mdr/authority/index.html.
[152] ODP – Publications Office datasets: http://open-data.europa.eu/en/data/publisher/f85e1721-1db8-4db7-84c4-fc798731ed18?tags=authority+data&page=1.
[153] E-TrustEx is a document exchange platform of the European Commission: http://ec.europa.eu/isa/ready-to-use-solutions/open-e-trustex_en.htm.

Other controlled vocabularies are included in additional 'extension' XSD schemas, which are defined for specific use-cases[154]:

- **OIB:**
  This format describes the metadata exchanges between the Office for Infrastructure and Logistics in Brussels (OIB), the contractor and the Publications Office for use in the project '*Digitisation of the historical archives of the European Commission's Historical Archives Service*'.
- **OP:**
  This format is intended to be used to describe publication requests from the OJ PRINTERS to the Publications Office.
- **LSEU:**
  IMMC transmission protocol rules between the Publications Office and the contractor for '*Summaries of EU legislation*'.
- **LA:**
  This format describes the metadata exchanges between the contractor and the Publications Office for the project '*Indexing and legal analysis of European Union documents*'.
- **GP:**
  This format describes the business level metadata transmitted in the General Publications view of the Publications Office.
- **CDJ:**
  This format describes the publication requests transmitted by the CJEU to the Publications Office.

Currently, the following institutions are involved:[155]

- Publications Office of the European Union
- European Parliament
- Council of the European Union
- European Commission
- Court of Justice of the European Union
- European Court of Auditors
- European Economic and Social Committee
- Committee of the Regions

It is important to note that the current state of metadata exchange between institutions does not solely rely on the IMMC since there are other methods of data exchange. For instance, the metadata concerning the inter-institutional procedure is sent via a database named 'Pegase' managed by the SecGen of the EC[156] and the documents to be published in OJ are sent by the FTP to the Publications Office (planJO service). The transition to full adoption of IMMC is therefore currently ongoing.

---

[154] Definitions taken from XSD schema descriptions in file cm3-20160316-0.zip, available on http://publications.europa.eu/mdr/core-metadata-schema/specific_versions_immc3.html. Additional XSD schemas can be found in the IMMC version 2 available on http://publications.europa.eu/mdr/core-metadata-schema/specific_versions_immc2.html http://publications.europa.eu/mdr/core-metadata-schema/specific_versions_immc2.html

[155] Involved organisational entities: http://publications.europa.eu/mdr/resource/core-metadata/IMMC_reu3_adoption_anx3.pdf_A-2011-764293.pdf.

[156] EC – Secretariat-General website: http://ec.europa.eu/dgs/secretariat_general/index_en.htm.

The current state of the IMMC implementation of abovementioned institutions can be summarized by following table:

Institution	IMMC implementation state
European Parliament	Ongoing preparation
European Council, Council of the European Union	Partial[157]
European Commission (Register of Commission documents)	Partial
Court of Justice of the European Union	Ongoing preparation[158]
European Court of Auditors	Ongoing preparation
European Economic and Social Committee	Ongoing preparation
Committee of the Regions	Ongoing preparation

Table 12: Current IMMC implementation state

Further information about IMMC and its schema can be found at the MDR website.[159]

4.2.10.1.2. Public access

The public has direct access to all published documents in their electronic form. Electronic access is free and no special authentication is necessary.

4.2.10.1.3. Time range covered

Generally speaking, EUR-Lex includes documents roughly from 1951 to present. Some documents are even older (e.g. some documents from Sector 7, see 4.2.10.2.1), but they are of different origin than the EU.

4.2.10.1.4. Overall volume of published documents

The empty search returns a total sum of 945 450[160] documents and the following figure shows their annual incremental growth in the last 20 years.

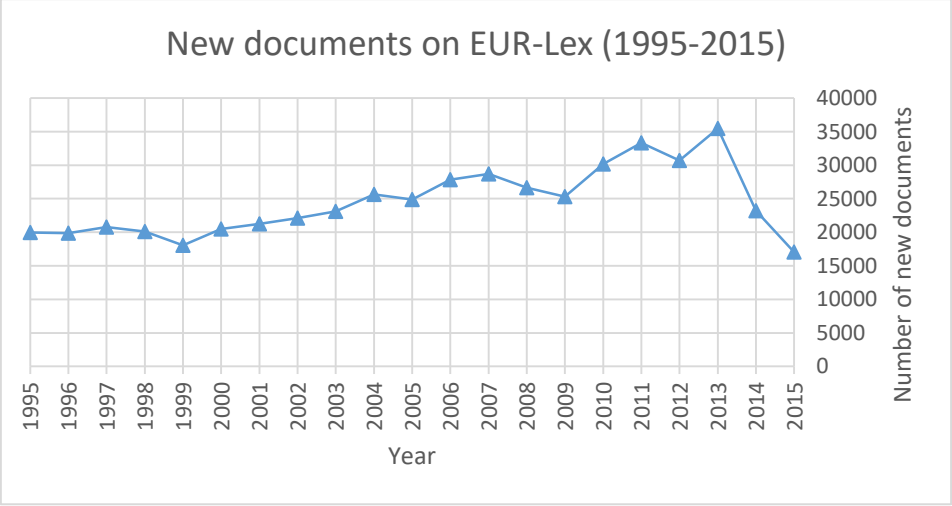


Figure 34: Annual increments of new documents in EUR-Lex (1995 – 2015)

The statistics related to the content in the period 1990 - 2016 can be accessed directly at EUR-Lex here: http://eur-lex.europa.eu/statistics/statistics.html.

[157] Some documents are still transmitted via other methods.

[158] The state of the IMMC implementation of the Court of Justice of the European Union is expected to change in near future (envisioned date is 1st of July 2016). All documents will be transmitted via IMMC with the exception for communications for the publication in the OJ. Therefore, the following state can be classified as "partial".

[159] MDR website: IMMC Core Metadata: http://publications.europa.eu/mdr/core-metadata/index.html.

[160] Based on empty search results 11. 3. 2016.

4.2.10.1.5. Brief Investigation of EUR-Lex

Figure 35 shows the results of the comparative analysis of EUR-Lex.[161]

Figure 35: Overview investigation of EUR-Lex

[161] Brief investigation of the EUR-Lex in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000035.

4.2.10.2.EUR-Lex Document types

More than one vocabulary may be considered Document types:

- Document Type (DTT)
- Form (FM)
- CDM Type

Both Document Type and Form are visible to the user on EUR-Lex. The CDM Type is accessible only through the SPARQL endpoint. Therefore, it can be considered 'internal'. The Form can be considered to provide more specific information than Document Type (DTT) in most cases. The CDM Type is even more specific than the Form. There are some exceptions to such relationships, for example: one type from the Form might be valid for multiple types of Document Type (DTT).

4.2.10.2.1. Document Type (DTT)

Document types (DTT) are listed on EUR-Lex. Now, 162 Document types (DTT) are directly involved in the creation of Celex numbers - a unique identifier for each document. All documents on EUR-Lex are divided into 12 groups called 'sectors'. Each document type belongs exactly to one 'sector'. The codes of document types are unique only in combination with the identifier of the sector (1A vs. 2A). There are 118 document types in all sectors in total.

Sector	Description	Sector	Description
1	Treaties	7	National implementing measures
2	International agreements	8	National Case Law
3	Secondary legislation	9	Parliamentary questions
4	Complementary legislation	0	Consolidated versions of sector 3 documents
5	Preparatory acts	E	EFTA documents
6	Case Law	C	Other documents published in the OJ C

Table 13: List all sectors used in EUR-Lex

4.2.10.2.2. Form (FM)

The Form (FM) or alternatively 'Type of act' is a vocabulary describing the document type. This type currently uses the 'resource-type' authority table[163]. The resource-type has 273 possible values to indicate a document type.

The empty search can be used to extract the information about the use of FM. The following table shows the Top 10 of document types (out of 168 currently in use).

[162] EUR-Lex – Types of documents: http://eur-lex.europa.eu/content/tools/TableOfSectors/types_of_documents_in_eurlex.html.

[163] MDR website – Resource Types table: http://publications.europa.eu/mdr/resource/authority/resource-type/html/resourcetypes-eng.html#description.

No.	Form	Count
1	Written question	173 681
2	Regulation	139 803
3	National implementing measures	138 460
4	Provisional data	71 736
5	Judicial information	40 431
6	Decision	33 365
7	Decision by national courts in the field of EU law	30 591
8	Question at question time	20 873
9	Consolidated text	20 012
10	Communication	19 122
-	Not listed (158 distinct entries)	220 153

Table 14 The Top 10 of document types in the Form vocabulary

4.2.10.2.3. CDM Type

The type defined directly in the CDM ontology can be considered as another document type. The CDM ontology uses the WEMI model, where the first level (Work) is an abstraction of a document regardless of the language version. The type Work is further specified by subclassing to a more specific class in the CDM ontology. This process is recursive, meaning that a subclass can be made more specific by further subclassing. An example might be a 'Treaty' class, which is a subclass of 'Legal resource' class, which is a subclass of a 'Work'. Therefore, the 'Treaty' class is both 'Legal resource' and 'Work' as well by a generic rule of inheritance.

The 'Work' class has 192 subclasses, of which only 11 are direct. The most commonly used subclass of 'Work' is 'Legal resource'. It has 148 subclasses (including indirect ones). The 'Work' hierarchy can be viewed in the study database.[164]

4.2.10.3.Metadata as relationships between documents and vocabularies

EUR-Lex data model uses a high number of vocabularies. Labels from all vocabularies are translated into all EU languages. MDR defines two types of vocabularies:

- ATTO
- NALs

ATTO tables can be considered obsolete vocabularies even though some of their tables are still in use. Unlike for NAL the ATTO tables do not provide any management features since the underlying XMLs contain codes only. There are currently 62 ATTO tables.

Both ATTO and NALs are published and regularly updated on the MDR website. Currently, MDR publishes 50 NAL tables (+ one in preparation). All NALs have been investigated by their 'use.context' property, which indicates the use of individual codified entries by various applications. The graph in Annex 2 shows that not all entries in the individual NALs are used by EUR-Lex. The 'use.context' attribute also indicates that only 15 out of 50 NALs are used within the EUR-Lex metadata.

The vocabularies used in the EUR-Lex documents can be viewed in the study database.[165]

---

[164] 'Work' hierarchy in the study database: http://atom.ts-publicaccess.eu/form/tree?TreeId=100027 (see 'I101 (OP_EUR-Lex) CDM 'Work' types').

[165] EUR-Lex vocabularies in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000035.

4.2.10.3.1. Document originator vocabularies

4.2.10.3.1.1. Vocabulary Author (AU)

The Author field indicates the name of the institution, the body or the country that produced the act. This metadata field currently uses two controlled vocabularies (NALs):

- Corporate bodies[166]
- Countries[167]

The empty search can be used to extract information about the use of 'Author' values. The following table shows the Top 10 contributing authors. Currently, there are 2 674 distinct author entries in use. Since there can be multiple authors for each document the total sum in Table 15 exceeded the total number of documents in chapter 4.2.10.1.4.

No.	Author	Count
1	European Parliament	226 988
2	European Commission	190 860
3	Council of the European Union	48 939
4	National Courts	27 691
5	Court of Justice	27 149
6	European Communities	25 617
7	Court of Justice of the European Union	15 000
8	Court of First Instance	12 933
9	United Kingdom	11 459
10	Austria	11 341
-	Not listed entities (2 663 distinct authors)	340 485
-	Listed as 'Other' entry	241 723

Table 15: The Top 10 contributing authors in EUR-Lex

4.2.10.3.2. Vocabulary of Topics

The vocabulary Topic represents a specification of what the document is about. EUR-Lex uses the following taxonomies for such descriptions:

- EuroVoc descriptor (DC)
- Subject matter (CT)
- Directory code (CC)
- EU Case Law directory code (RJ_NEW)

4.2.10.3.2.1. EuroVoc

The EuroVoc[168] is a multilingual and multi-disciplinary thesaurus which covers the field of activity of the EU. It was originally built up specifically for processing the documentary information of the EU institutions, but it is used widely outside the EU institutions as well.

[166] EUR-Lex vocabulary of Corporate bodies in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1106231.
[167] EUR-Lex vocabulary of Countries in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1108515.
[168] EuroVoc website: http://eurovoc.europa.eu.
The hierarchy of EuroVoc in the study database: http://atom.ts-publicaccess.eu/form/tree?TreeId=100096.

EuroVoc has a hierarchical structure. There are several 'layers' within it. The first (top-most) layer is 'Domain' which contains 21 values, the second 'layer' is 'Microthesaurus' which has 127 entries. There are several other layers between 'Concepts' (6884 values).

EuroVoc descriptors are assigned in a consistent manner to documents within all sectors except for sectors 0, 8 and partially sector 6. The usual number of descriptors covering each document varies from one to ten.

4.2.10.3.2.2. Directory code

Currently, there are 475 values within the ATTO table FD_555 which is used for the Directory code. The Directory code is attributed to sectors 1-5 and E (as well as for the special editions of the OJ).

The number of attributed codes should not be higher than three codes for each document. The codes in FD_555 are organised into a hierarchical structure.[169]

Users of EUR-Lex can browse the attributed documents through existing Directories:

- Directory of European Union legislation[170]
- Directory of European Union preparatory acts[171]
- Directory of European Union consolidated acts[172]
- Directory of international agreements[173]

4.2.10.3.2.3. Subject matter

The Subject matter is a classification tool containing an alphabetically structured list of over 200 keywords (ATTO table FD_070). It is based on the subdivisions of the treaties and on the areas of activity of the institutions. The descriptors are less specific than those used in the Directory code but they provide a general overview of the content of the document. The structure of the Subject matter is not hierarchical.

According to the current methodology, one document cannot be attributed with more than three Subject matter descriptors.

4.2.10.3.2.4. EU Case Law directory code

The EU Case Law directory code is a specific directory classification for EU Case Law. It uses FD_578 and is attributed to some documents within the sector 5. The structure of the codes is also hierarchical and there are 2 241 codes within the table. The hierarchy can be viewed in the study database.[174]

Users of EUR-Lex can browse the attributed documents in the Directory of European Union Case Law.[175]

4.2.10.3.3. Other codified metadata

The following Table 16 shows various codified metadata with their accompanying vocabularies.

---

[169] Hierarchy of the FD_555 in the study database: http://atom.ts-publicaccess.eu/form/tree?TreeId=100027 (select the 'I101 (OP_EUR-Lex) FD_555' node).
[170] Directory of EU legislation: http://eur-lex.europa.eu/browse/directories/legislation.html.
[171] Directory of EU preparatory acts: http://eur-lex.europa.eu/browse/directories/legislation-preparation.html.
[172] Directory of EU consolidated acts: http://eur-lex.europa.eu/browse/directories/consleg.html.
[173] Directory of international agreements: http://eur-lex.europa.eu/browse/directories/inter-agree.html.
[174] Hierarchy of the EU Case law directory code in the study database: http://atom.ts-publicaccess.eu/form/tree?TreeId=100027 (see I101 (OP_EUR-Lex) FD_578).
[175] Directory of EU case law: http://eur-lex.europa.eu/browse/directories/new-case-law.html.

Codified metadata	NALs or ATTO
Miscellaneous information	
Service responsible (RS)	corporate-body
Associated service (AS)	corporate-body
Political group (AF)	FD_290, country, corporate body
Miscellaneous information (MI)	FD_400
Parliamentary term (LG)	FD_285
Depositary (DP)	FD_040
Addressee (AD)	FD_050
Authentic language (LF)	language
Rapporteur (RAPPORTEUR)	FD_013, FD_014
Additional information (IC)	FD_301
Number of session (NS)	FD_345
Publication reference	
Official Journal collection (OJ_SERIES)	document-collection
Official Journal special edition chapter (OJ_CHAPTER)	FD_555
EU Case Law properties	
Applicant (AP)	FD_110, country, corporate body, role qualifier
Defendant (DF)	FD_110, country, corporate body, role qualifier
Observations (OB)	FD_160, country, corporate body, role qualifier
Nationalities of parties (NA)	country, role qualifier
Type of procedure (PR)	FD_100, legal proceedings, legal proceedings results
Judge	Rapporteur (JR), FD_130
Advocate general (AG)	FD_100
National court (NC)	FD_100
Country (COUNTRY)	country

Table 16: Codified metadata in EUR-Lex with their accompanying vocabularies

4.2.10.3.4. Document reference

The Celex identifier (DN)[176] is considered the most important identifier of the documents published on EUR-Lex. The structure of Celex identifiers is intuitive enough (at least for commonly used documents) that the users on EUR-Lex may be able to create a Celex identifier for a searched document and use it as a search term to find it. This identifier is also commonly used for the creation of hyperlinks to documents on EUR-Lex. The value of the Celex is composed from several parts which are maintained within appropriate number of separated metadata (components of the Celex identifier value). These metadata can be used separately in the Expert search.

- Celex identifier (DN)
  - Document type sector (DTS)
  - Document type year (DTA)
  - Document type (DTT)
  - Document natural number (DTN)
  - Corrigendum number (DT_CORR)
  - Document sequence number (DSN)
- Obsolete Celex identifier (DN_old)

[176] EUR-Lex website: 'What are Celex numbers composed of?' http://eur-lex.europa.eu/content/help/faq/intro.html#help10.

4.2.10.3.5. Metadata related to the document text

- Title (TI)
- Alternative title (NOM_USUAL)
- Text (TE)

The subdivision of the document's text into several separate parts can be considered a special kind of metadata. These parts of the document's text can be individually searched through an 'Expert search':

- Keywords (IX)
- Parties (I1)
- Subject of the case (I2)
- Grounds (MO)
- Endorsements (VS)
- Decision on costs (CO)
- Operative part (DI)

Subparts are currently available for part of sector 6 (Special EU Case Law).

4.2.10.3.6. Dates

EUR-Lex uses relatively large amounts of dates; in addition to these, the values may also include additional information such as detailed specification of the document type or specification of the relevant location in a document. Such additional information is called 'Annotation' within CDM. The values in annotations usually use codified values from controlled vocabularies.

Dates:

- Date of document (DD)
- Date of publication (PD)
- Date of effect (IF)
- Date of end of validity (EV)
- Date of notification (NF)
- Date of transposition (TP)
- Date of signature (SG)
- Date of vote (VO)
- Date of debate (DB)
- Date lodged (LO)
- Date of dispatch (DH)
- Date of deadline (DL)
- Date of reply (RP)
- Date of debate

4.2.10.3.7. Other document attributes

- In force indicator (VV)
- Directory indicator (REP)
- Internal reference (RI)
- Internal comments (CM)
- Notes relating to the decision (NO)

4.2.10.4.Relations between documents

4.2.10.4.1. Internal relations

The internal relations can be viewed in the 'Linked documents' tab within the document's detail page.

There can be such relations as:

- Treaty (TT)
  - A codified relationship, uses NAL Treaties
- Legal basis (LB)
- Amendment to (MS)
- Amended by (MD)
- Earlier related instrument (EA)
- Subsequent related instrument (SP)
- Case affecting (AJ)
- Affected by case (CD)
- Instruments cited (CI)
- Related documents (RD)

Some relations are bidirectional ('SymmetricProperty'[177]) and some relations have a counterpart. For example, the 'Amendment to' has an inverse relation 'Amended by'. The latter is an inferred relation, which means that 'Amended by' is not actually an inserted value, but when an 'Amendment to' is inserted, the 'Amended by' is automatically created to supplement the information from the targeted document.

The relationships use 'Annotations' (similar to dates) to further specify the information in the created relationships. The annotations can contain values of mixed type. Such values can consist of combinations of a free-text and a codified value (in a specific format). For example, an 'amendment to' relationship can be annotated within the underlining XML data format to specify a relevant location as follows:

<annot:reference_to_modified_location>{AR|http://publications.europa.eu/resource/authority/fd_370/AR} 27</annot:reference_to_modified_location>

The XML value above is presented on EUR-Lex (English version) as 'Article 27'.

4.2.10.4.2. External relations

A part of EUR-Lex is a common access portal for sources of national law called N-Lex[178]. N-Lex provides search forms for all the Member States of EU. These searches provide results in the form of links to the national laws websites.

4.2.10.5.Procedure metadata

EUR-Lex also publishes (as briefly mentioned above) 'Legislative procedures', which follow the life cycle of a legislative procedure from the first moment it is launched until the final law is adopted. Legislative procedures have their own metadata. The other document sources (institutions) may refer to the same procedure within their register of documents. The procedure can also be published on different websites as a duplicate (see 4.2.1.2).

[177] W3C Recommendation: SymmetricProperty: https://www.w3.org/TR/owl-ref/#SymmetricProperty-def.
[178] N-Lex website homepage: http://eur-lex.europa.eu/n-lex/index_en.htm.

From a CDM perspective, procedures are subclasses of a 'Dossier' class. A 'Dossier' may contain several 'Events' and further on the 'Event' may contain several document references (even to documents outside of EUR-Lex).

4.2.10.5.1. Dossier related metadata

The 'Dossier' and its subclasses can have the following attributes:

- Identifier of the dossier
- Comment (on the dossier as a whole)
- Title
- Year of procedure
- Year of reference
- Number of reference
- Legal basis of the procedure (as text)
- Domain (as text)
- Procedure reference
- Reference type
- Number of procedure

Dates:

- Legacy date of creation
- Dossier was withdrawn on a given date
- Dossier was adopted on a given date

Relationships to vocabularies:

- Dossier is about dossier type concept (uses FD_612[179])
- Procedure has type concept type procedure (NAL Interinstitutional procedures[180])

Relationships to documents:

- Dossier initiated by preparatory act
- Dossier produces legal resource
- Procedure based on legal resource
- Dossier contains work

Relationships to events:

- Dossier contains event

The CDM model defines many more attributes and relationships but those not listed are not currently in use (no data has been found).

4.2.10.5.2. Event related metadata

Each 'Dossier' can contain multiple 'Events'.

Events can have attributes such as:

[179] FD_612 in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1226742.
[180] Interinstitutional procedures in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1110685.

- The type of the legal event
- Date
- Comment
- Initiating event
- Council session number
- Document reference (external document's reference as text)
- Legal basis

Relationships to documents:

- Legal event based on a legal resource
- Legal event containing a work

Relationships to vocabularies are listed in the Table 17.

Relationships name	Uses NAL or ATTO
Legal event addresses institution	corporate-body
Legal event associated with institution	corporate-body
Legal event consults institution	corporate-body
Legal event formally addresses institution	corporate-body
Legal event gets the opinion of an institution	corporate-body
Legal event gets the report of an institution	corporate-body
Legal event has remark	FD_603, FD_600
Legal event has type concept type event	event
Legal event informs institution	corporate-body
Legal event initiated by institution	corporate-body
Legal event is about prelex concept	fd_600
Legal event under joint responsibility of an institution	corporate-body
Legal event mandatorily consults institution	corporate-body
The event references to a council agenda	fd_600
Legal event optionally consults institution	corporate-body
Legal event under responsibility of an institution	corporate-body
The event operates under a given decision mode	fd_606
The event uses a given decision type	fd_602
Dossier was about historical dossier type concept	fd_612
Procedure had type concept type procedure	procedure

Table 17: Event relationships to vocabularies

Events also have relationships to a 'Person' type:

- Legal event under joint responsibility of person
- Legal event taken over by person
- Legal event under responsibility of person
- Legal event drafted by person
- Legal event reported on by person

Relationships to documents:

- Legal event based on legal resource

- Legal event containing a work

4.2.10.6.End-user search possibilities

This section analyses the search options that are available to the end-user in EUR-Lex. It covers three main areas in line with section 4.1, i.e. the search form, the list of results, and the document detail along with a brief conclusion.

4.2.10.6.1. Search form

EUR-Lex provides the users three ways to search. The first one is the standard quick search (a textbox with a suggesting (auto-complete) functionality). The second one is called 'Advanced search'[181], which enables the user to search using several metadata. The metadata that can be chosen depend on the selected 'Collection' filter. For example, when 'EU Case Law' is selected, all metadata belonging to this collection are shown (i.e. 'defendant' and 'applicant') and the metadata that cannot be used for such documents are hidden.

Another option for the creation of a search query is the use of 'Expert search'[182] . The Expert search enables the user to search over a large number of metadata and with a large degree of freedom. This is achieved by specifying a query in the form of a 'search query' in a specific query language[183]. This query uses specialised EUR-Lex metadata codes and is validated prior to its execution.

4.2.10.6.2. List of results

It is possible to configure the metadata to be displayed in rows in the list of results.

It is possible to export (a maximum of 100) query results to various formats (CSV, TSV, XLS, XML and PDF). The range of exported metadata may be managed in the same way as when displaying the results.

The list of results can be sorted by:

- Document title
- Document identifier
- Document date
- Document type (dependent on the query)
- Document author (dependent on the query)
- Celex sector (dependent on the query)

Further refinement of the search results can be performed using facets such as:

- Domain
- Subdomain
- Year of document
- Type of procedure
- Author
- Type of act

The resulting query may be saved or transformed into its own RSS on the search results page.

---

[181] EUR-Lex website: Advanced search: http://eur-lex.europa.eu/advanced-search-form.html.
[182] EUR-Lex website: Expert search (available only after login): http://eur-lex.europa.eu/expert-search-form.html.
[183] EUR-Lex website: Help – Expert search: http://eur-lex.europa.eu/content/help/search/intro.html#help3.

4.2.10.6.3. Document detail

The document detail enables the user to view all language versions of the document's text as well as an extended set of the document's metadata. Within the document's page there is also the possibility to download the different types of the document's text and a metadata export within 'notice' XML file.

The document's data is presented within several tabs[184]:

- 'About this document' (basic metadata information about the document)
- 'Text'
- 'Linked documents' (various relationships to the other documents (i.e. 'legal basis'))
- 'All' (metadata and the text of the document together)
- 'Procedure' (displays the legislative procedure which is linked to the document)
- 'Summary' (applicable to a subset of documents)

An important feature from the perspective of the end user is a multilingual view of the document's text, which allows up to three language versions to be displayed side by side.

4.2.10.6.4. General evaluation of the Search functionality

The search functionality of EUR-Lex can be considered as very good.

It is characterized by both user friendliness (quick and advanced search) and the rich set of search possibilities (expert search). The searches can be saved for the later use and/or transformed into an RSS feed. The possibility to bulk export the search results to various formats should also be considered useful for the end user. On the other hand, the response time of the searches performed sometimes took longer than expected (more than 10 seconds) even when the quick search was used. Further investigation of these delays is outside of the scope of the study but since long delays can be considered to be problematic from the 'user-experience' point of view it might be beneficial to perform detailed analysis of this phenomenon.

4.2.10.7.Sample document

To see documents from different document sources through the same lens, a randomly selected document from EUR-Lex was re-created in the unified structure of the study database. This could help to design the future integrated access solution or help to understand the treatment of metadata from different document registers at one location. The sample document from EUR-Lex is shown in Figure 36 and Figure 37 and is directly accessible in the study database.[185]

[184] EUR-Lex website: Help – Consulting search results: http://eur-lex.europa.eu/content/help/consulting-search-results/intro.html.

[185] Sample document from the EUR-Lex in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1234281.

Sample document **EUR-Lex**

BASIC INFORMATION

Title	Regulation (EU) 2015/2219 of the European Parliament and of the Council of 25 November 2015 on the European Union Agency for Law Enforcement Training (CEPOL) and replacing and repealing Council Decision 2005/681/JHA
URL	http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32015R2219
Register	Target: I101 Publications Office - EUR-Lex

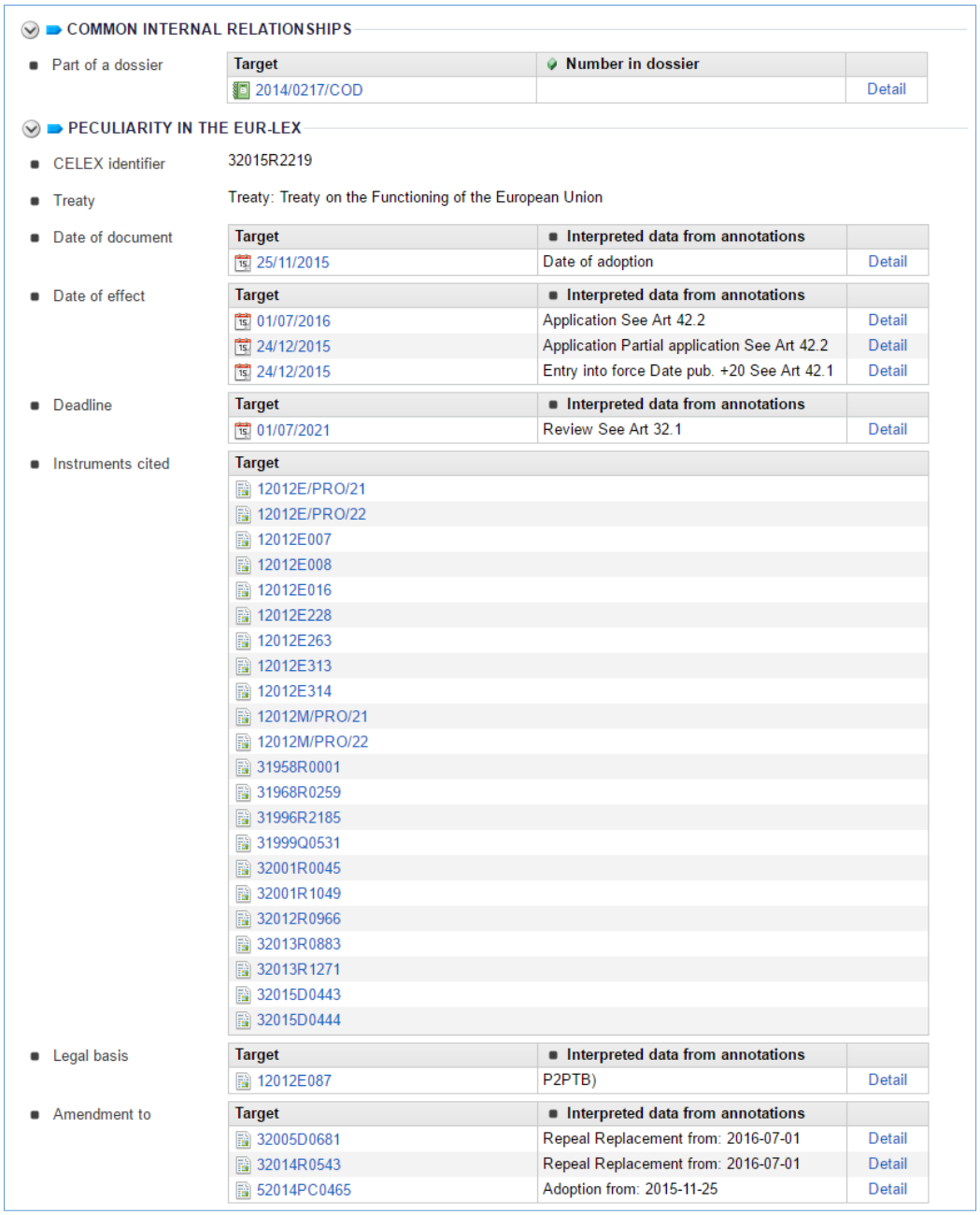
COMMON TYPES OF METADATA

Date of the document	25.11.2015

COMMON TYPES OF VOCABULARIES

Year	Target: 2015
Originator	Target: Council of the European Union; European Parliament
Topic(s)	Target: application of the law; EU office or agency; European Police College; Justice and home affairs (Subject matter); operation of the Institutions; Police cooperation (Directory code); Provisions governing the institutions (Directory code); Provisions governing the Institutions (Subject matter); seat of Community institution; vocational training
Type(s)	Target: Regulation
Language(s)	Target: Bulgarian; Croatian; Czech; Danish; Dutch; English; Estonian; Finnish; French; Gaelic; German; Greek; Hungarian; Italian; Latvian; Lithuanian; Maltese; Polish; Portuguese; Romanian; Slovak; Slovenian; Spanish; Swedish

Figure 36: Sample document from EUR-Lex (part 1)



COMMON INTERNAL RELATIONSHIPS

	Target	Number in dossier	
Part of a dossier	2014/0217/COD		Detail

PECULIARITY IN THE EUR-LEX

CELEX identifier: 32015R2219

Treaty: Treaty: Treaty on the Functioning of the European Union

Date of document	Target	Interpreted data from annotations	
	25/11/2015	Date of adoption	Detail

Date of effect	Target	Interpreted data from annotations	
	01/07/2016	Application See Art 42.2	Detail
	24/12/2015	Application Partial application See Art 42.2	Detail
	24/12/2015	Entry into force Date pub. +20 See Art 42.1	Detail

Deadline	Target	Interpreted data from annotations	
	01/07/2021	Review See Art 32.1	Detail

Instruments cited	Target
	12012E/PRO/21
	12012E/PRO/22
	12012E007
	12012E008
	12012E016
	12012E228
	12012E263
	12012E313
	12012E314
	12012M/PRO/21
	12012M/PRO/22
	31958R0001
	31968R0259
	31996R2185
	31999Q0531
	32001R0045
	32001R1049
	32012R0966
	32013R0883
	32013R1271
	32015D0443
	32015D0444

Legal basis	Target	Interpreted data from annotations	
	12012E087	P2PTB)	Detail

Amendment to	Target	Interpreted data from annotations	
	32005D0681	Repeal Replacement from: 2016-07-01	Detail
	32014R0543	Repeal Replacement from: 2016-07-01	Detail
	52014PC0465	Adoption from: 2015-11-25	Detail

Figure 37: Sample document form EUR-Lex (part 2)

4.2.10.8. Re-use of EUR-Lex in view of integrated access solution

The possibility of reusing EUR-Lex for PublicAccess.eu should be considered in terms of its metadata storage: CELLAR. EUR-Lex publishes a web service which could be used as a data source for PublicAccess.eu, but it is more suitable to use a direct access to the CELLAR via the SPARQL endpoint. Some functionalities are accessible only through this endpoint (for example an attribute denoting document's last modification date).

From the metadata perspective the CELLAR database and its ontology system (CDM) can be easily used for solution based on non-unified data inputs. The range of use of the controlled vocabularies and the overall structure of the data model shows promising possibilities for all types of re-use of its metadata.

There are also further possibilities for solutions based on an extension of the CDM and usage of the CELLAR database as a store for all documents in the scope of PublicAccess.eu.

4.2.11. Tenders Electronic Daily

Tenders Electronic Daily ('TED') is the online version of the 'Supplement to the Official Journal of the European Union' dedicated to European public procurement.

4.2.11.1.General information

On average, approximately 1 700 public procurement notices are published on TED every day from Tuesday to Saturday.

Information about every procurement is published in the 24 official EU languages. All notices from the EU's institutions are published in full in these languages.

The content of the notices consists of predefined structures. Notices are not primarily stored in PDF format but in a structured XML. Such a format provides numerous advantages for multilingual data presentation.

Other documents can be found in the e-Tendering platform (https://etendering.ted.europa.eu). This platform publishes documents from EU institutions only. Every 'Call for tender' may include documents such as: Invitation, Draft contract, Technical Specifications, Price schedule, etc. In March 2016, there are about 7 200 documents associated with about 1 000 calls for tenders.

4.2.11.1.1. Public access

TED is accessible at this general URL address:
http://ted.europa.eu/

A list of all documents is available at this URL address:
http://ted.europa.eu/TED/search/searchResult.do?sort=PUBLICATION_DATE&dir=desc

The public has direct access to all of these documents in their electronic form. Electronic access is free and no special authentication is necessary.

The documents are also available in bulk XML exports.[186] Bulk exports allow users to download packages from 2011 and later.

Moreover, these packages can be found also on the FTP (ftp://ted.europa.eu) accessible with generic credentials (guest/guest), dating back to 1993.

4.2.11.1.2. Time range covered

Documents are archived (and published on the website) for a period of five years. Older documents are not available on the Web but are stored on the FTP.

4.2.11.1.3. Overall volume of published documents

The volume of published documents at the beginning of March 2016 was 2 184 760.[187]

[186] XML exports of the TED (available only after login):
http://ted.europa.eu/TED/misc/xmlPackagesDownload.do

[187] Relevant for the documents available on the webpage only.

4.2.11.1.4. Brief Investigation of the TED

Figure 38 shows the results of the comparative analysis of TED.[188]

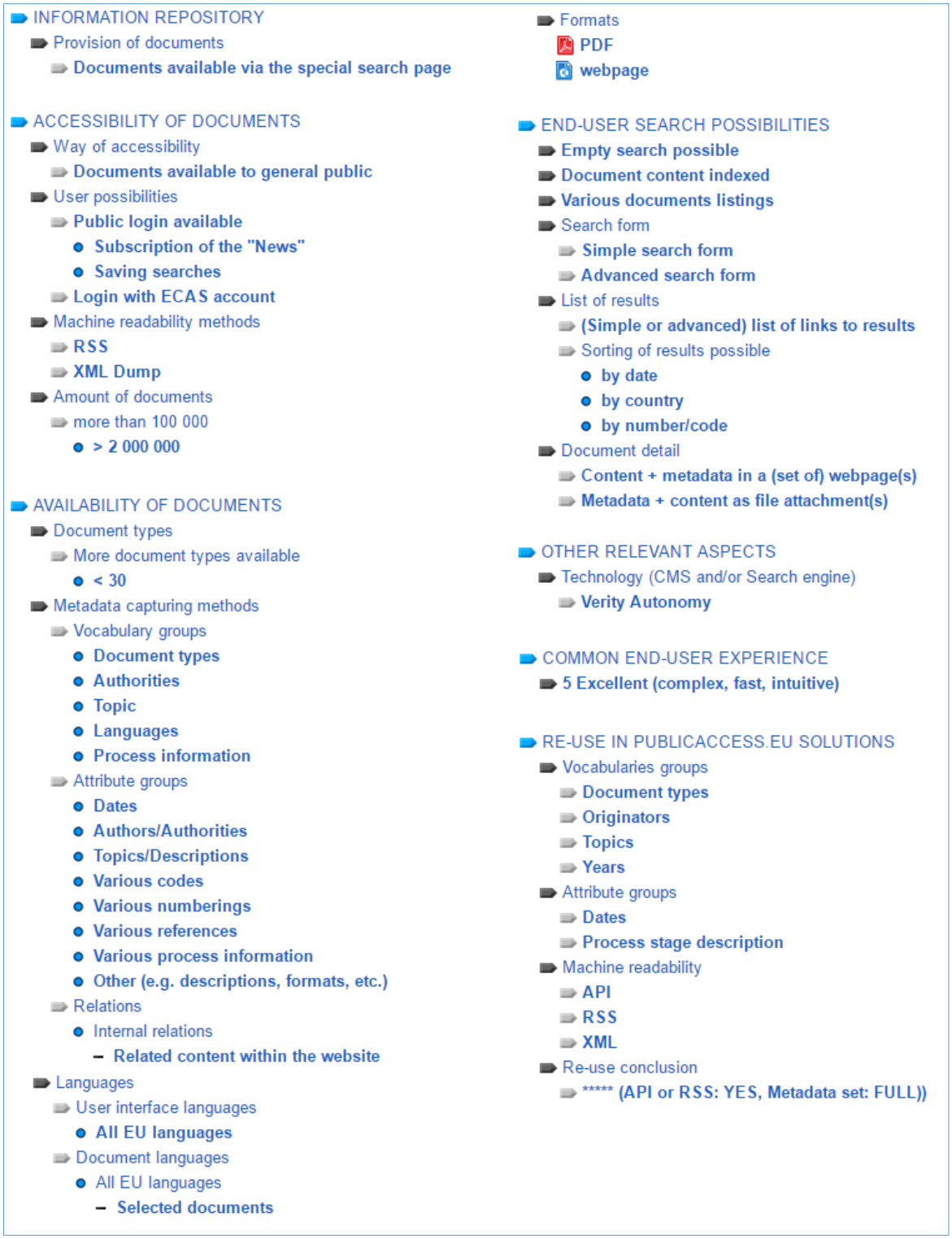


Figure 38 Overview investigation of TED

[188] Brief investigation of TED in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000037.

4.2.11.2.TED Document types

The documents displayed on the TED website are divided into 27 types. The most important document types, based on frequency, are the following: **Contract notice, Contract award notice,** and **Additional information** which account for 93 % of all documents published on TED. Some document types began to be used only in recent years. For example, 'Subcontracts in the fields of defense and security' (used from 2014) or 'Modification of a contract/concession during its term' (used from 2016).

Growth in document numbers over the past five years ranges from between 0 – 6 %. Thus, it cannot be said that there is an upward trend in the number of published documents.

Document types and their frequency is shown in the following Table 18 (the source data is from an XML export on 29 February 2016):

Type of the document	2011	2012	2013	2014	2015	2016	Total
Contract notice	175 397	174 090	179 200	179 605	187 232	24 883	920 407
Contract award	157 787	157 499	160 356	163 245	173 366	28 560	840 813
Additional information	49 430	51 794	70 579	72 634	71 322	9 789	325 548
Prior Information Notice	11 845	15 520	17 396	16 780	17 522	3 331	82 394
Voluntary ex ante transparency notice	10 694	9 934	10 154	9 231	9 242	1 029	50 284
Design contest	1 810	1 506	1 427	1 163	1 310	185	7 401
Results of design contests	1 251	1 107	1 023	851	768	180	5 180
Periodic indicative notice (PIN) without call for competition	844	616	279	299	470	77	2 585
Qualification system with call for competition	728	724	674	676	759	135	3 696
Public works concession	415	404	344	242	272	42	1 719
Qualification system without call for competition	367	408	417	564	454	74	2 284
European company	312	401	453	270	219	34	1 689
Buyer profile	171	215	247	248	281	66	1 228
Works contracts awarded by the concessionaire	138	68	43	79	70	7	405
Call for expressions of interest	133	38	44	37	33	3	288
Periodic indicative notice (PIN) with call for competition	122	122	102	95	63	7	511
Dynamic purchasing system	110	188	149	203	239	32	921
European economic interest grouping (EEIG)	99	92	78	96	96	14	475
General information	96	15	2	1	0	0	114
Corrigenda	92	86	101	76	77	13	445
Prequalification notices	9	9	11	19	10	1	59
Subcontracts in the fields of defence and security	0	0	0	5	9	4	18
Prior information notice with call for competition	0	0	0	0	3	6	9
Services concession	0	0	0	0	4	9	13

Modification of a contract/concession during its term	0	0	0	0	0	13	13
Concession award notice	0	0	0	0	0	2	2
Total	411 850	414 836	443 079	446 419	463 821	68 496	2 248 501
Yearly growth	-	0,73%	6,81%	0,75%	3,90%	-	-

Table 18: Number of documents in TED Document types

4.2.11.3. Metadata as relationships between documents and vocabularies

It can be deduced from the codified metadata that there is a clear emphasis on multi-lingual presentation in preparation of the data. Codified metadata include for example the following:

- Heading (combination of type of tender procedure, contract and applicable regulation)
- Country[189]
- Country groups[190]
- Type of authority (sector or awarding authority)[191]
- Contract type (market code)[192]
- Procedure type[193]
- Document type[194]
- Regulation type[195]
- Type of tender, division into lots[196]
- Award criteria[197]
- Common Procurement Vocabulary ('CPV') code[198] (current and original)
- Main activity[199]

[189] TED vocabulary of Country in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062460.
[190] TED vocabulary of Country groups in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062463.
[191] TED vocabulary of Type of authority in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062465.
[192] TED vocabulary of Contract type in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062467.
[193] TED vocabulary of Procedure type in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062469.
[194] TED vocabulary of Document types in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062283.
[195] TED vocabulary of Regulation type in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062473.
[196] TED vocabulary of Type of bid in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062475.
[197] TED vocabulary of Award criteria in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062477.
[198] TED vocabulary of CPV codes in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062479.
More about CPV codes can be found at: http://simap.ted.europa.eu/en/web/simap/cpv.
[199] TED vocabulary of Main activity in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062483

- NUTS code[200] (current and original)
- Languages[201]
- Extended CPV code (Additional vocabulary)[202]

4.2.11.3.1. Document originator vocabularies

The Authority name is isolated within the rest of the data but no indication was found that there is a vocabulary including the specific authors. The vocabulary 'Type of authority' is used here instead.

4.2.11.3.1.1. Vocabulary of Type of authority

Table 19 below outlines how the 'Type of authority' vocabulary is used. Currently, 8 codified values are used. 'Not specified' and 'Not applicable' are the special 'types' denoting an absence of data.

Type of authority	Count	Percentage
Regional or local authority	614 105	28,12%
Body governed by public law	466 294	21,35%
Other	399 131	18,28%
Ministry or any other national or federal authority	239 975	10,99%
Utilities entity	217 100	9,94%
Not specified	111 531	5,11%
Regional or local Agency/Office	57 727	2,64%
National or federal Agency/Office	46 922	2,15%
European Institution/Agency or International Organisation	28 386	1,30%
Not applicable	2 582	0,12%
Total	2 183 753	

Table 19: Vocabulary Type of authority (from the search results)

4.2.11.3.2. Vocabulary of Topics

4.2.11.3.2.1. Heading

The heading is a special type of vocabulary whose five-digit codes represent the following:

- The first component, consisting of two characters, is the origin of the document.
- The second component (and third character) signifies what kind of contract it is and is equivalent to 'Contract type'.
- The fourth character denotes the 'Document Type'.
- The final character denotes the 'Procedure type'.

Given the fact that the last three characters are represented as separate metadata, only the first component is interesting from the study's point of view. For example, if the first character is not equal

[200] TED vocabulary of NUTS codes in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062485.
More about NUTS codes can be found at: http://ec.europa.eu/eurostat/web/nuts.

[201] TED vocabulary of Languages in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062487.

[202] TED vocabulary of Extended CPV codes in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1062489.

to '0' it is a European institution. Therefore, based on the heading code, it is possible to determine from which European institution the document originates.

4.2.11.3.2.2. CPV classification

The CPV establishes a single classification system for public procurement. It aims to standardise the references used by the contracting authorities and entities to describe the subject of the procurement contracts.

Assigned CPV codes are diversified into two types of metadata depending on whether their use was prior to or after the adoption of Regulation (EC) No 213/2008.[203]

CPV consists of 9 455 codes and 904 supplementary codes in additional vocabulary. All text code names are translated into the 24 languages. CPV codes are arranged in a tree hierarchy (division, group, class, category, and item).

4.2.11.3.2.3. NUTS classification

NUTS stands for 'Nomenclature of Territorial Units for Statistics'.[200] Similarly to the CPV, NUTS is also a hierarchical system, which defines the location, where the contract is to be performed.

4.2.11.3.2.4. Procedure type

At present, the vocabulary 'Procedure type' uses 13 values (+2 denoting empty values). It was found that the Procedure type also includes the value 'Unknown' and blank values which are reported in Table 20 below as 'not applicable'.

Procedure	Count
Open procedure	1 631 137
Negotiated procedure	147 555
Restricted procedure	131 003
Contract award without prior publication	73 837
Negotiated without a call for competition	60 443
Accelerated restricted procedure	10 790
Competitive dialogue	9 039
Not specified	7 716
Accelerated negotiated procedure	6 875
Direct award	485
Competitive procedure with negotiation	57
Other	46
Concession award procedure	16
Unknown	9 548
Not applicable	96 213
Total	2 184 760

Table 20: Procedure types (from the search results)

4.2.11.4.Metadata as document attributes

The following metadata may be considered as document attributes:

- Name of the awarding authority

[203] Commission Regulation (EC) No 213/2008 on the EUR-Lex: http://eur-lex.europa.eu/eli/reg/2008/213/oj.

- Place
- Internet address (URL) of the awarding authority
- Date document sent to the Publications Office
- Deadline for request of documents
- Deadline for receipt of tenders
- Publication date
- Number of reference document(s)
- Document number
- Edition number of supplement to the OJ
- Origin of invitation to tender (applicable regulation for procurement)
- Title of document
- Title of the main activity
- Directive

4.2.11.5. Relations between documents

4.2.11.5.1. Internal relations

The information displayed in the 'Document family' tab, which displays all documents related to the same tendering procedure, may be considered to the internal relations.

For a certain subset of documents, one may find a relation between the document and the 'Call for tender' section published at https://etendering.ted.europa.eu.

4.2.11.5.2. External relations

No existing external relations have been found.

4.2.11.6. End-user search possibilities

This section analyses the search options that are available to the end-user in the TED, and it covers three main areas in line with section 4.1, i.e. the search form, the list of results, and the document detail along with a brief conclusion.

4.2.11.6.1. Search form

The TED website offers two ways of searching. The first one allows the user to search using a selected set of metadata (including full-text search)[204]. The second one is the 'expert search'[205], which offers a search form with a wider array of metadata (30 in total) to filter the results. It uses a structured query language ('SQL') methodology, or common command language ('CCL').[206]

An important functionality of the search is the possibility to save the query to a list of favourite queries in the user's account. These queries can be reloaded later for reuse or one can create a customized RSS from it. Due to the nature of the published queries, this functionality is considered very beneficial for the end-users.

4.2.11.6.2. List of results

The list of results consists of links to documents in TED, where each document entry is characterized by the following metadata:

- Document number

[204] TED website – General search: http://ted.europa.eu/TED/search/search.do
[205] TED website – Expert search: http://ted.europa.eu/TED/search/expertSearch.do.
[206] A standard text retrieval query language proposed by the International Organization for Standardization.

- Description
  - Title
  - Type of authority
  - Document type
  - Procedure
  - Contract
- Country
- Publication date
- Deadline

The list of results is equipped with additional options for narrowing the volume of documents by filtering through the facet filters for which the vocabularies: 'Document type', 'Country' and 'CPV code' are used. The facets used are gathered in a special box from which they can easily be removed.

The list of results differs based on the chosen scope of search. The scopes of the search options are:

- Last edition
- All current notices
- Archives

When the option Archives is set then all documents are searched. However, for this option the facet filters cannot be used. The facet filtering is only available for the 'Last edition' and 'All current notices' scopes of the search.

The list of results also offers the specialized view: 'Statistic search result'. This can create tables with the numbers of documents based on the selected metadata on the X and Y axis.

4.2.11.6.3. Document detail

After clicking on the document title in the list of results the user receives additional information about the document. Documents data are displayed in a tabbed manner. The following tabs may be shown in the document details page:

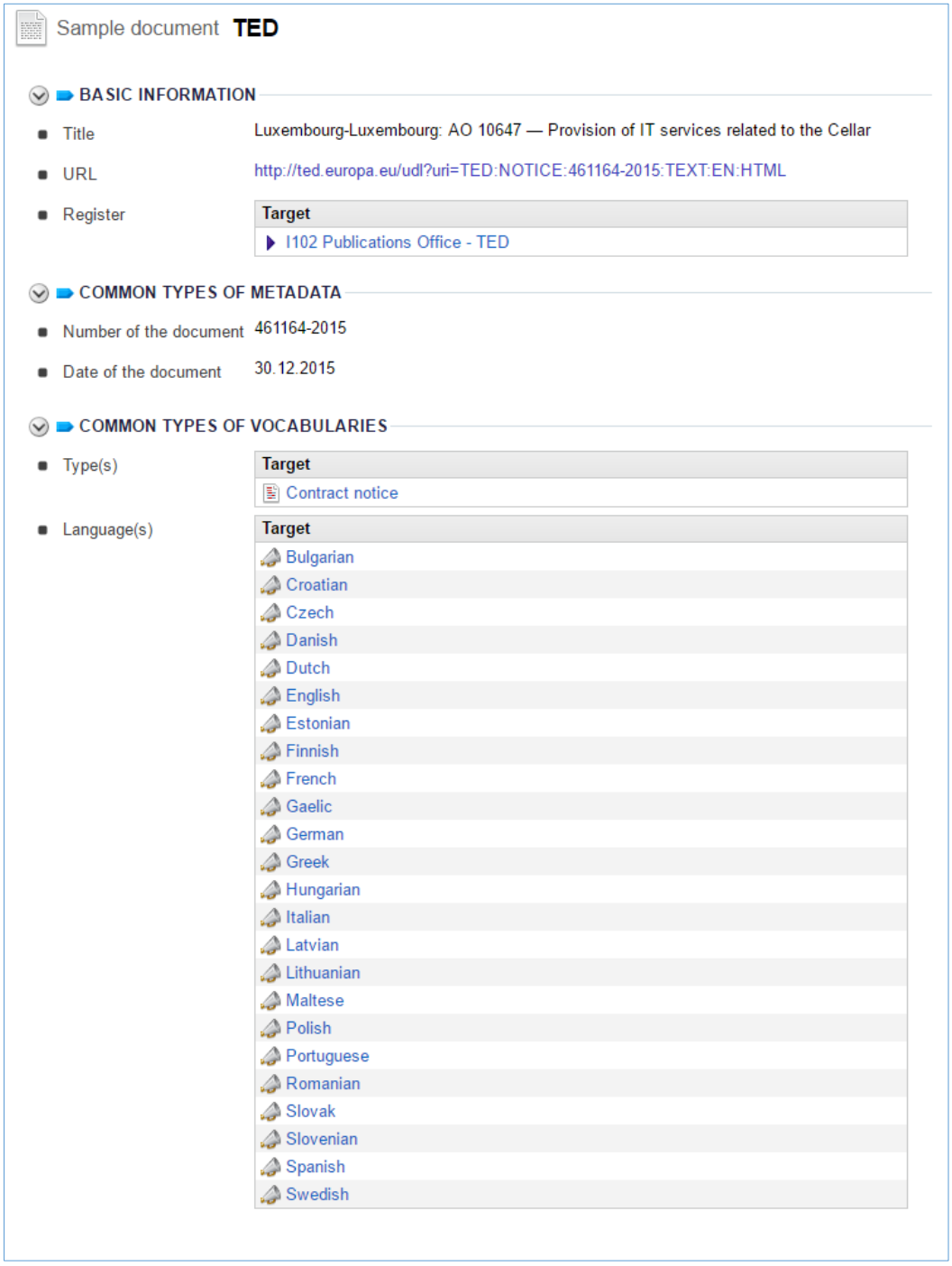
- Current language
  - Displays a partial set of information about the document in the selected UI language
- Original language
  - Displays a complete set of information about the document in original language(s)
- Summary view
  - Displays a complete set of information in selected UI language
- Data
  - Displays catalogue data about the document
- Document family
  - Displays all the notices that refer to the same tendering procedure
- Machine translation
  - Displays translation of the document

4.2.11.6.4. General evaluation of the Search functionality

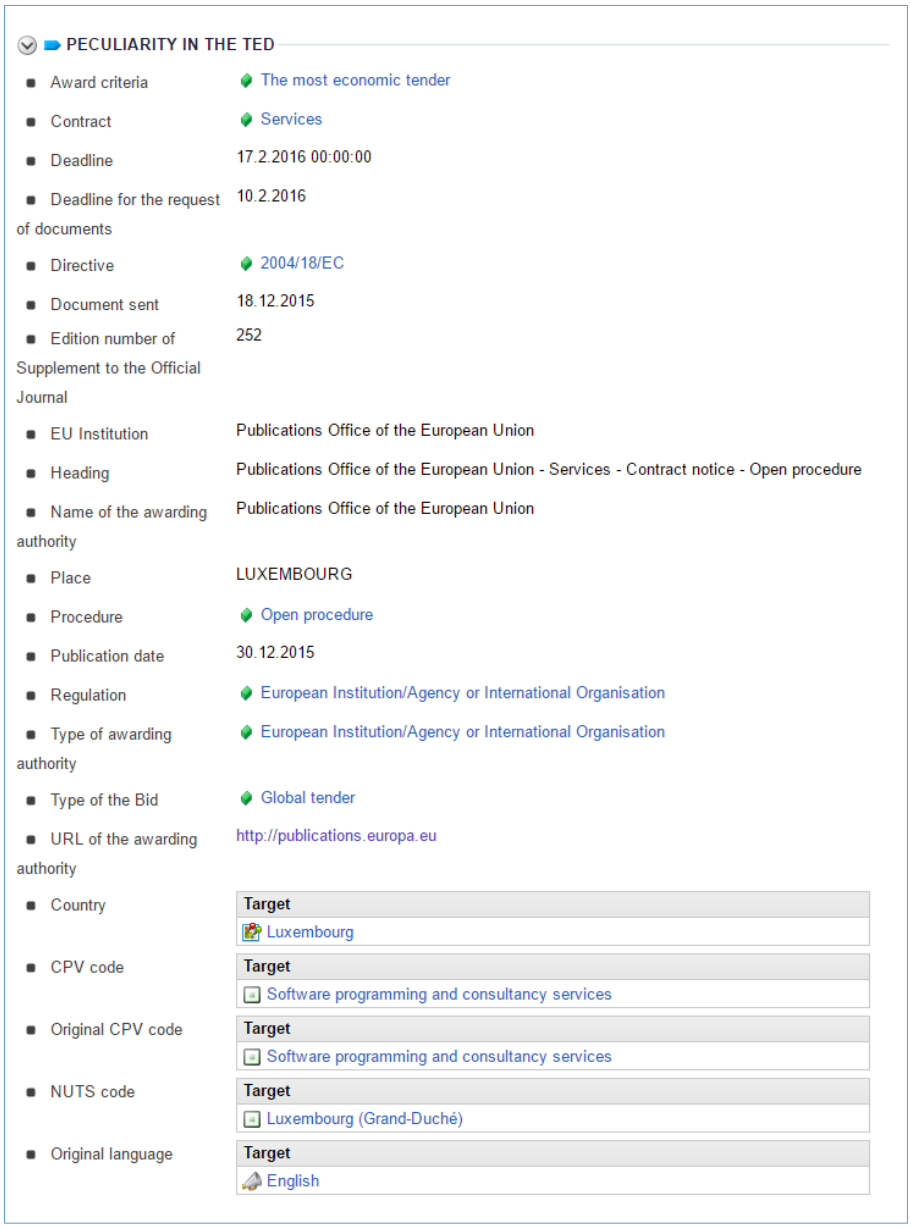
The search functionality works very well and provides a rich functionality for users.

4.2.11.7.Sample document

To see documents from different document sources through the same lens, a randomly selected document from TED was re-created in the unified structure of the study database. This could help to design the future integrated access solution or help to understand the treatment of metadata from different document registers in one location. The sample document from TED is shown in Figure 39 and is directly accessible in the study database.[207]



[207] Sample document from the TED in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1065438.



PECULIARITY IN THE TED

Award criteria	The most economic tender
Contract	Services
Deadline	17.2.2016 00:00:00
Deadline for the request of documents	10.2.2016
Directive	2004/18/EC
Document sent	18.12.2015
Edition number of Supplement to the Official Journal	252
EU Institution	Publications Office of the European Union
Heading	Publications Office of the European Union - Services - Contract notice - Open procedure
Name of the awarding authority	Publications Office of the European Union
Place	LUXEMBOURG
Procedure	Open procedure
Publication date	30.12.2015
Regulation	European Institution/Agency or International Organisation
Type of awarding authority	European Institution/Agency or International Organisation
Type of the Bid	Global tender
URL of the awarding authority	http://publications.europa.eu

	Target
Country	Luxembourg
CPV code	Software programming and consultancy services
Original CPV code	Software programming and consultancy services
NUTS code	Luxembourg (Grand-Duché)
Original language	English

Figure 39: Sample document from the TED

4.2.11.8.Re-use of TED in view of integrated access solution

TED exhaustively covers relevant metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT (is the document about)	Document type; Topics
WHEN (was the document published)	Years; Dates
WHO (is responsible for the document)	Authorities
WHY (purpose of the document)	Process information; Other metadata

TED XML exports provide well-structured data. Therefore, the exported data can meet the requirements for solution alternatives based on non-unified data inputs.

4.3. Analysis of the document registers of the agencies

The EU agencies which were selected for the study have specific document registers, where they publish documents for public viewing on their websites.

It must be noted that some EU agencies do not store their documents in a document register but publish them scattered across their websites. For example, the Executive Agency for Small and Medium-sized Enterprises and the Innovations and Networks Executive Agency do not have a publicly accessible document register available on their websites at all.

The goal of the following analysis is to investigate the ways in which the documents are published in each given agency, and to analyse the selected document registers. A focus is put on reviewing the possibilities of the document registers for re-use in a future integrated access solution.

Each agency and its document register is described in a separate subchapter. Because of the relatively small volume of documents included, the analysis is processed in more **general** way than for the EU institutions in chapter 4.2. However, it still includes the most important points from the structure specified in the chapter Method for the analysis of the document sources (see 4.1).

4.3.1. Executive agencies

4.3.1.1. Education, Audiovisual and Culture Executive Agency

The Education, Audiovisual and Culture Executive Agency ('EACEA') is responsible for the management of certain parts of the EU's funding programmes in the fields of education, culture, audiovisual, sport, citizenship and volunteering.

All the information about its activities and responsibilities are available on the EACEA website (http://eacea.ec.europa.eu).

The main focus of the EACEA website is on the programme management sections. There are 5 of them:

1. Erasmus+
2. Creative Europe
3. Europe for Citizens
4. EU Aid Volunteers
5. Eurydice

Although each programme has its own dedicated section they share the same structure with the following sections:

- General programme information
- Actions
- Funding
- Beneficiares space
- Selection results
- Library
- News
- Events contacts

The content is presented in the form of richly structured articles. These articles contain several attached files with detailed information. It is hard to imagine that these articles could be used separately (without the accompanying articles). Because of this reason and also because of the relatively small volume of documents included, the following analysis will be processed in a **general** way.

4.3.1.1.1. Public access

The EACEA website is accessible at this general URL address: http://eacea.ec.europa.eu.

The general EACEA document register is accessible at this URL:
http://eacea.ec.europa.eu/about-eacea/document-register_en.

The document register consists of a single web page with links to documents on:

- EACEA legal background
- EACEA activities (work programmes, activity reports)
- EACEA financial information
- Calls for proposals
- Audit procedure
- Other

The links to the documents route in most cases to EUR-Lex.

The documents in the document register are public, no authorization is required.

There is a general search accessible from all website pages and accessible also at the URL: http://eacea.ec.europa.eu/search.

The results are organised into the following categories:

- Programmes
- Author
- Metadata (keywords)
- Language
- Content type

Part of the documents accessible on the search page are public and part require authorization.

4.3.1.1.2. Overall volume of published documents

The number of documents published in EACEA document register is circa 100 primary language versions (mostly English) and circa 200 other language versions.

The volume of documents accessible via search is circa 900.

4.3.1.1.3. Time range covered

The EACEA website covers documents produced in the period 2006 - 2016.

4.3.1.1.4. Brief investigation of the EACEA

Figure 40 shows the results of the comparative analysis of the EACEA website.[208]

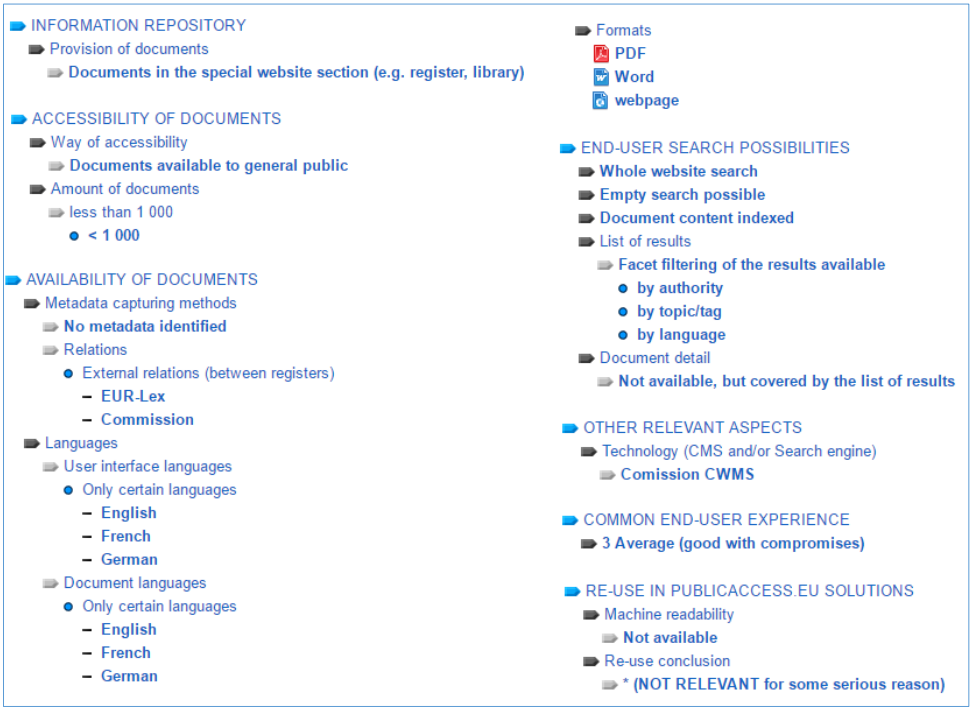


Figure 40: Overview investigation of the EACEA website

[208] Brief investigation of the EACEA in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000170.

4.3.1.1.5. Re-use of the EACEA in view of integrated access solutions

The EACEA document source does not cover the metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT (is the document about)	Not available
WHEN (was the document published)	Not available
WHO (is responsible for the document)	Not available
WHY (purpose of the document)	Not available

It is quite difficult to imagine that it could be used in any of the integrated access solution alternatives. There are two main reasons for this difficulty:

1. Absence of any machine readable format.
2. Information is displayed as website articles on the EACEA website and for this reason the documents attached to the articles cannot be presented independently.

The re-use options of the EACEA website are very limited or even impossible. Serious technical improvements need to be implemented in order to grant automatic access to documents contained in the EACEA website.

4.3.1.2. European Research Council Executive Agency

European Research Council Executive Agency ('ERCEA') implements and manages operations of the European Research Council ('ERC') and is responsible for all aspects of the administrative implementation and execution of the ERC Programme. The mission of the ERC is to encourage the highest quality research in Europe through competitive funding and to support, in particular, young scientists and investigators. Its main goal to find and fund the best researchers in Europe and those from outside Europe who want to come to work in Europe. Their vision is to change 'science in Europe' into 'European science.'

The ERC approach is 'investigator-driven' rather than 'politician-driven'.

All information about ERC activities, responsibilities and projects are presented on the ERC website. ERCEA as an organisational part of ERC does not have its own website, so the general ERC website has been analysed.

The ERC website is divided into 3 sections:

1. Funding and Grants
2. Projects and Results
3. Media and Events

Each section has its own dedicated structure.

The section 'Projects and Results' possesses a standalone database of projects from which information can be found through filtering. It is, however, out of the scope of the study since it does not contain Documents (as defined in Section 4.1).

Content is presented in the form of richly structured website articles. These articles contain several document attachments as files which contain detailed information. It is hard to imagine that these articles could be used separately (without the files bearing them). Because of this a **general** analysis was carried out.

4.3.1.2.1. Public access

The ERC website is accessible at this general URL address: https://erc.europa.eu.

Information about ERCEA, its tasks and organisational structure is accessible at this URL: https://erc.europa.eu/about-erc/organisation-and-working-groups/executive-agency.

The general ERC document library is accessible at this URL: https://erc.europa.eu/document-library.

The documents in the document register are public, no authorization is required.

4.3.1.2.2. Overall volume of published documents

The volume of documents published in the ERC document library is ca 320. Several documents have an ISBN code assigned to them. The library does not provide options to select only documents belonging to the ERCEA, just the common library for all ERC documents is available. For this reason, analysed information covers all documents from ERC included in the common library.

4.3.1.2.3. Time range covered

The ERC website covers documents produced in the period 2007 - 2016.

4.3.1.2.4. Brief investigation of the ERC website

Figure 41 shows the result of the comparative analysis of the ERC website.[209]

Figure 41: Overview investigation of the ERC website

4.3.1.2.5. Document types in the ERC documents library

The *document type* is determined by the relationship between one specific item in the vocabulary of the document categories and the document itself.

The investigation revealed that 'document categories' substitute for 'document types.' The vocabulary of document categories[210] is organised as a hierarchy with 10 entries in the 1st level and another 19 entries in the 2nd level.

The titles of the document categories also replace the role of 'Topic.' In other words, they can describe the document.

[209] Brief investigation of the ERC in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000175.

[210] ERC vocabulary of Hierarchy of document categories in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1090903.

4.3.1.2.6. Metadata as document attributes

Only one meta-information was investigated – the date of document.

4.3.1.2.7. Re-use of the ERC document library in view of integrated access solution

The ERC document library does not cover enough metadata to answer basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT **(is the document about)**	Document types Topics not available (but basic structure of topics can be derived from Document types)
WHEN **(was the document published)**	Dates
WHO **(is responsible for the document)**	Not available
WHY **(purpose of the document)**	Not available

It is quite difficult to imagine that the ERC document library could be used in any integrated access solution. The main reason is the absence of any machine readable format.

A lot of improvements both from the of content (metadata) perspective as well as from the technical point of view need to be implemented in order to re-use the ERC document library for any integrated access solution.

4.3.2. Regulatory agencies

4.3.2.1. Body of European Regulators for Electronic Communications

The Body of European Regulators for Electronic Communications ('BEREC') objective is to contribute to the development and improved functioning of the internal market for electronic communications networks and services. Furthermore, BEREC assists the EC and the national regulatory authorities (NRAs) in implementing the EU regulatory framework for electronic communications. It provides advice on request and on its own projects to the European institutions and complements at European level the regulatory tasks performed at national level by the NRAs. It was established in 2009.[211]

4.3.2.1.1. General information about the BEREC website

On its website, the BEREC provides information on its organisation, activities, events, as well as press releases, news and much more.

Only a **general** analysis of the BEREC website document source was carried out.

4.3.2.1.1.1. Public access

The BEREC website is accessible at this general URL address: http://berec.europa.eu.

The majority of documents available on the BEREC website are contained in the Document Register ('RD-BEREC') available at the URL address: http://berec.europa.eu/eng/document_register.

The documents published on the RD-BEREC are mostly accessible by the general public without any special authentication. However, some documents are only available upon request.

4.3.2.1.1.2. Overall volume of published documents

The number of documents published in the RD-BEREC to the end of February 2016 is ca 4 000. Ca 60% of them are equipped with a PDF attachment. This means some ca 40% of the content is presented in the form of website articles only.

4.3.2.1.1.3. Time range covered

The RD-BEREC covers documents from the beginning of its operation. The first document published on the RD-BEREC is dated 1.1.2010.

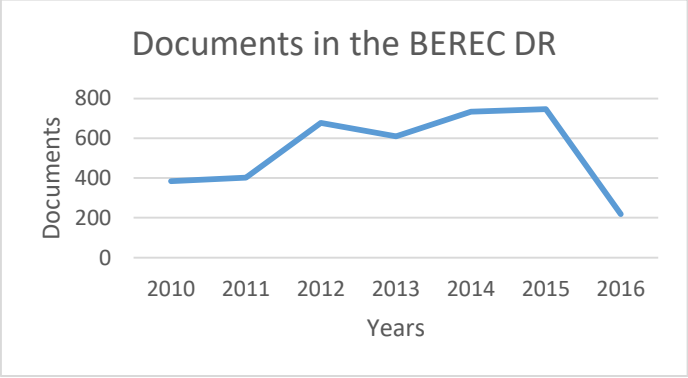


Figure 42: Volume of published documents published in the BEREC Document Register by years

[211] Source of the characteristics: http://berec.europa.eu/eng/about_berec/what_is_berec.

4.3.2.1.1.4. Brief investigation of the RD-BEREC

Figure 43 shows the results of the comparative analysis of the RD-BEREC.[212]

Figure 43: Overview investigation of the BEREC

4.3.2.1.1.5. RD-BEREC Document types

The Document types in the RD-BEREC [213] are represented by the vocabulary of Subject matter containing a list of 43 document types. This vocabulary is organised as a hierarchy in 3 levels.

The investigation revealed that there are duplicates in the 2nd level of the Document types hierarchy.

The relationship between a document and entries in the vocabulary of RD-BEREC Document types has 1:1 cardinality. This means that the document must be of exactly one document type.

4.3.2.1.1.6. Metadata as document attributes

The following document attributes were investigated and they are mandatory for each document in the RD-BEREC:

- Document number

[212] Brief investigation of the RD-BEREC in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000086.

[213] RD-BEREC vocabulary of Document types in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1090989.

- Document date
- Date of registration
- Author/Authority (based on the investigation it is not a vocabulary)

4.3.2.1.1.7. End-user search possibilities

This section analyses the search options that are available to the end-user in the RD-BEREC. It covers three main areas in line with the section 4.1, i.e. the search form, the list of results, and the document detail along with a brief conclusion.

4.3.2.1.1.7.1. Search form

There is an advanced search form available for retrieving entries from the RD-BEREC.

This advanced search form is exhaustive and allows filtering by the document types described above (see 4.3.2.1.1.5) and all metadata attributes (see 4.3.2.1.1.6):

- Document number
- Words in title
- Text in document
- Subject matter (Document type)
- Document dates (from-to)

The use of more than one criterion is processed as a chain linked by the logical AND operator. This means that the more criteria are used, the fewer results are produced.

4.3.2.1.1.7.2. List of results

This consists of the links to entries in the RD-BEREC organised in the grid with the following columns:

- Document number
- Document date
- Document title
- Document author

The list of results is not paginated. It always displays the full scope of documents found by the search query: this could be seen as an advantage.

4.3.2.1.1.7.3. Document detail

The document detail is extremely user friendly. Aside from all the metadata and links to PDF attachments it is also short with comprehensively written annotations of the information shown. Sometimes they are accompanied by links to video recordings of the meeting etc.

4.3.2.1.1.7.4. General evaluation of the Search functionality

The search functionality works very well. It is also very intuitive for the end-users.

4.3.2.1.1.8. Sample document

To see documents from different document sources in a common pattern, a randomly selected document from the RD-BEREC was re-created in the unified structure of the study database. This could help to design the future integrated access solution or help to understand the treatment of metadata from different document registers at one location. The sample document from the RD-BEREC is shown in Figure 44 and is directly accessible in the study database.[214]

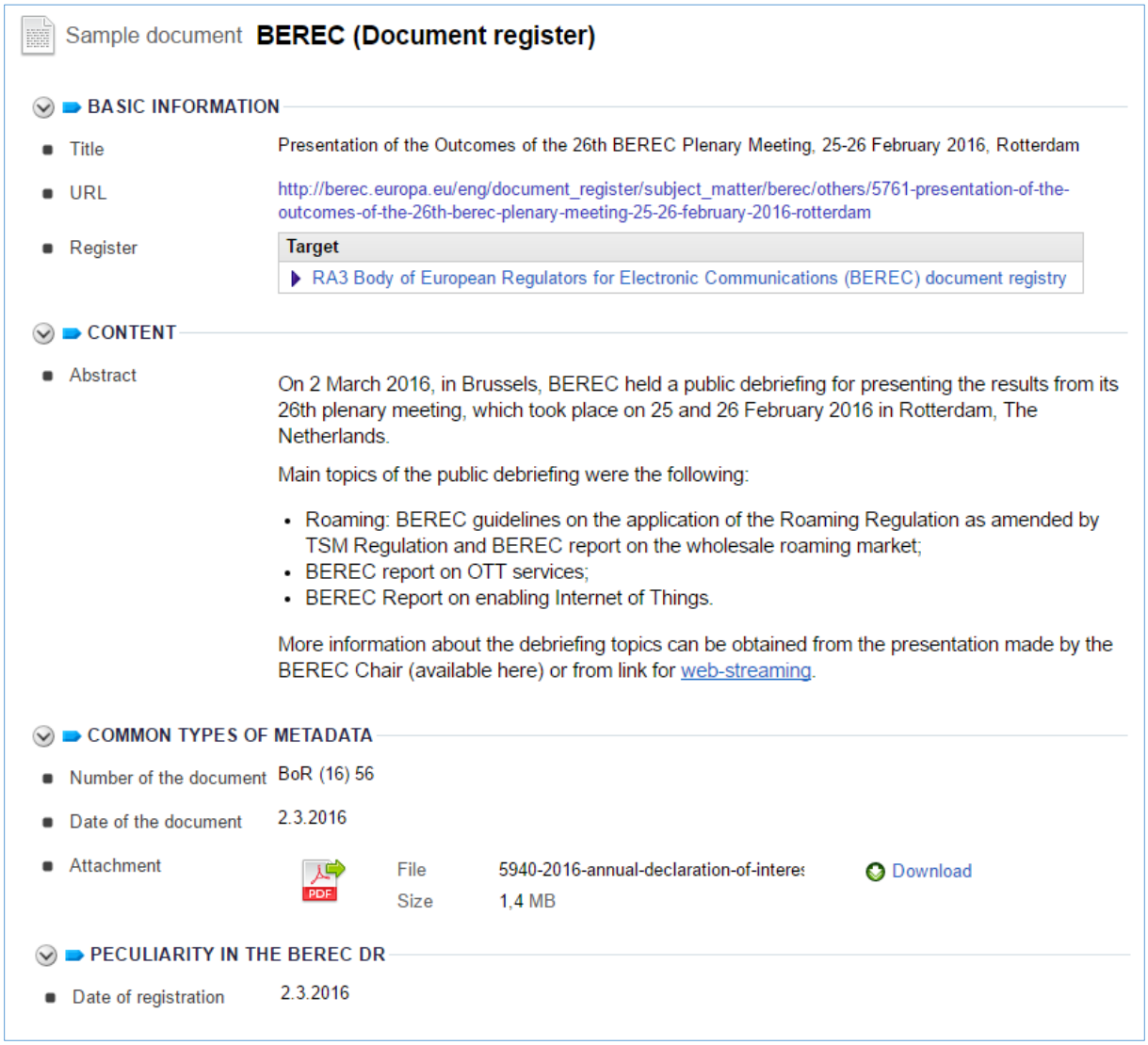


Figure 44: Sample document from the BEREC Document Register

[214] Sample document from the RD-BEREC in the unified study database structure: http://atom.ts-publicaccess.eu/form/item?ItemId=1091077.

4.3.2.1.1.9. Re-use of the RD-BEREC in view of integrated access solution

The RD-BEREC does not cover enough metadata necessary for answering basic PublicAcess.eu project questions (WHAT, WHEN, WHO, WHY).

QUESTION	ANSWERS BY EXISTING VOCABULARIES or ATTRIBUTES
WHAT **(is the document about)**	Document types Topics not available
WHEN **(was the document published)**	Dates
WHO **(is responsible for the document)**	Authors
WHY **(purpose of the document)**	Not available

Although the RD-BEREC works well from the end-user's point of view, it is quite difficult to envisage that it could be used in any of the integrated access solution alternatives.

The main disadvantage for its re-use is the absence of machine readability solution (API, etc.) Additionally, the design of the RD-BEREC is implemented towards BEREC processes (PDF document attachments are not self-explanatory and could not be used without the rest of the information provided in the document detail, especially the abstract).

4.3.2.2. European Union Intellectual Property Office (EUIPO)

The European Union Intellectual Property Office ('EUIPO') is the EU agency responsible for managing two important vehicles for the protection of creativity and innovation – the Community trade mark and the registered Community design.[215]

4.3.2.2.1. EUIPO website general information

The EUIPO has a very user-friendly website exhaustively covering all end-user needs. It is accessible at this general URL address: https://euipo.europa.eu/ohimportal/en/home.

It contains the following four information sources applicable to the study:

1. **eSearch plus**: Comprehensive information about trademarks, designs, owners, representatives and bulletins.
2. **eSearch Case Law**: EUIPO decisions, judgments of the General Court, Court of Justice and national courts.
3. **EuroLocarno**: Classification and terms for the naming of products in each of the official EU languages.
4. **Verify certified copies**: Entry of an Identification Code (ID) allows a CTM certificate to be viewed.

Sources 1, 3 and 4 provide dynamic information generated from the database so they will not be analysed.

The 2nd source **eSearch Case Law** contains documents on decisions and judgements and will be further analysed in a **general** way.

4.3.2.2.2. EUIPO Case Law Register general information

The EUIPO Case Law Register ('EUIPO-CLR') contains the following independent sections:

1. Trade mark decisions
2. Design decisions
3. National courts' judgements
4. Preliminary rulings

Functionalities of the different sections are similar, so they will be investigated together.

4.3.2.2.2.1. Public access

The EUIPO-CLR is accessible at this general URL address:
https://euipo.europa.eu/eSearchCLW.

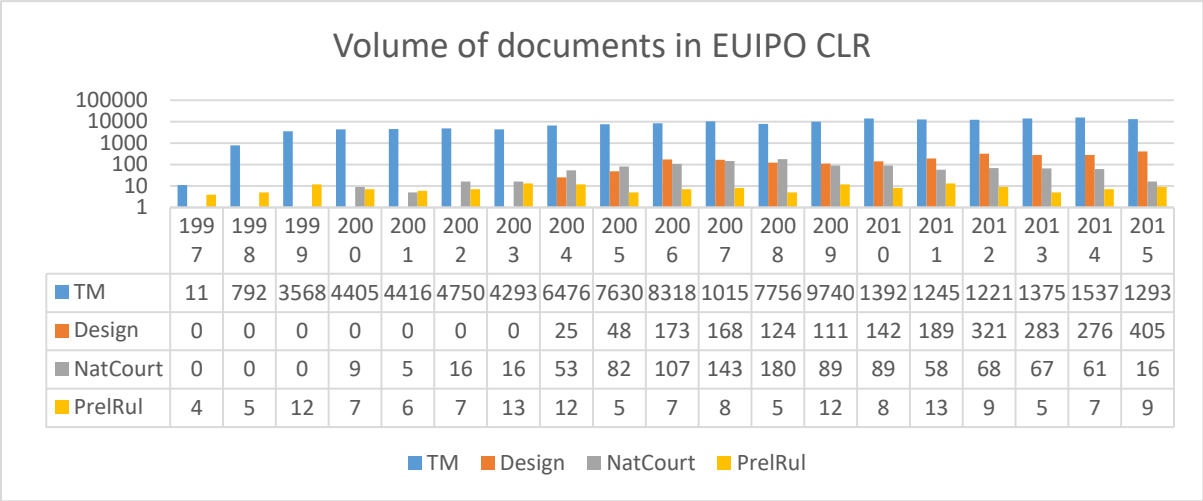
The published documents are accessible to the general public without any restrictions.

4.3.2.2.2.2. Overall volume of published documents and time range covered

The number of documents published up to the end of 2015 is 156 469. Different language versions of the documents are not taken into account.

The distribution of the documents in sections by year is shown in Figure 45. The vertical axis, displaying the volume of documents uses a logarithmic scale.

[215] Source of the description of the EUIPO mission: https://euipo.europa.eu/ohimportal/en/about-ohim.



	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
TM	11	792	3568	4405	4416	4750	4293	6476	7630	8318	1015	7756	9740	1392	1245	1221	1375	1537	1293
Design	0	0	0	0	0	0	0	25	48	173	168	124	111	142	189	321	283	276	405
NatCourt	0	0	0	9	5	16	16	53	82	107	143	180	89	89	58	68	67	61	16
PrelRul	4	5	12	7	6	7	13	12	5	7	8	5	12	8	13	9	5	7	9

Figure 45: Volume of published documents in EUIPO Case Law Register

4.3.2.2.2.3. Brief investigation of the EUIPO-CLR

Figure 46 shows the result of our comparative analysis of the EUIPO-CLR[216].

Figure 46: Overview investigation of the EUIPO Case Law Register

[216] Brief investigation of the EUIPO-CLR in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1000144.

4.3.2.2.2.4. EUIPO-CLR Document types

The Document types in the EUIPO-CLR are basically identical to the sections. The trade mark decisions and Design decisions both contain subtypes. As a result, there is the following simple hierarchy of document types in the EUIPO-CLR[217]:

- Trade marks decisions
  - Examination decisions
  - Cancellation decisions
  - GC/CJEU judgments
  - Opposition decisions
  - Board of Appeal decisions
- Design decisions
  - Invalidity decisions
  - GC/CJEU judgments
  - Board of Appeal decisions
- National courts' judgements
- Preliminary rulings

The document type 'Preliminary rulings' includes some overlaps with InfoCuria.

The relationship between a document and entries in the vocabulary of EUIPO Document types has 1:1 cardinality. This means that each document must be of only one document type.

4.3.2.2.2.5. Metadata in the EUIPO-CLR

The metadata structure of the EUIPO-CLR is the most detailed of all the registers examined in the study. It provides sufficient inputs for an in-depth analysis. However, a detailed analysis would lead to this chapter being unfeasibly long. Therefore, the analysis of metadata in EUIPO-CLR will be presented only summarily and include all Document Types:

- Total number of vocabularies: 47
  - Out of which 16 vocabularies are used in more document types.
- Total number of metadata attributes: 85
  - Out of which 14 are used in more document types.

The data model of the EUIPO-CLR is backed by a detailed ontology.

This ontology was inserted in the study database:

1. In a simplified tree structure (i.e. without internal/external relations)[218]
2. In the form of a clear picture[219]

[217] EUIPO-CLR Document types hierarchy in the study database: http://atom.ts-publicaccess.eu/form/item?ItemId=1091941.

[218] EUIPO-CLR ontology in the simplified tree structure in the study database: http://atom.ts-publicaccess.eu/form/group?TreeId=100353

[219] EUIPO-CLR ontology as a figure in the study database: http://atom.ts-publicaccess.eu/form/imagedownload.ashx?ImageId=57.

4.3.2.2.2.6. End-user search possibilities

This section analyses the search options that are available to the end-user in the EUIPO-CLR. It covers three main areas in line with the section 4.1, i.e. the search form, the list of results, and the document detail along with a brief conclusion.

4.3.2.2.2.6.1. Search form

Search form possibilities (and also the metadata shown in the lists of results) depend on the 1st document type level. This means that they are different for each of the main document types. This is another sign of the existence of the ontological basis.

There are two types of search forms:

- Basic search is a simple search with
  - Date range filtering
  - Language filtering
  - Scope for a textual search
- Advanced search for each 1st level Document type
  - Four independent search forms
  - Wide possibilities for search throughout the whole ontology (see figure in footnote 219), vocabularies, attributes with many possibilities for combination

The use of more than one criterion is handled as a chain combined with logical AND/OR/NOT operators.

4.3.2.2.2.6.2. List of results

As stated in the chapter above, there are separate lists of results for each 1st level document type according to following structure:

- Trade marks decisions
  - Case number
  - 2nd level document type
  - Date
  - IP right
  - Trade mark number
  - Trade Mark name
  - Preview
  - Link for document(s) download
- Design decisions
  - Case number
  - 2nd level document type
  - Date
  - Design number
  - Nickname
  - Preview
  - Link for document(s) download
- National courts judgements
  - Country
  - Court name
  - Court instance
  - Date
  - IP Right

    - Parties
    - Link for document(s) download
- Preliminary rulings
    - Date
    - Case number
    - IP Right
    - Parties
    - Nickname
    - Keywords
    - Link for document(s) download

Entries in the list of results are sortable by the metadata that has been selected.

4.3.2.2.2.6.3. Document detail

There are no document details available in the lists of results. The document is directly downloaded after clicking on the language version.

4.3.2.2.2.6.4. General evaluation of the Search functionality

The search functionality works very well. Everything is in right place and works quickly. The two search forms cater to both casual users and professionals perfectly well.

4.3.2.2.2.7. Re-use of the EUIPO-CLR in view of integrated access solution

It is important to say that the EUIPO-CLR works so well because its functionality is closely aligned with the decisions about intellectual property and all the other information of the EUIPO website. Any other application without these dedicated functionalities could not bring such benefits to users.

However, it is quite difficult to imagine that the documents from the EUIPO-CLR could be used in any of the integrated access solution. The main reasons for this are:

1. Absence of any machine readable format or method for retrieving both the documents and the metadata.
2. Exhaustive coverage by the metadata (47 vocabularies and 85 attributes, see 4.3.2.2.2.5) from which no relevant subset covering the content area could be used.

4.4. Conclusions from the investigation of the current situation

This chapter summarizes the results of the analysis. These summaries were written mainly with the formulation of options for the possible future integrated access solution alternatives in mind (discussed in chapter 6).

All results were then presented in an interactive way on the study database environment described above. The study database is filled with a large amount of information that could be further utilized in various ways, e.g. to create the architecture and to design the selected solution. All of the information contained in this chapter is based on different issues pertaining to the study database.

This chapter consists of the following subchapters:

1. *Conclusions from the brief investigation* provides an overview of all analysed document sources in an interactive form.
2. *Number of document sources by institutions* provides the basic statistical overview of the number of documents. It may indicate their benefits for the future integrated access solution.
3. *Vocabularies summary* provides a general overview and analysis of the metadata used by a specific document source in the form of vocabularies.
4. *Attributes summary* provides detail on and analysis of the metadata captured in the form of specific values of variables, i.e. attributes.
5. *Re-use summary* looks at the technical feasibility of the various document sources for the automatic reading of the content and context of the documents. Only the document sources indicated as being suitable for re-use in the future integrated access solution are evaluated here.
6. *Sample document* provides insights into the fundamentals of the future solution, by presenting a sort of universal document. It describes which metadata can be considered as shared and, conversely, which are dedicated to specific document sources.
7. *General conclusion* then provides a summary to all the chapters above.

It should be noted that, as well as an analysis of each document source, this chapter also includes the views and opinions of the authors resulting from:

- The premises defined in the beginning of study
- Detailed knowledge of information systems similar to those described in the chapter ‘Integrated access solution’ and long-term practical experience in their design, architecture and implementation.

All statements in the analyses of the document sources and conclusions that follow are adequately documented and can be proven.

4.4.1. Conclusions from the brief investigation

There were 27 document sources analysed in total - 21 document sources originally established in the specifications and 6 additional document sources found to be significant during the main analysis. The results of the analysis were described in detail in the previous chapters.

The overview of findings for each individual document source is summarized in the chapter entitled 'Brief investigation.' Each document source was profiled based on the taxonomy created for this purpose. The profiling was performed using the study database. An advantage of this approach is that it allowed an overview of all the document sources to be created. The results are described below.

Legend:

- The rows list specific criteria according to the taxonomy (in the tree structure).
- The last column shows the number of document sources which meet the criterion.

The individual items are also hyperlinks. They link to a list in the study database of the document sources that meet the given criterion. However, it should be noted that the number is only representative. In some cases, the application interface of the selected document sources was not sufficiently specific to state with certainty whether it meets the given criterion or not.

- INFORMATION REPOSITORY (27)
  - Provision of documents (27)
    - Documents spread across website (2)
    - Documents available via the special search page (20)
    - Documents in the special website section (e.g. register, library) (4)
    - Documents only as a 'side effect' of the other functionality (1)
  - Complementary features/applications (7)
    - More standalone website sections (1)
    - Special database (2)
    - More special databases (2)
    - Process & workflow database (2)
- ACCESSIBILITY OF DOCUMENTS (27)
  - Way of accessibility (27)
    - Documents available to general public (27)
    - Certain documents available upon request (6)
  - User possibilities (9)
    - Update subscription (2)
    - Public login available (4)
      - Subscription of the 'News' (3)
      - Saving searches (3)
    - Restricted login (3)
    - Login with ECAS account (2)
  - Machine readability methods (14)
    - Public API (2)
    - RSS (12)
    - Datasets on the Open data portal (2)
    - XML Dump (2)
  - Social networks used for broadcasting information (9)
    - YouTube (6)
    - Facebook (4)

- Twitter (8)
- LinkedIn (5)
- Google+ (3)
- Flicker (1)
- Amount of documents (26)
  - less than 1 000 (5)
  - 1 000 - 9 999 (6)
  - 10 000 - 99 999 (10)
  - more than 100 000 (5)

AVAILABILITY OF DOCUMENTS (27)

- Document types (24)
  - Not identifiable (5)
  - More document types available (19)
  - Hierarchical organisation (6)
- Metadata capturing methods (25)
  - No metadata identified (2)
  - Vocabulary groups (23)
    - Document types (18)
    - Authorities (17)
    - Topic (16)
    - Years (or other periods) (12)
    - Languages (8)
    - Process information (7)
  - Attribute groups (19)
    - Dates (18)
    - Authors/Authorities (2)
    - Topics/Descriptions (3)
    - Various codes (10)
    - Various numberings (12)
    - Various references (8)
    - Various process information (4)
    - Other (e.g. descriptions, formats, etc.) (7)
  - Relations (16)
    - Internal relations (13)
      - Related content within the website (13)
    - External relations [between registers] (9)
      - EUR-Lex (8)
      - Commission (2)
      - EP Register of Documents (1)
- Languages (27)
  - User interface languages (21)
    - All EU languages (15)
    - Only one language (2)
      - English (2)
    - Only certain languages (4)
      - English (4)

- French (2)
- German (1)

Document languages (26)
- All EU languages (11)
  - Selected documents (8)
- Only one language (2)
  - English (2)
- Only certain languages (16)
  - English (5)
  - French (4)
  - German (1)

Formats (27)
- PDF (25)
- Word (12)
- Excel (2)
- PPT (1)
- ePub (2)
- webpage (10)

END-USER SEARCH POSSIBILITIES (27)
- Search possibility not available (1)
- Whole website search (2)
- Empty search possible (17)
- Document content indexed (17)
- Various documents listings (2)
- Search form (24)
  - Simple search form (11)
  - Advanced search form (20)
  - More search forms (6)
- List of results (25)
  - (Simple or advanced) list of links to results (5)
  - Results equipped with content abstracts (3)
  - Facet filtering of the results available (15)
    - by document type (10)
    - by date/year (10)
    - by authority (11)
    - by topic/tag (8)
    - by country (1)
    - by language (5)
  - Sorting of results possible (11)
    - alphabetically (2)
    - by document type (3)
    - by date (9)
    - by authority (4)
    - by country (1)
    - by number/code (7)
    - by relevance (3)

- Document detail (19)
  - Not available, but covered by the list of results (7)
  - Content + metadata in a (set of) webpage(s) (4)
  - Metadata + content as file attachment(s) (8)
  - Metadata + link to content placed in another source (1)
- OTHER RELEVANT ASPECTS (25)
  - Documents duplicated from their primary sources (Overlaps) (4)
    - Register of Commission documents (3)
    - Duplicates from the same website (1)
  - Technology (CMS and/or Search engine) (25)
    - Not identifiable (12)
    - Microsoft SharePoint (6)
    - Adobe ColdFusion (2)
    - Commission CWMS (2)
    - Jahia (2)
    - Google Search (1)
    - Apache Solr Search engine (2)
    - Verity Autonomy (1)
- COMMON END-USER EXPERIENCE (22)
  - 1 Basic (as book reading)
  - 2 Below average (abstruse, but working) (2)
  - 3 Average (good with compromises) (5)
  - 4 Above average (close to perfection) (4)
  - 5 Excellent (complex, fast, intuitive) (11)
- RE-USE IN INTEGRATED ACCESS SOLUTION (20)
  - Vocabularies groups (18)
    - Document types (15)
    - Originators (15)
    - Topics (12)
    - Years (12)
  - Attribute groups (18)
    - Dates (18)
    - Process stage description (11)
  - Machine readability (20)
    - Not available (11)
    - API (3)
    - RSS (8)
    - XML (2)
  - Re-use conclusion (20)
    - * (NOT RELEVANT for some serious reason) (9)
    - ** (Very laborious, but POSSIBLE at the end) (4)
    - *** (API or RSS: NO, Metadata set: PARTIAL) (2)
    - **** (API or RSS: NO, Metadata set: FULL) (3)
    - ***** (API or RSS: YES, Metadata set: FULL)) (2)

4.4.2. Number of documents by institutions

Table 21 summarizes the number of documents by institutions/agencies and by specific document sources. It includes only the primary language versions, not the translations. The numbers do not reflect the public documents available on request.

The purpose of this table is to generally outline the scope in which the future integrated access solution would be carried out.

Institution/Agency	Document source	Documents total	
European Parliament	Register of documents	585 000	**679 000**
	Legislative Observatory	14 000	
	IPEX	60 000	
	Website	20 000	
Council of the European Union	Register of Council documents	350 000	**379 000**
	CASE	27 000	
	Council database of agreements and conventions	2 000	
European Commission	Register of Commission documents	60 000	**115 000**
	Comitology register	48 000	
	Register of Commission expert groups	7 000	
Court of Justice of the European Union	Register of Case Law - cases	35 000	**123 000**
	Register of Case Law - documents	88 000	
European Central Bank (website)	Research and Publications	11 000	**11 000**
European Court of Auditors (website)	Audit reports and opinions	5 000	**5 000**
European Economic and Social Committee	Register of documents	37 000	**37 000**
Committee of the Regions	Register of documents	27 000	**48 300**
	Members Portal	1 200	
	Studies/Brochures register	100	
	General website	20 000	
European Ombudsman	Register of case	5 000	**6 000**
	Register of resources	1 000	
Publications Office	EUR-Lex	900 000	**3 100 000**
	TED	2 200 000	
The Education, Audiovisual and Culture Executive Agency	Document Register	1 000	**1 000**
European Research Council	Document Library	320	**320**
The Body of European Regulators for Electronic Communications	Document Register	4 000	**4 000**
The Office for Harmonization in the Internal Market	eSearch Case Law - Trade marks decisions	157 000	**157 000**
Total sum:			**4 665 620**

Table 21: Number of documents by institutions

4.4.3. Vocabularies summary

Table 22 includes a total of 149 vocabularies used in the document sources of the institutions and agencies investigated.

However, this analysis is not complete because in some cases the web applications using the vocabularies do not allow (either partially or completely) the extraction and analysis of individual vocabularies.

Therefore, this chapter should be seen more as an overview of vocabularies than as the definitive list. All inputs are available in the study database.[220]

Institution	Vocabularies investigated
European Parliament	13
Council of the European Union	6
European Commission	16
Court of Justice of the European Union	14
European Central Bank	3
European Court of Auditors	3
Committee of the Regions	10
European Economic and Social Committee	5
European Ombudsman	14
Publications Office – EUR-Lex	47
Publications Office – TED	14
Education, Audiovisual and Culture Executive Agency	0
European Research Council Executive Agency (ERC Executive Agency)	1
Body of European Regulators for Electronic Communications (BEREC)	1
European Union Intellectual Property Office (EUIPO)	2
Total:	149

Table 22: Total number of analysed Vocabularies by Institutions/Agencies

All entries of the investigated vocabularies were inserted into the study database where:

- Entries are assigned to the specific document sources.
- Entries are structured into nine logical groups[221], mainly according to their designation and the purpose for which they are used.

The results are summarized in the following tables.

The total number of specific Vocabulary groups arranged separately for EUR-Lex and for all the other institutions/agencies are included in Table 23.

Vocabulary Type	Vocabulary description	EUR-Lex	Other than EUR-Lex
Document Type	Vocabularies specifying the types of documents, in many cases 'document type' also contains information about its reason for being or process stage etc.	2	18
Authority	Vocabulary describing Originator or Origin of the document (Author, Department, Rapporteur, Unit, Plenary session etc.)	14	27

[220] List of all analysed Vocabularies in the study database: http://atom.ts-publicaccess.eu/form/class?ClassId=100171.

[221] Vocabularies arranged by their types in the study database: http://atom.ts-publicaccess.eu/form/option?AtomId=100357.

Topic	Vocabularies specifying of what the documents are about, the themes or areas of interest covered (subject matter, policy area, field of law, classification etc.)	5	25
Year/Time	Vocabularies describing the time periods (Years, Parliamentary terms, Months etc.)	1	13
Language	Helping vocabularies of languages specify in which languages the documents exist	1	5
Place/Location	Localisation aspects describing place of origin of the document, place of its signature etc.	1	0
Process information	These vocabularies are very dependent on the source and typically specify process stage in which the document was produced	14	12
Container	This type of vocabulary is very diverse and dependent on the document source and serves as an organisational entity	0	15
Helpers	Assistance vocabularies like types of addresses, types of structures in legislation acts etc.	7	0
Not identifiable	It was not possible to assign the type of Vocabulary type	2	0
	Total:	**47**	**114**

Table 23: Sums of Vocabularies by logical Vocabulary type

The Vocabulary types 'Document Type', 'Authority' and 'Topic' are considered important and useful from the perspective of the future integrated access solution. They provide answers to the questions WHAT (what is the document about), WHO (who is responsible for the document) and partially WHY (what is the reason for the document). The prospective users of the integrated access solution would be able to find the answers to these questions through the document search.

The Vocabularies Year/Time and *Language* are considered only as support for the following reasons:

- Year/Time vocabulary is clearly also important because it responds to another fundamental question: WHEN (when the document was published). It can always be easily obtained through this.
- Through the vocabulary Language other language versions of a document can always be accessed.

The vocabulary type *Process information* and *Container* are always tied to a specific application and they are not applicable without them.

The vocabulary types *Place/Location, Helpers* and *Not identifiable* were identified only in EUR-Lex.

Table 24 shows the coverage of the Vocabulary type (Document Type, Authority, Topic) by the individual web applications of the institutions and agencies. In addition, the presence of groups of three important vocabularies in the source document is indicated in separate columns.

Institution/Agency	Document source	Document Type	Authority	Topic
European Parliament	Register of documents	*	*	*
	Legislative Observatory	*	*	*
	IPEX		*	*
	Website	*	*	*
Council of the European Union	Register of Council documents	*		*
	CASE	*		
	Council database of agreements and conventions			
	Website	*		*
European Commission	Register of Commission documents	*	*	
	Comitology register	*	*	*
	Register of Commission expert groups	*	*	*
Court of Justice of the European Union	Register of Case Law – cases			
	Register of Case Law – documents	*	*	*

European Central Bank	Research and Publications	*	*	*
European Court of Auditors	Audit reports and opinions	*		*
European Economic and Social Committee	Register of documents	*	*	
Committee of the Regions	Register of documents	*	*	
	Members Portal		*	
	Studies/Brochures register	*		*
	General website	*		
European Ombudsman	Register of cases	*		*
	Register of resources	*	*	*
Publications Office	EUR-Lex	*	*	*
	TED	*	*	*
The Education, Audiovisual and Culture Executive Agency	Document Register			
European Research Council	Document Library	*		
The Body of European Regulators for Electronic Communications	Document Register	*		
The Office for Harmonization in the Internal Market	eSearch Case Law - Trade marks decisions	*	*	*

Table 24: Usage of vocabulary groups in document sources

Notes:

- The more groups of vocabularies the source document uses, the more it is prepared for the integrated access solution.
- In some cases, the column Authority is empty, but this does not necessarily signify an absence of this meta-information because each institution may be taking the role of Authority - especially in the case of the agencies.

Table 25 shows the number of entries in each Vocabulary type:

- EUR-Lex is excluded because it largely contains entries already contained in other vocabularies.
- Vocabularies are not defined further by the individual institutions (as is the case for MDR used by EUR-Lex for instance).
- EuroVoc is also not included, because it is described in detail.

Vocabulary type	Entries in Vocabularies (EUR-Lex excluded)
Vocabulary type: Document type	550
Vocabulary type: Authority	6 547
Vocabulary type: Topic	16 910
Total	24 007

Table 25: Sum of entries in Vocabulary types Document Type, Authority, Topic

Conclusions from the investigation of the Vocabularies

The main problem is that the Vocabularies and their entries are not specifically defined (e.g. as in the case of the MDR used in EUR-Lex).

The custom names of the individual entries are not sufficiently informative. Although in many cases they are identical or very similar, it cannot be stated with certainty that the individual document sources cannot have the same or at least similar meanings. The same problem may be the fact that different vocabularies as a whole describe an identical type of data, but their entries are different (for example the vocabulary 'Types of document' from EUR-Lex which is used for assigning Celex numbers seems to be the same as the vocabulary 'Celex category' used in IPEX, as they describe the same issue. However, they differ greatly.) Almost all the vocabularies are also composed as flat lists which do not allow them to have a clear connection to their superiority and inferiority.

The integrated access solution would require a large amount of documentary work. The framework of this documentary work is outlined in the chapter 'Integrated access solutions'.

4.4.4. Attributes summary

Table 26 includes a total of 61 attributes used in the document sources of the EU institutions and agencies which were investigated.

However, this analysis is more of an overview, because the metadata attributes are designed for use in the view of each institution's or agency's document source process. Additionally, in many cases it was not technically feasible to find out all the attributes and only the information received directly from the institution's representatives was helpful.

Therefore, this chapter should be regarded as an illustrative overview of attributes used rather than as the ultimate list. All inputs are available in the study database.[222]

Institution	Vocabularies investigated
European Parliament	17
Council of the European Union	4
European Commission	14
Court of Justice of the European Union	6
European Court of Auditors	3
Committee of the Regions	7
European Ombudsman	6
Body of European Regulators for Electronic Communications (BEREC)	4
Total:	61

Table 26: Sums of analysed attribute types by Institutions/Agencies

All investigated attributes were inserted into the study database where:

- Attributes are assigned to the specific document sources;
- Attributes are structured into seven logical groups[223] included in Table 27 according to their designation and purpose.

Institution/Agency, document source	Date	Code	Number	Authority	Process	Reference	Other
European Parliament Register of documents	*	*	*			*	
Council of the European Union Register of documents	*		*			*	*
European Commission Register of Commission documents	*		*		*	*	
European Commission Comitology register	*	*	*		*	*	*
Court of Justice of the European Union Register of Case Law	*	*	*				*
European Court of Auditors Register of publications	*	*					
Committee of the Regions Document Manager	*	*	*		*	*	*
European Ombudsman Register of Cases	*	*					
Body of European Regulators for Electronic Communications (BEREC) Document registry	*		*	*			

Table 27: Metadata attributes grouped by common purpose

[222] List of all analysed Attributes in the study database: http://atom.ts-publicaccess.eu/form/class?ClassId=100359.

[223] Presence of Attributes in specific document sources based on their types in the study database: http://atom.ts-publicaccess.eu/form/option?AtomId=100364.

Apart from the meta-information from the group 'Date', all other groups of metadata attributes can be considered dedicated to a specific application.

Information about the time parameters of documents can be reconstructed from the attribute group Date. This would be useful for the prospective users of the integrated access solution, i.e. to know WHEN.

The attributes in this group can be easily re-used from a technical perspective. However, similarly to the Vocabularies, a large amount of documentary work would be necessary for the integrated access solution. The framework for this documentary work is outlined in the section Integrated access solution.

4.4.5. Document sources re-use possibilities in view of the integrated access solution

This section summarizes the re-use possibilities of the document sources in terms of their machine readability, which is very important from the perspective of feasibility for the future integrated access solution.

Table 28 summarizes the machine readability of the document sources, which is based on the detailed analysis definition stated in chapter 4.1.

Institution/Agency/Body	Document source	IMMC	API	RSS	Other
European Parliament	Register of documents				Limited or EUR-Lex
Council of the European Union	Register of Council documents				Limited or EUR-Lex
European Commission	Register of Commission documents			*	EUR-Lex
	Comitology register				
Court of Justice of the European Union	Register of Case Law - documents				EUR-Lex
European Central Bank	Research and Publications			*	
European Court of Auditors	Audit reports and opinions			*	
European Economic and Social Committee	Register of documents			*	
Committee of the Regions	Register of documents			*	
European Ombudsman	Register of case				
Publications Office	EUR-Lex	*	*	*	*
	TED		*	*	

Table 28: Document sources machine readability options

Table 28 demonstrates that the opportunities for machine readability offered by the listed document sources are very limited. Based on this fact, it can be concluded that improvements in this area would require a lot of effort in the document sources' operators and administrators.

4.4.6. Common and dedicated metadata

The final part of the analysis of each document source is the chapter titled 'Sample document'. It includes a typical example of a document from each document source together with its metadata. The metadata are divided into those that can be considered common and those that are dedicated to a particular document source.

The purpose of the chapter Sample document is to provide a bird's-eye view of the vision for a 'common document' which would then be presented in the integrated access solution. It should primarily determine whether such a common view is feasible.

A common document should provide answers to the following questions:

- WHAT (what is the document about);
- WHO (who is responsible for the document);
- WHY (what is the reason of the document);
- WHEN (when the document was published).

Table 29, on the next page, shows the metadata which are included in all the document sources (left side of the table), as well as the specific metadata for each document source (right side of the table). These latter must be taken into account for the overall profile of the document.

Table 29 includes only the document sources which meet the basic definition criteria for the document source stated in chapter 4.1.

Finally, the creation of such a common document, which could be presented in the future integrated access solution, is feasible. However, a substantial amount of work both from the operators/administrators of the document sources and from the service provider of such a solution (integration and development services) would be required.

BASIC INFORMATION	
	Title
	Register
CONTENT	
	Abstract
	URL to content
COMMON TYPES OF VOCABULARIES	
	Year
	Originator
	Topic(s)
	Type(s)
	Language(s)
	Procedure
	Successor(s)
	Predecessor(s)
COMMON TYPES OF ATTRIBUTES	
	Number of the document
	Date of the document
	Date of the meeting
	Attachment
COMMON INTERNAL RELATIONSHIPS	
	Part of a dossier

PECULIARITY IN THE EUROPEAN PARLIAMENT REGISTER OF DOCUMENTS	
	Date of entry
	Date of event
	EP other references
	EuroVoc
	Parliamentary term
	Author(s)
PECULIARITY IN THE PUBLIC REGISTER OF COUNCIL DOCUMENTS	
	Interinstitutional file
	Addressee(s)
PECULIARITY IN THE EUROPEAN CENTRAL BANK	
	Note
	Author(s)
PECULIARITY IN THE EUROPEAN COURT OF AUDITORS REGISTER OF PUBLICATIONS	
	Originator
	ECA free text
	ECA identifier
	ECA keywords
PECULIARITY IN THE INFOCURIA	
	Documents in case
	Subject matter
	References
	Date lodging
	Parties
	Academic writing
	Formation of the Court
	Judge-Rapporteur
	Procedure and result
PECULIARITY IN THE EUROPEAN OMBUDSMAN REGISTER OF CASES	
	Document structure
	Document content
PECULIARITY IN EUR-LEX	
	Amendment to
	Date of Document
	Date of effect
	Deadline
	Instruments cited
	Date of effect
	Deadline
PECULIARITY IN THE TED	
	Award criteria
	Contract
	Deadline
	Deadline for the request of documents
	Directive
	Document sent
	Edition number of Supplement to the Official Journal
	EU Institution
	Heading
	Name of the awarding authority
	Place
	Procedure
	Publication date
	Regulation
	Type of awarding authority
	Type of the Bid
	URL of the awarding authority
	Country
	CPV code
	Original CPV code
	NUTS code

Table 29: Common/dedicated metadata in all/particular document sources

4.4.7. General conclusions

All analysed document sources are taken from the europa.eu domain.

Therefore, it is evident that europa.eu represents a comprehensive information ecosystem. This ecosystem is very granular, diversified and decentralized. Despite various partial solutions and lists of resources no overall map describing the individual parts could be found. Unfortunately, some parts of the ecosystem are well hidden and the users discover them either accidentally or through a Google search – but only if they know what they are looking for.

Providing the information contained within the documents (see the definition of a document in the view of the study in the chapter 4.1) is just one part of this ecosystem. Much of the information is released during the course of its development, subject to change and various refinements. These pieces of information cannot, therefore, be regarded as documents.

The diversification of the europa.eu ecosystem is well known. Each institution/agency is autonomous and responsible for conducting its own processes and so it provides information in the technical format of its choice. On many occasions, it is clear that the institution/agency publishes the information as an extract from its internal systems, which includes even more metadata (vocabularies, attributes, relations).

This study is focused on these documents, i.e. only on a part of the complex information ecosystem of europa.eu.

The analysis of the document sources clearly showed that the usual approach to document publication separates the content and the context of the documents.

The most commonly applied method is the direct access to the document context (with some exceptions, e.g. InfoCuria and EO-RC). The URL is a unique identifier while the document's context is caught in HTML notation. HTTP is used to locate and access the document through the URL. The content of the document is accessed through its context. In most cases the content is attached as a file attachment in PDF and DOC formats, both of which are difficult for machines to read. However, sometimes the document's context (i.e. metadata) is displayed on the lists of results.

The document is represented as a site with the document metadata in the form of a unique address, which includes a link to the content of the document in the form of an attachment.

The reasons for this approach are clear: the presentation of the content in the form of a file attachment is the most cost-effective method whilst ensuring its durability without changes.

On the other hand, the popularity and extensive use of the commercial legal systems may be explained because of the advanced presentation of the content and the lack of separation between context and content.

However, this method of connecting the content as an attachment to the context can be considered a major limitation on further development long-term. For example, it does not allow the content of a document to be worked with effectively, e.g. to target parts of the document through hyperlinks or to come closer to the Linked Open Data principles (and to change from the web of documents to the web of data).

It can be concluded that EUR-Lex is the most advanced system in terms of its presentation (both of context and content). Every user can display the form that is most appropriate for them. It has a good solid foundation to improve even further towards advanced content structuring (Formex).

PublicAccess.eu is an integration project. Its aim is to look at the documents from one common system. In order for the users to benefit from the results of PublicAccess.eu, it would be necessary to unify and aggregate metadata and perhaps follow a new approach to the content. Naturally, it can be divided into several phases. The different approaches and possibilities are outlined in Chapter 6 'Integrated access solution'.

5. Integrated access from practical point of view

This chapter provides two different views on the problematics of integrated access to EU documents, both based on real-life experience. The first subchapter describes the outcomes from the attempts of the general public to find the EU documents from a particular legislative process. The second subchapter summarizes the experience of the authors of the study gained on the project under a contract with the Publications Office: "PublicAccess.eu – Conceptualisation, design and development of an on-line showcase" ('PublicAccess.eu showcase"), whose main aim was to highlight and test new, innovative ways of presenting EU legal information, on the basis of five specially pre-selected legislative acts; within this project, different types of EU documents from different document sources with the different metadata were gathered into one common place and are presented via a single website.[224]

5.1. Conclusions from the users' point of view

As part of the study, several registers, documents sources and even websites of the European institutions and agencies were analysed (as described in the previous chapters). The analysis was carried out by experienced users, who are generally familiar with the existing document databases at the European level. When designing a new integrated access solution, there is another important point to consider. It is the viewpoint of the 'general public' that all future integrated access solution alternatives should be accessible to this group.

A group of people from different EU countries were randomly selected. They were asked to search for particular documents. The outcome of this public survey contains interesting results and this chapter sums up the results and conclusions drawn from the practical searches performed by these participants.

Who were the respondents?

When looking for the respondents, the aim was to cover several areas of their activities, where they had an interest in EU documents. Therefore, the respondents found included law students (Palacký University Olomouc, Czech Republic), teachers, civil servants working for a national Ministry, certified experts on public procurement, etc. The goal was not to undertake a huge EU survey but to get feedback from a few real people who would take the search seriously.

What was the task?

The respondents were asked to search for all documents relating to the life cycle of one legislative act of the European Union. They were provided with a set of basic acts and asked to select one (but they were allowed to select any other act as well). They were asked to search for as many documents regarding the legislative process of the chosen act as possible (such as a proposal, other accompanying documents, final publication, etc.) The respondents did not receive any other instructions or guides as to where to search, the kind of document sources available or any directions on the websites that may contain relevant documents. The reason was so as not to influence their search approach.

What are the results and conclusions?

At prima facie, the results of the searches were very favorable. Even the respondents unfamiliar with the EU legal system were able to find a lot of documents related to their selected act. The results implied that a combination of currently available search tools could lead to a positive search even if the integrated access solution would not be implemented.

[224] http://showcase-publicaccess.publications.europa.eu/.

However, such a conclusion could be oversimplified. It must be noted that the participants were able to find information which was already connected together in some way or which are collected in one place.

The method of obtaining the documents (information) was similar. Law students had the advantage that they had access to various commercial databases with the legislation. They usually began their search in such databases. It should be noted that they were not very successful in obtaining the required documents (i.e. documents related to the legislative process of a particular act) from the commercial tools. The commercial tools (various commercial law databases) usually only contain the final wording of the European legislation published in the OJ.

After this unsuccessful attempt, the law students mostly switched to another way of collecting the documents. The next step was to use the Google search tool. At the same time, most of the other respondents started with Google search as their first option.

The respondents searched keywords related to their chosen act in their own language and in English, and almost always found links to the Act on EUR-Lex among the first results.

Therefore, it may be determined that in terms of Google ranking, EUR-Lex has an irreplaceable position. It is rated as the most relevant source for information on EU legislation and it reliably displays the searched document at the top. But the fact that the ordinary respondents without the legal background and even the law students did not look in EUR-Lex automatically, but only indirectly through Google, should be considered for the future promotion of EUR-Lex. Only a negligible proportion of the respondents searched for the information exclusively via EUR-Lex.

This points to the fact that, although EUR-Lex is an excellent portal which provides access to the legislative information (including the only legally binding form in the OJ), a large portion of the general public are not aware of its existence, which is clearly not a good thing. One can conclude that any integrated access solution, which goal is much broader than providing legislative information, must not only be state-of-the-art technically and in terms of content, but also give priority to the promotion and marketing of its existence to the general public. On the other hand, it is necessary to use all the techniques to hand to make the integrated access solution as 'Google friendly' as possible (so that the integrated access solution would be shown amongst the top results when searching for documents just as today is the case with EUR-Lex).

It should be added that even those users who started their search directly on the EUR-Lex portal moved on to Google search to search for additional documents and did not use the opportunity to look, for example, in the Legislative Observatory or the RD-CEU, etc. (although Google search ultimately lead them to the information from the Legislative Observatory).

Again, this points to the fact that the greater the interconnectivity between the information, the better the find. Conversely, the more documents that are included in only one document source, the harder it is to find them. During the experiment, the vast majority of respondents did not search beyond the first 1-2 pages of Google search results. This shows that if the users did not find the information quickly and clearly, they were not willing to go much further. This means that if they receive the information from the interconnected resources (e.g. EUR-Lex), they were not willing to continue to search in the special systems (e.g. RD-CEU). Most likely, they felt that the document was non-existent or unavailable. For example, there are a number of documents listed in EUR-Lex in the 'Procedure' section only as a document number without a direct link to its text or source (mostly documents of the CEU). In such cases, the users did not search further for this document number, but instead automatically assumed that the document was unavailable (even though, by using the document number mentioned in the procedure it would be easily found in the RD-CEU).

The respondents eventually identified EUR-Lex (and its section 'Procedure') as the source for accessing the necessary information. They could find all the necessary information (proposal, opinions of other institutions, the final act, etc.), which were considered to be relevant for their assignment. Only a few respondents went further with their search (and used, for example, the Legislative Observatory available at the EP website, or other more specific document sources.)

The key result from the finding is that the user will only find as many documents as the EU institution and agencies 'decide' to be relevant. The number of publicly available documents that accompany the legislative process of EU acts should be counted in the hundreds. However, only a fraction of them are displayed through the 'Procedure' section in EUR-Lex (e.g. the Green Papers, various records and meeting minutes of the authorities, national scrutinies information, etc. are missing). If all these documents were made accessible in a clear and transparent form (currently in the Procedure section in EUR-Lex), the users would find all the available documents using Google search, EUR-Lex, and the Procedure section. Therefore, it is only up to each EU institution to decide how much and what information would be accessible in any integrated access sulution in a simple and interconnected manner, and how many of them would remain isolated in separated registers or hidden subpages of websites.

Additionally, the citizen survey also revealed that despite the prevailing trend of online information, there were respondents who began their search in a traditional library.

5.2. Conclusions from the PublicAccess.eu showcase

This subchapter describes the practical experience gained from the PublicAccess.eu showcase[225]. This practical experience illustrates some of the problems to solve in the future integrated access solution. The PublicAccess.eu showcase covered only the documents related to the EU legislative process. Therefore, it does not cover all the document types that are the subject of study. However, the results indicate some common issues to be solved, as there are several documents originating from many document sources which are described in the previous chapters.

The following text explains the issues that need to be considered when the documents are obtained, categorized and indexed, and it describes how the four basic questions (What, Who, Why, When) were answered for the PublicAccess.eu showcase.

5.2.1. Basic specification of the PublicAccess.eu showcase

The goal of the PublicAccess.eu showcase was to assemble the documents from the whole legislative process life cycle (i.e. from the drafting, proposing, observation, amendments and publications of the selected EU legal acts originating from the multiple document sources) in one place and to prepare a simple and transparent means by which to aggregate, visualize and search inside these documents. The documents were published on a single website with a common Content Management System.

5.2.2. Collection of documents

The documents were gathered from several different document sources. All these document sources were analysed in detail in the study (see chapter 4 of study). The PublicAccess.eu showcase proved that the creation of a common methodology, which defines the scope of documents to be part of a common database, has a substantial advantage. Following the methodology means that the process of collecting the documents will not be just gathering as many documents as it is possible to discover. Such a methodology could be extended or narrowed over time, but it should be applied uniformly for all similar documents to maintain the consistency of the common database. During the PublicAccess.eu

[225] http://showcase-publicaccess.publications.europa.eu

showcase some double-checks or triple-checks were performed to provide the most consistent database as is currently possible. Similarly, the future integrated access solution would need to focus on the methodological aspects with regards to the scope of documents in the beginning of the project.

In the case of not following the unified methodology, the user would be able to find via the future integrated access solution a certain type of document from one source (institution, agency), but the same type of document would not be provided by other sources, even though such a document type is available from the original document source.

These important findings from the analysis were verified during the collection of the documents within the PublicAccess.eu showcase. For example, some document sources are ready for automatic retrieval using specific and custom-developed tools which utilize the special conditions of each document source. In such cases, it was easy to follow the scope of the required documents and acquire them by mass download. Conversely, some document sources have proved to be unsuitable for machine readability or automated search and download of documents. In this case, the documents had to be manually retrieved one by one. Obviously, this is a very time-consuming approach requiring lots of personnel (the fewer the people involved in obtaining the documents, the longer it takes to acquire them). From practical experience, it may be concluded that the methodology for defining the scope of documents for the future integrated access solution should be accompanied by an in-depth analysis of the resources available for their future automated re-use. In other words, the methodology should accommodate, to a reasonable extent, the required scope of the documents and the technical feasibility (to avoid an excessive number of documents, which can only be obtained manually and in a multiple-year timeframe).

During the PublicAccess.eu showcase, attention was paid to the possibility of having as many language versions for each document as possible. However, the first challenge is that only a select number of documents exist in various languages. A large number of documents originating from EU activities are available only in selected languages (most frequently in English and French). The second challenge is associated with the problem described in the paragraph above – i.e. in the automated collection of specific (existing) language versions. If all the language versions are considered, the total number of documents increases significantly. The PublicAccess.eu showcase showed that without an automated method of acquiring the documents' language versions, the completion of the database is an almost impossible task (or a task taking several months or even years to complete). Therefore, an important precondition is to enable machine-based data processing and to develop various software tools to retrieve the required documents. If the document source does not allow such functionality, it is necessary to decide whether or not to acquire (existing) multiple language versions (and perhaps to accept the excessively long time and large number of personnel needed to obtain them).

5.2.3. Processing of documents

Looking at the hypothetical use of the existing document sources along with their contents, the analytical section of the study focused on four basic questions. These should become the backbone of any integrated access solution.

As already mentioned several times in the study, the answers to the following questions need to be established:

- WHAT (is the document about)?
- WHO (is responsible for the document)?
- WHY (does the document exist)?
- WHEN (was the document published)?

The PublicAccess.eu showcase also sought answers to all of these four questions and the following subchapters describe the practical experience gained during the PublicAccess.eu showcase.

5.2.3.1. WHAT (is the document about)?

It was necessary to define the number of documents to be gathered during the PublicAccess.eu showcase in terms of their content and the possibility of their further use. In this document categorization, the most attention was paid to two key components with which the user came into immediate contact – the Title of the document and the Type of the document.

Titles

The PublicAccess.eu showcase showed that the individual institutions/agencies and other various internal organisational units within the institutions/agencies have a different approach to formulating titles. A unified methodology for title formulation is applied at the EU level only for the key legislative documents (regulations, directives, decisions, etc.). Therefore the Interinstitutional Style Guide describes how to form the title[226], what the title should include and in what order the individual elements of the title are listed. This contributes to the uniformity of titles regardless of their authors. Such uniform titles are then easily used for any integrated access solution. However, the vast majority of documents do not have a unified method of creating titles. The same types of documents originating from various sources have a wide variety of names. This may not cause problems as long as the documents are made available only within a particular document source. However, in the PublicAccess.eu showcase, it was clearly confirmed that when documents of the same type but with different methodologies for creating titles are gathered in one integrated place and then provided to the user as 'one database', the user often has to sift through the mix of these documents. To provide the user with the maximum comfort when viewing his search results, most of the titles used for web presentation were manually edited for the PublicAccess.eu showcase. Almost no automated ways of data processing could be used here. It would be necessary to invest in a lot of skilled personnel to ensure the various documents from various sources have unified titles. At the same time, the title must still reflect the content of the document whilst the title editing rules cannot be completely free, and any modification of titles should follow the methodology to consistently ensure the equal treatment of the same document types. It should be also noted that these edited titles are presented (for example) in the search results or web-titles, and editing titles do not modify the documents themselves (they are kept in their original versions). However, it modifies the title as one of the metadata. All these activities together represent a very difficult task which requires a significant amount of effort to process. There were more than 1 000 documents of various types processed in the PublicAccess.eu showcase and editing their titles in one language version resulted in several dozens of man-hours. When considering this issue in the future integrated access solution, which foresees the availability of tens of hundreds of thousands of documents in all languages in which the documents are available, the total effort for manually editing these documents is difficult to imagine and to estimate. However, without title editing the overall impression from the documents provided to the users would be of lower level because the user would have to decide whether the displayed document does or does not meet their requirements, whereas it would be better to have a situation where the essential information is already provided clearly and directly in the title.

One of the alternatives, which was tested during the PublicAccess.eu showcase, was the possibility to create the title automatically through shared metadata. For example, the existing title (unmodified)

[226] Online version of the Interinstitutional Style Guide (chapter 2.1. Title) available at: http://publications.europa.eu/code/en/en-120100.htm.

could be automatically modified so that it includes the document's author/type/date. Then, the user would be able to identify the required document much more easily. However, from the end-user's perspective, this is still much less comfortable than the result of manual title modification. Eventually, the automatic modification of the titles was abandoned in the PublicAccess.eu showcase. Non-unified document titles could be partially substituted from the user's perspective by clearly stated basic document metadata with the name of the document. However, like all compromises, this affects the usability of the system.

This also raises an important question as to what extent the integrated access solution should refer to existing documents already published in the document sources, i.e. the so-called 'historical part of the timeline.'

Types of document

The most important indicator which tells the user about the content of the document is the metadata describing the document type. During the PublicAccess.eu showcase, it was necessary to choose the methodology for the assignment of document types to the individual documents which have been collected in the process of collecting data. Given the fact that the goal of the PublicAccess.eu showcase was to test the possible ways of presenting the documents of the EU, the final decision was to use a unified vocabulary describing the types of document which are available through MDR. MDR registers and maintains the definition data used by the different EU institutions gathered in the IMMC and by the Publications Office. Specifically, the vocabulary Resource Types NAL[227] was chosen.

The PublicAccess.eu showcase showed that this choice was correct since the vocabulary was able to cover the majority of the necessary document types. The PublicAccess.eu showcase also proved that the use of such a common code list, one to which several European institutions are accustomed, is definitely the right step. However, during the document processing, it was found that the entries used in this code list are not sufficient to describe all types of documents that have been collected for the purposes of the PublicAccess.eu showcase. If this vocabulary should be utilized for the description of all types of documents, to be made available through the future integrated access solution, it would be necessary to extend it. The use of Resource Types NAL was also extremely useful from the perspective of other language versions since the individual entries are available in all EU languages.

5.2.3.2. WHO (is responsible for the document)?

Another essential criterion for the categorization of documents and their classification is the question of who created the document, and who is the original author?

The same methodological principle for the document types was used during the processing of the documents collected for the purpose of the PublicAccess.eu showcase, i. e. the available vocabularies from MDR were used. In this case, it was the vocabulary Countries NAL[228], Corporate bodies NAL[229], Rapporteur EP[230], Rapporteur EESC/COR[231]. Again, the advantage is that they cover a substantial number of EU authors and the individual entries are available in various languages.

Again, the advantage of having a methodology for indexing the authors proved very helpful. For any integrated access solution, it is also necessary to decide on certain basic issues, for example, define

[227] Accessible at: http://publications.europa.eu/mdr/authority/resource-type/index.html.
[228] Accessible at: http://publications.europa.eu/mdr/authority/country/index.html.
[229] Accessible at: http://publications.europa.eu/mdr/authority/corporate-body/index.html.
[230] Labelled as FD_013.
[231] Labelled as FD_014.

whether the author of the document may be considered as the institution or further, as the organisational unit involved in the preparation of the document. Similarly, whether an institution or just one specific person who prepared the document should be labeled as the (co)author (for example the rapporteur in the case of the documents of the EP, COR or EESC). This problem can be solved in several ways, for example, include all available actors (institutions, organisational units, named persons) among authors or if necessary, to create several categories of authors in the form of a hierarchical list of authors. Then, the categories such as 'institution', 'organisational unit' or 'person' would constitute a separate metadata (a similar principle can be observed in EUR-Lex) and then they would subsequently be used in this way in results filtering, for example.

The PublicAccess.eu showcase demonstrated that any integrated access solution would need to consider the fact that existing vocabularies are not sufficient to fully cover all documents. This issue is particularly associated with the specific persons, who should be identified as the authors (for example, the name of a particular European Commissioner).

5.2.3.3. WHY (does the document exist)?

The PublicAccess.eu showcase dealt with the legislative documents specifically, and the question 'Why' was solved by the document being assigned to a particular stage of the legislative process. Then, the reason for which the document was created was clear (e.g. a proposal of the legislative act at the beginning of the legislative process, national parliaments' comments on the proposal, etc.). This topic shows the importance of the question 'Why'. If the future integrated access solution consisted not only of a simple search of documents from the various document sources in one place, but the search request options were also accompanied by explanatory notes and accompanying information (as in the case of various information on e-Justice portal), then question 'Why' would have a more accurate information value. Then, the user (EU citizen) would find documents with detailed accompanying explanatory information in one place.

5.2.3.4. WHEN (was the document published)?

This question seems clear at first glance, but the PublicAccess.eu showcase has shown that having a methodology would benefit this area as well. The methodology of the future integrated access solution needs to define which date (of those connected to the document) should be considered as the key date. The PublicAccess.eu showcase showed that it is necessary to deal with the question of what exactly the 'date of the document' is because many documents are accompanied with numerous dates. For example, in the case of 'Minutes of the meeting', this may be the actual meeting date, but also the date when the minutes are drawn up, which is generally a much later date than the date of the meeting. Also, the date of signing of the final text or the date of its publication may be considered. The same issue arises in judicial decisions (decision date vs. date of publication of the decision). Obviously, this issue can be resolved by introducing more categories of dates. However, the PublicAccess.eu showcase has shown that, for the purposes of clarity and understanding from the user's perspective, it is better to choose one of the available dates as the 'master date', and then use this in facet filters, etc.

6. Integrated access solution

The main goal of this chapter is to analyse and evaluate the options for establishing a new kind of service represented by the future integrated access solution and focused on the end-users. The integrated access solution would serve as a 'high-level entry point' to the other existing document sources. However, the integrated access solution would not be limited only to a simple retrieval of documents from another web site (like Google search). It would have to combine the advantages of a simple search of various document sources with coprehensive web portal functionality (i.e. speedy access to the document search, advanced user assistance, guides for effortless solving of the users' needs when seeking out the relevant documents and rich personalisation features).

6.1. PublicAccess.eu project context

Analysis of the document sources revealed that the native web applications are (in most cases) very well adjusted to the processes of the institutions or agencies that operate them. The idea of replacing them with a new universal integrated access solution may seem to be counterproductive in this regard. An integrated access solution would probably never go into as much detail as the current customized applications operated by their stakeholders (the institutions and agencies).

When designing the optimal integrated access solution one fundamental question needs to be addressed before all others: What is the main purpose of the integrated access solution? Seeking the best possible answer should be instantly focused on the users –to provide them with publicly available documents in an easily accessible and understandable manner via the one single point which does not have the ambition to replace the original document sources, but which guides the user to them if further or more detailed information is required.

There is an analogy between designing the overall concept of the integrated access solution and the perception of the legislative process documents within the gradual development of commercial legal information systems. In the past, the legal information systems included information about the new legislation. Later, it turned out that this was not enough. Thus, the legal information systems were enhanced with new functions showing the citizens the law as it was enforced and the consequences thereafter. Originally, the documents were presented in their original format. Later, the documents were re-processed to be presented in various ways per the end-user's needs. However, this is once more not sufficient anymore. Today's technical capabilities allow the user to better understand the law as it is enriched with accompanying information to create synergies for the users allowing them to work more effectively with the law and deal with specific life situations[232].

Similarly, the integrated access solution should not limit itself to merely presenting the documents. The final solution should go further to provide the users with a simple solution that meets their needs on how to obtain the documents and add the value with the provision of various assistance tools to understand them. In this view, the integrated access solution should serve essentially as a single contact point for the user, i.e. it might operate as: an 'Access EU documents' link placed for example in the europa.eu domain. It should be emphasized that the location for such a web application is an important issue. Therefore, the Documents and publications section on the main entry page of the europa.eu portal (i.e. http://europa.eu/publications/index_en.htm) could be a suitable location.

[232] A workshop organised within the meeting of the European Forum of Official Gazettes Legal (Fribourg, 2014) also came to similar conclusions about the development of a new role for information systems.

Alternatively, the EU Law and Publications portal managed by the Publications Office would also be a suitable candidate (access to EU law, publications and administrative documents).

The analysis proved that the specific web applications including the original document sources are built with a diverse level of detail and offer a different range of document content and context. Some applications are designed to provide only the mandatory information and they contain only the basic information about the given institution. Usually, they do not have an advanced search or the search options are very limited. Conversely, other institutions (or agencies) provide very extensive search capabilities with a variety of different document types, differentiated co-authors, a variety of descriptive information (EuroVoc, various kinds of thematic vocabularies) etc.

The analysis points to the fact that some institutions (agencies) would benefit from the creation of an entirely new and comprehensive integrated access solution in the form of a standard complex information system. At the same time, it should be noted that for others, this would be an unnecessary duplication of their existing systems, which their regular users are accustomed to, and towards which significant funds have been put. These systems serve their purpose particularly well. It can also be expected that these institutions (agencies) would openly (and probably rightfully) refuse this new initiative, because from their standpoint it would not deliver any positives to their users. Therefore, the planned integrated access solution should not aspire to be a substitute for these existing applications, which work well.

The new integrated access solution must be focused on a new form of added value that would justify its creation whilst respecting the status of the existing document source applications. The main role of the new solution from the 'business point of view' would be to bring 'new customers' to the existing EU institutions' document source applications. This does not mean that some of the basic search options and selected functionality should not be provided directly by the integrated access solution.

Support from the institutions

If the final integrated access solution should truly become a gateway to all EU documents available to the public, it needs to get strong political support from the EU institutions. Without this support, there is a risk that the expected benefits would not occur.

A parallel can be seen with the e-Justice portal, which received strong support and finally became an excellent example of how to provide a wealth of information about the EU, the individual Member States and their national legal systems. Additionally, it provides various added value features (forms, etc.) for the users so that it is not only a simple repository of information.

6.2. Integrated access solution design methods

As mentioned in the last chapter of the analytical part of study (see 4.4.7) the foundations of the integrated access solution exist in the form of an information ecosystem in europa.eu, in a very granular, diversified and decentralized form.

Users have access to all available public documents. The problem is that the user must go through a number of different systems and the particular document retrieval process must meet the following conditions:

1. The user must know WHAT to look for.
2. The user must know WHERE to look.
3. The user must know HOW to look.
4. The user must be sure that he/she has found everything and has not missed something.

In summary, the user must have considerable expertise and experience to handle the document retrieval process.

On the contrary to this approach, the integrated access solution should represent a new level. It should provide all the documents in one place (according to the principle of a one-stop-shop), and it should not require much expertise from the user.

The creation of an advanced search form covering all the metadata of all the institutions would not be the goal of the solution. Such a 'mega search' would be confusing and complicated. By combining the characteristics of the author and the topics the users will be navigated further to address their needs.

Therefore, a simplification is needed. The integrated access solution should provide the user with retrieved documents that answer the four basic questions:

1. WHAT (is the document about)?
2. WHO (is responsible for the document)?
3. WHY (does the document exist)?
4. WHEN (was the document published)?

All other options would essentially be an extension of these main criteria.

This principle of Document retrieval can also be depicted schematically using a Cause & Effect diagram, see Figure 47. A *Cause* is represented by the user need to obtain a document, an *Effect* by the retrieved document.

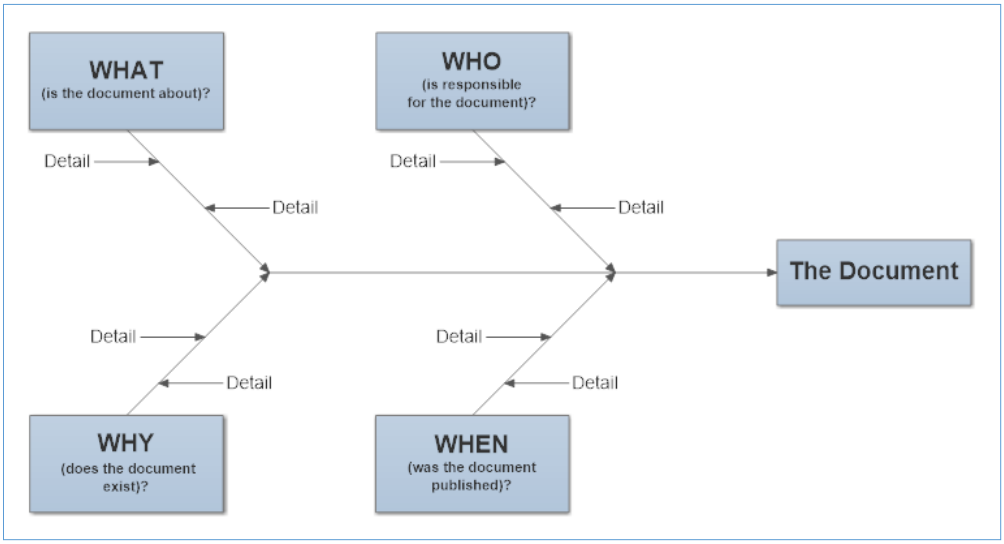


Figure 47: WHAT/WHO/WHY/WHEN document retrieval principle

For this principle to succeed it is necessary to prepare and re-process the context and, arguably, the content of the documents obtained from their original document sources. The full sets of metadata from the original document sources (provided by the original applications) would be used to cover the additional users' needs.

In this respect, it is particularly important to retreat from the established and generally applied paradigm that 'document = file attachment'.

Due to the differing quantity and quality (i.e. the context) in the detail of the documents about the legislative processes and other documents, it seems appropriate to treat these two groups separately, with respect to the available metadata.

Nowadays, the focus of all web applications is their user interface as this is the main gateway which identifies the needs of the users and fosters a positive experience.

One can guess that the deployment of any integrated access solution would be gradual and dependent on the cooperation of many stakeholders. Therefore, it must be designed with this in mind and must be prepared for a step-by-step integration.

6.3. Integrated access solution architecture

This chapter presents a high-level architecture of the integrated access solution together with an outline of its main components. This chapter also presents a template for the description of various solution alternatives. The following chapter 6.4 then explores the specific solution alternatives each with selected components specified in this chapter.

The solution architecture outlines the following components:

1. Automated content & context retrieval from the stakeholders' document sources;
2. Automated context reprocessing into the single ontology focused to users' needs;
3. Automated content parsing (file attachments → HTML5);
4. Intuitive management of both the content and context;
5. Front-end web and mobile application.

These components are highlighted in green in Figure 48 and described in the following chapters.

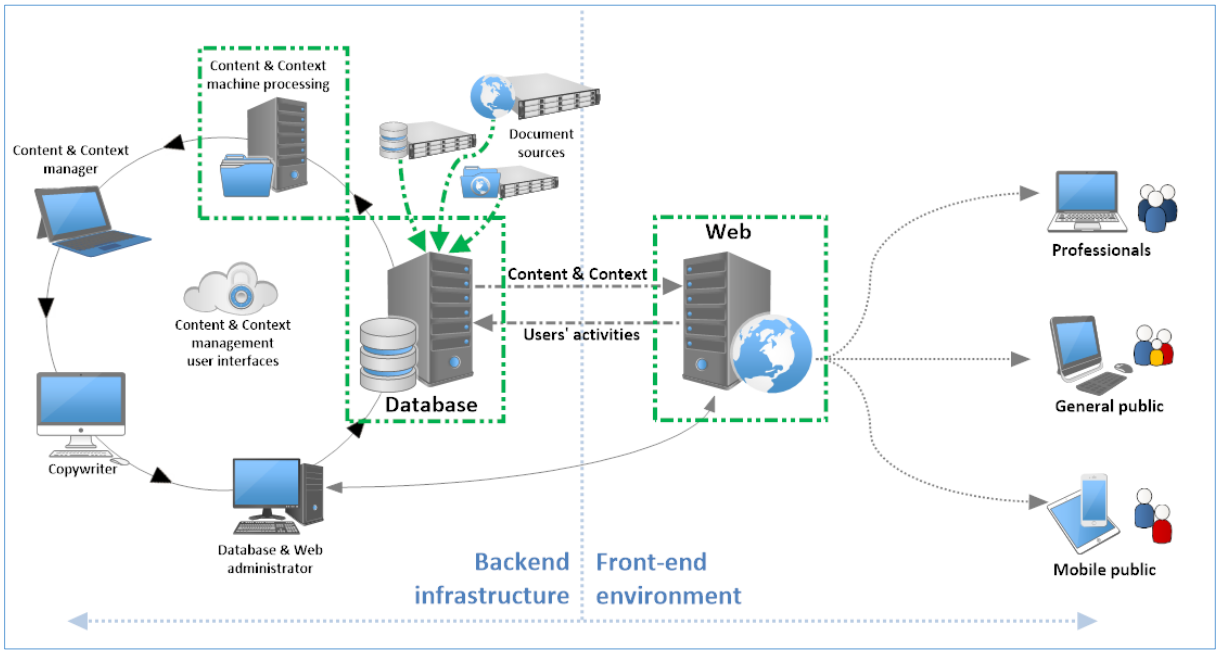


Figure 48: Five basic components of the integrated access solution architecture

It should be emphasized that integrated access solution would be a new project which represents an extension to the document content & context stored in the existing document sources. As for each new project, it would be appropriate to consider its sustainability and staffing: this includes the solution management staff as depicted on the left part of Figure 48 above.

6.3.1. Automated content & context retrieval

The analysis of the document sources (see Table 28) showed that the possibilities for the automated content & context retrieval from the stakeholders' document sources are currently limited. In line with the conclusion of chapter 4.4.5, it is necessary to note that it would be necessary to cooperate with the stakeholder of each document source on the following topics:

- To determine which documents from the particular document source would be displayed within the integrated access solution.
    - It is also necessary to decide whether the documents not available from CELLAR
        - should be either stored to CELLAR at first and then used,

    - or such documents would be stored directly in the database of the integrated access solution without previous storing to CELLAR,
    - or such documents would be presented by the integrated access solution directly from the storage of their original document sources;
- To define the necessary context of documents in scope and thus, to determine the set of metadata for the specific document sources and their subsequent use in the integrated access solution;
- To identify the technical method by which the content and context of the document could be transmitted to the integrated access solution for their further presentation;
- To carry out the necessary development both on the source document side and on the integrated access solution side;
- To ensure maximum automation;
- To define a way of updating content and context.

Only the document sources administrated by the Publications Office (CELLAR/EUR-Lex and TED) provides an API and offers the full capabilities of machine readability.

In other cases, the transfer method between the document source and the integrated access solution would have to be developed from scratch.

There are two basic automated transmission methods:

1. PUSH transfer methods, where the document source is in an active role. It actively provides the data (both content and context) and directly sends them, with the possibility of bidirectional communication (sending acknowledgment of receipt).
2. PULL transfer methods, where the document source is in the passive role and the active role is carried out by the integrated access solution, which obtains the data needed directly from the document source.

Both ways are further specified in the following chapters including both their advantages/disadvantages and implementation issues.

6.3.1.1. PUSH transfer methods

In the case of PUSH transfer methods, the document sources are in an active role and initiate the data transfers.

PUSH transfer methods are already in place and widely used for inter-institutional data transmissions. They were created and are maintained by IMMC.

PUSH transfer methods are represented mainly by the IMMC Core Metadata Exchange Protocol. The protocol can be described with a large degree of simplification as follows:

- In the document source, the document file and the package containing the XML file in FRBR notation (see 4.2.10.1.1 and 4.2.10.1) are prepared:
    - each document source creates an XML file per a predetermined, defined and agreed XSD schema;
    - the Work section contains metadata compatible with CDM ontology;
    - the Item section contains references to files (streams) either embedded in the package or referenced to their original locations.
- The document source initiates data transmission to the destination, which is typically a data store in CELLAR (for example E-TrustEx or FTP).

- In CELLAR, the package is automatically processed and placed into the storage.
- Document source is informed about the success/failure of the transaction processing with other detailed information.
- IMMC Core Metadata Exchange Protocol also defines the procedure for updating the metadata structures that arise on the side of the document source (i.e. the cases where the source document will, for example, publish new document types or new item in metadata structures through provisional values called DATPRO[233]).

The IMMC Core Metadata Exchange Protocol has a distinct advantage - it is an established way of data communication between EU institutions. Some institutions already have the functional implementation in some of the document sources and are used to working with the IMMC Core Metadata Exchange Protocol in daily practice and they are further developing it's implementation at both the technical and the process level.

For the purposes of the integrated access solution, the data acquisition (i.e. the content and context of the documents) from the document sources based on IMMC Core Metadata Exchange Protocol would be applicable in two ways:

1. The integrated access solution would retrieve the data from CELLAR. Thus, it would connect with the existing method of communication between the institutions:
    - Advantage: ease of implementation;
    - Disadvantage: slim chance of acquiring data other than those which are already the subject of data communication between the institutions. Its implementation would be necessary to expand the CDM and CELLAR in collaboration with the document sources' stakeholders.
2. The integrated access solution would implement a new IMMC gateway on its side:
    - Advantage: integrated access solution could be fully customized;
    - Disadvantage: a parallel IMMC sending packages from the institutions would be needed, hence the need to extend their implementation.

The advantage of the latter (customized integrated access solution) seems to be obvious – it could address exactly the context and content which would be presented by the integrated access solution to the end-users with no redundancies above this range. Nevertheless, it would be necessary to have a managerial decision on which of the two options described above will be chosen.

To repeat, it is necessary to point out the fact that only some institutions and only certain document sources have a functional implementation of the IMMC Core Metadata Exchange Protocol. Other institutions do not have such implementations. Besides that, there are 46 agencies where a functional implementation barely occurs. Full implementation of the IMMC Core Metadata Exchange Protocol by all institutions would be very expensive and time-consuming. Because of the diversity in their IT environment, it could be assumed that a significant multiplication of expenses would be needed (since analogous and even the same tasks would have to be realized more often.)

It is also worth mentioning that the continual development of CELLAR is on-going. This should bring new possibilities, tools and scenarios for the smooth integration of the data from document sources with CELLAR. In the end, the document sources' stakeholders should be able to use new methods of transferring their data to CELLAR, which could suit them more than IMMC PUSH transfer methods.

[233] DATPRO: Surviving without a concept (http://publications.europa.eu/mdr/core-metadata-schema/official/documentation/DATPRO-solution_20150521.pdf).

6.3.1.2. PULL transfer methods

The limitations outlined in the previous paragraph could be resolved by the integrated access solution taking over an active role and acquiring the data (both document content and context) from the document sources. The integrated access solution would initiate and perform the data transfers by pulling them (hence 'the PULL method') from various document sources. For this task, the solution would include tools customized for the given document source ('data pumps'.)

There are basically three options to use the PULL transfer methods:

1. New API on the side of each document source, unified data pump on the side of the integrated access solution;
2. Adaptation of the integrated access solution to the current situation so that it can initiate the specific data pump for each document source;
3. Document sources would publish the datasets with metadata and references to the content of their documents on the Open Data Portal and the integrated access solution would be adjusted accordingly.

These three options are described in the following subchapters.

6.3.1.2.1. New APIs at document sources side

The ideal solution on the side of the document sources would be the deployment of a new unified API interface, through which the integrated access solution would obtain the data (both content and context).

However, the complexity of deploying such solution is comparable to full implementation of the IMMC Core Metadata Exchange Protocol (see 6.3.1.1) on the side of institutions and agencies.

6.3.1.2.2. Integrated access solution adjustment to current possibilities of document sources

The results of the analysis of all document sources have indicated the possibility to develop such data pumps for data transfers of the data from the specific document source to the integrated access solution.

The data pump types can be generally summarized as follows:

1. The automated and periodically generated queries to search engines of various document sources (GET Method e.g. sequence of queries with incremental dates of the documents day-by-day);
2. Receiving the information from RSS channels provided by some document sources;
3. The use of specific methods offered by some document sources (e.g. periodically generated XML archives).

Of course, the individual types of data pumps could be mutually combined when retrieving the data from a single document source. The implementation of the data pump is relatively easy and it is easily adaptable to changes or improvements over time.

The data pumps work with the document source interfaces and they assume that the document source stakeholder (EU institution or agency) has already published all the necessary content and context through the dedicated web application.

The number of data pumps would be relatively large and their functionality would be diverse. Therefore, it would be necessary to develop a common collaborative environment for the administrators of the document sources and the administrator of the integrated access solution where

the content and context of each document source will be registered and maintained in relation to the implemented data pump. Such an environment could also serve to monitor the data acquisition (i.e. which data was obtained through the data pump and when, monitoring for and solution of problems, etc.).

6.3.1.2.3. Usage of the Open Data Portal

Another possibility in PULL data transfer is the widespread use of Open Data Portal[234] (ODP). For example, the EC publishes and periodically updates a list of documents available through the RD-EC (see 4.2.3.1), including links to these documents on the ODP. If all institutions and agencies would do the same, it would bring the following advantages:

- Simplify the development and implementation of the data pumps by the integrated access solution supplier;
- Provide transparent information to the public about the content and context, which would be available in the integrated access solution whereby this information may also be used independently by the public.

This solution of acquisition of the context from the ODP and the content from the original document sources would be a typical application of the way in which Open Data should work and what they are intended for. However, this way of obtaining data duplicates new scenarios and tools currently implemented in CELLAR, so it is not considered as an appropriate method in the remaining parts of the study.

6.3.1.3. Conclusion on automated obtaining the content and context

It seems obvious that the decision of which method to choose is a matter for managerial decisions.

The main question is whether to:

- Continue with the current approach of in strengthening the position of CELLAR as a central storage for documents' content and metadata by developing more feasible and user-friendly interfaces for document sources stakeholders on how to transmit their data for presentation in the future integrated access solution;
- Build a brand new process and infrastructure level for gathering the data for the chosen integrated access solution.

Possible specific scenarios of the integrated access solution are outlined in the chapter integrated access solution alternatives (see chapter 6.4).

6.3.2. Automated context reprocessing into the single ontology

6.3.2.1. Integrated access Ontology

In the opening chapters of this section the vision and the expected benefits of the integrated access solution are outlined. It is obvious that a new way of understanding the context of the documents (i.e. adding the new level of meaning and usability to the current document metadata layer) is a basic prerequisite.

[234] ODP website homepage: https://open-data.europa.eu.

At several places in the study there is a paradigm postulated that the end-user searches for answers (in the form of documents) by the gradual refinement of the search criteria represented by the questions:

1. WHAT (is the document about)?
2. WHO (is responsible for the document)?
3. WHY (does the document exist)?
4. WHEN (was the document published)?

Therefore, there would be a need for one ontology to have the same understanding of the selected metadata provided by specific document sources. A high-level vision of such ontology is shown in Figure 49. It is evident that this ontology is compatible with the ontology of study database used for the analysis of document sources (see Annex 1: Description of the study database ontology).

A basic class in the ontology is the document. It is linked to the final vocabularies by associations.

These vocabularies represent the search criteria for the end-user:

- WHAT – Document Type and Document Theme
- WHO – Responsible body
- WHEN – Timeline
- WHY – Procedures, Dossiers etc.

This is indicated in Figure 49 by the group *Retrieval* in the left part of the figure.

Therefore, the association represents the relationship between a specific document and the specific vocabulary entries. There may be more kinds of associations or properties of associations between a document and a specific vocabulary, e.g. between Document and Timeline there can be Date of creation, Date of signature, Date of publishing, etc.

The vocabularies should be created by the supplier of the integrated access solution based on a thorough analysis of the documents and document sources in collaboration with the stakeholders of the document sources. It should be noted that the main purpose of these vocabularies is to adjust the large number of current metadata to the needs and expectations of the integrated access solution users. Again, a possible method of this aggregation is indicated in the study databases developed during the analysis of specific document sources.

The organisation of ontology as a set of taxonomies and/or thesauruses assumes interconnections by various pre-defined associations. The purpose of such associations is to enable the creation of the recommendation tool. For example, the responsible body may produce only certain types of documents or certain document types are only used in some time periods, etc. This recommendation tool is indicated in Figure 49 by the group *Recommendation* in the right-hand side of the figure.

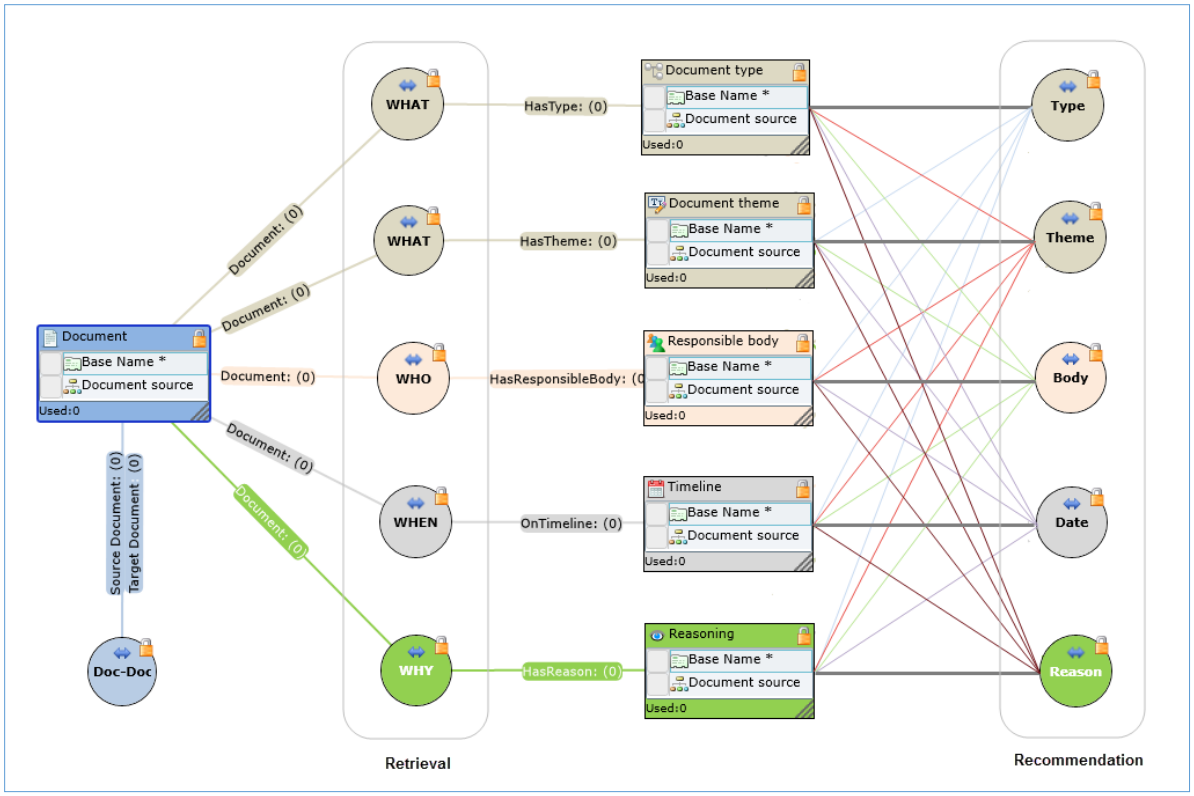


Figure 49: Vision of the integrated access solution ontology

An important requirement is an effective data model for meeting the rules outlined in the ontology above accompanied by the necessary infrastructure.

6.3.2.2. Selected aspects of context reprocessing

Preparation of the new vocabularies for Document type, Responsible body and Timeline does not raise any major questions nor does it indicate problems during their creation. Conversely, the vocabularies Document theme and Reasoning deserve special attention.

6.3.2.2.1. Document Theme vocabulary

The analysis of document sources revealed that EuroVoc thesaurus seems to be the optimal solution for the description of Document theme. It is used in the RD-EP (see 4.2.1.1.3.2), IPEX (see 4.2.1.3.2.7) and EUR-Lex (see 4.2.10.3.2.1).

Other document sources use proprietary methods for the description of their document themes or do not use any. There are two potential reasons for this:

1. Either EuroVoc is too large or complex for them;
2. Or vice versa, EuroVoc does not cover in detail all of their documents.

Given the importance of the structured notation for the Document theme, the use of EuroVoc is highly recommended even within the document sources that do not use it yet. The afore-mentioned reasons for not using EuroVoc could be overcome for example by:

1. Encapsulation of EuroVoc into a looser structure of the integrated access solution ontology, which would include also other methods of document classification (e.g. InfoCuria classifications) as well.
2. Better specification of some of its components with respect to the specific needs of the document sources and presentation of their data (context and content of their documents) in the integrated access solution architecture.

However, it must remain clear which parts are generic EuroVoc components and which extend it.

The main task in this field would be finding a sustainable way of document indexing according to EuroVoc. Due to the large number of documents anticipated, it would be necessary to deploy the automatic indexation according to EuroVoc in the document sources that still do not use EuroVoc.

The creation of the Document theme vocabulary and the method of EuroVoc usage would be a major task for the supplier of the integrated access solution architecture. This task must be solved in cooperation with the document source stakeholders.

EuroVoc automatic indexation

The performance of human indexing compared to machine indexing is always superior. But with regards to overall costs, the method of a machine-made indexing could be also considered. Especially for documents not currently indexed.

There have been numerous publications about automatic EuroVoc classification in the last decade[235]. One of the most interesting publications is 'JRC EuroVoc Indexer (JEX)' developed by European Commission - JOINT RESEARCH CENTRE in 2012[236].

These techniques have one serious downside - their precision and recall are significantly lower compared to human indexation. For instance, JEX precision and recall are around 0.5 (therefore 50% of identified nodes are false positive and 50% has not been found). Nevertheless, it is safe to say that machine learning is a dynamically progressing field and the performance achieved in 2012 could be bettered nowadays. Moreover, the machine learning techniques could benefit strongly:

- From better understanding of the way of human indexation (i.e. which parts of documents are considered by determining the suitable EuroVoc concepts);
- From better understanding the so-called training input set (the human indexed data used to 'train' the algorithm).

It is also worth noting that even human indexing does not have 100% accyracy since the attribution of EuroVoc descriptors can be subject to the opinion of the indexer.

6.3.2.2.2. Reasoning vocabulary

Reasoning is a specific vocabulary, which should specify the reasons for the existence of a document. These reasons can include for example:

- Documents of the legislative process which are part of the procedure.
- Other documents which are part of dossiers that have a common theme.
- Documents for meetings are connected to the event of the specific meeting.

Therefore, the vocabulary Reasoning will be very complex and it will contain a large number specific items (e.g. Procedure files) and this number will certainly rise over time.

---

[235] Luis M. de Campos, Alfonso E. Romero, Bayesian network models for hierarchical text classification from a thesaurus, International Journal of Approximate Reasoning, v.50 n.7, p.932-944, July, 2009;
Boella, G., Caro, L.D., Lesmo, L., Rispoli, D., & Robaldo, L.. (2012). Multi-label Classification of Legislative Text into EuroVoc. JURIX;
Villena Román, Julio and Collada Pérez, Sonia and Lana Serrano, Sara and González Cristóbal, José Carlos (2011). Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization. In: 'Twenty-Fourth International Florida Artificial Intelligence Research Society Conference', 18/05/2011 - 20/05/2011, Palm Beach, Florida, EEUU. pp. 323-328.

[236] https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer

6.3.2.2.3. RDFa and SPARQL

The integrated access solution ontology could also be published in the form of RDFa snippets to open the integrated access solution architecture for SPARQL querying. In the case of content parsing from the file attachment to HTML5 (see 6.3.3), it is also worth considering, in the future, embedding RDFa triplets directly into the HTML5 content for further description of its meaning.

6.3.3. Automated content parsing (file attachments to HTML5)

The study repeatedly states that the documents are typically published as attachment files mainly in PDF format and sometimes also in the other formats such as DOC, XLS etc. Chapter 4.4.7 of the study details the advantages and disadvantages of publishing these documents in this manner.

If the integrated access solution architecture should improve the current situation, it is necessary to pursue the unification of the presentation of documents. This means their conversion from the binary and machine-unreadable format to a structured form described by a predetermined markup language.

Formex markup language has been used within EUR-Lex for decades. Considerable funds were invested into the conversions of the documents from their source formats to Formex. The use of Formex for marking the documents within the integrated access solution would be the best solution from a technical point of view. From an effectiveness and practicability point of view it would mean a significant multiplication of those costs and therefore, the use of Formex for the integrated access solution should not be considered.

The issue of publication of the textual content (structured versus binary formats) has been discussed for some time in the professional community[237]. In accordance with the results of these discussions as well as with the development of web presentation techniques, it can be postulated that the most suitable format is HTML5, which is also regarded as the standard[238].

Therefore, it would be necessary to develop tools for the conversion of the textual content published in binary and machine un-readable formats (PDF or DOC) to HTML5 for their usage within the future integrated access solution. Because tens of millions of the documents in all their language versions are estimated (see 4.4.2), those instruments must be fully automatic and 100% reliable. The complexity of the development of such conversion tools is obvious.

The integrated access solution must not be considered a temporary solution to a current problem, but a long-term project whose outcome should be:

- A next-generation solution meeting both the current and future needs of the end-users.
- Usability on the mobile devices, whose use is increasing by tens of per cent each year.
- Possibilities to link or cite the specific parts of a document.
- Opening new possibilities for working with the text content (semantic web, embedding RDFa snippets etc.).

The development of new IT technologies is accelerating from year to year, new products and services appear every day and therefore it could be stated that the textual content in binary and non-structured format will be a substantial obstacle to their implementation.

---

[237] Serge Huber: 'After Flash, why PDF must die!' http://community.aiim.org/blogs/serge-huber/2012/04/25/after-flash-why-pdf-must-die-.

[238] W3C Recommendation: HTML5: https://www.w3.org/TR/html5/.

To convert all file attachments into HTML5 might be impossible. There are a number of documents where formatting and typesetting is the ultimate bearer of their meaning (such as annual reports with almost "artistic" way of typesetting). A conversion to HTML5 could decrease their understandability. Such documents should be clearly marked out and left out of the automated conversion described above.

It is also worth considering applying HTML5 conversions for all future documents while the documents from the past will be processed only selectively.

Based on aforementioned statements it may be concluded that in this area there are a few key managerial decisions needed:

- Whether to apply the conversion or not.
- If so, to what extent and how far back into history.
- If so, whether and how to involve the other document sources' stakeholders.
- If so, how to specify the objectives to be achieved in detail and how to realize these processes (a document sources' side versus an integrated access solution level).

Several other possibilities and benefits from the conversion of file attachments into HTML5 are later described further.

6.3.3.1. Use of Natural Language Processing (NLP) of texts of documents

The progress in the field of NLP has been noted by some of the EU institutions. For instance, a tender for automatic annotation of patent texts has been published in 2015 by the European Patent Office: http://ted.europa.eu/udl?uri=TED:NOTICE:183511-2015:TEXT:EN:HTML.

The reason for including this information is to point out that several specific projects related to this matter and the information systems within the europa.eu ecosystems has already taken place or are currently running. It might be useful to use the results of these projects if cooperation with the sponsors of these projects would be possible.

6.3.3.2. Named Entity Recognition (NER)

Named entity recognition ('NER') is a process by which a text is automatically analysed to locate and classify elements in the text into pre-defined categories such as the names of persons, organisations, locations, expressions of times, etc. One of the interesting uses of NER is the Europe Media Monitor ('EMM')[239], which provides a means to analyse various news articles with multilingual aspects. As a by-product, a very useful resource of multilingual named entities has been published on ODP.[240]

Extraction of entities could provide additional features for searches within the web application as well as for external search engine optimization ('SEO').

NER can be used in order to contribute to a 'find similar document' feature as well.

6.3.3.3. Citation annotation

The texts of documents can contain references to other documents which are currently not annotated in any way. One example is the following sentence taken from 'Commission Implementing Regulation (EU) 2016/569' published on EUR-Lex[241]:

[239] Europe Media Monitor: http://emm.newsbrief.eu/overview.html.
[240] ODP – Datasets by JRC: https://open-data.europa.eu/en/data/dataset/jrc-names.
[241] Commission Regulation (EC) 329/2007 on EUR-Lex: http://eur-lex.europa.eu/eli/reg_impl/2016/569/oj.

'*Having regard to Council Regulation (EC) No 329/2007 of 27 March 2007 concerning restrictive measures against the Democratic People's Republic of Korea*'

The sentence contains a reference to another document and the formulation of the reference has a predictable machine readable format '*(EC) No 329/2007'.* Such text can be annotated with a reference to the target document concerned. Moreover, the annotation can be brought to the user interface so that the relevant part of the sentence includes a hyperlink.

For the extraction of such citations, a classical annotator is usually better suited. An annotator based on regular expressions[242] can be considered representative of such annotators. These annotators usually yield very good precision rates compared to various NLP solutions.

6.3.4. Intuitive management of both the content and the context

The institutions/agencies in the role of the document source owner create the content and context of the documents published by them to the public which would be in turn provided to the integrated access solution. Presentation of this content and context by the integrated access solution and their management will be ensured by the newly established agile integrated access solution team. Such a team is considered in all integrated access solution alternatives (see 6.4).

The following list represents the basic requirements for the management of both the content and the context of documents:

1. Content processing
    - The original documents in PDF/DOC/DOCX formats must be fully automatically re-processed into the HTML5 for their further presentation in the integrated access solution. There should exist fast, transparent and convenient tools for such re-processing operated by the integrated access solution team.
2. Context processing
    - The original, highly diversified metadata must be converted to a new common ontology (i.e. the new vocabularies, thesauruses, ontologies, etc.) by the automated tools operated by the integrated access solution team in a convenient and sustainable way to meet the described 'WHAT/WHO/WHY/WHEN' principles.
3. Users' activity evaluation
    - The integrated access solution must be equipped with tools to evaluate the activity of users with the goal of permanently improving the services for the users.

    Vision
    - The tools for the users' activity evaluation should include elements of artificial intelligence, for example to be able to analyse the queries put forward by the users to obtain the results, the assistance provided by the solution and to permanently improve the future user experience. Of course, the best practices that ensure personal data protection must be followed.
4. Readiness to extend the content and context
    - The integrated access solution would undoubtedly be deployed gradually with a permanent evaluation of the success of deployed elements. In this regard, the gradual extension of completeness and quality of the content and context presentation must be considered.

[242] Definition of 'Regular expression' on Wikipedia: https://en.wikipedia.org/wiki/Regular_expression.

6.3.5. Front-end web and mobile application

The key success driver is primarily the optimal user interface. The integrated access solution must provide the users with the common functionality frequently used in today's web applications and it must also apply the latest trends enabled by current technologies and web design techniques. Support for seamless usage on various devices (namely smartphones and tablets) is also essential.

The following section summarizes the fundamental requirements for any of the alternatives of the integrated access solution. It relies on the fulfilment of the previously specified requirements of the integrated access solution high-level architecture. It primarily focuses on the results provided by the integrated access solution to the end-user. The front-end environment is logically separated from the back-end infrastructure, which may be technically implemented in various ways.

6.3.5.1. Usability and user experience

The target audience is the user that does not have any special experience in IT. At the level of ergonomics, usability and the user experience, the integrated access solution must meet the following requirements:

1. Full multilingualism
    - The user interface must be available in all EU languages.
    - The documents should be available in as many language versions as possible. If there is no document present in the selected language, the same document in another language should be offered to the user.
2. It will not be a necessity to study a user manual or documentation beforehand
    - The basic requirement in today's web and mobile applications is intuitiveness so that the user can assimilate everything from the very beginning.
    - The user must always be aware of the composition of the user interface and what the application could offer as well as the next step to reach the desired result.
3. Emphasis on the performance
    - The final solution must handle the user requests in the fastest way possible.
    - This requirement may seem to be obvious, but when one considers the expected large number of documents, it is necessary to explicitly emphasize this requirement.
4. Stability without outages
    - It is necessary that the final solution gains the trust of the users and they will not experience failures or unanticipated responses.
5. Device independence
    - Diversification of the users' devices, for which the resulting solution should be built, is significant. It can be noted, that at least 10% of users currently browse the web from their mobile devices. According to the principles described by Sir Geoffrey Moore in his book Crossing the Chasm[243], the time is approaching when the share of mobile devices will grow exponentially. Therefore, the planned solution must anticipate this fact and be prepared for it from the beginning. Technically speaking, the website design must be responsive.

[243] https://www.amazon.com/Crossing-Chasm-3rd-Disruptive-Mainstream-ebook/dp/B00DB3D81G

6.3.5.2. Intuitive search process

1. Support of document retrieval by the WHAT/WHO/WHY/WHEN principle
    - The user must be able to start searching from either side.
    - The analysis showed the great complexity of the document contexts, which is also dependent and dedicated to the individual document sources. The integrated access solution must lead the user in an intuitive way, and it should not assume any elementary knowledge or skill from the user.
2. Gradual guiding of the user to the result
    - The user must continuously receive the actual results based on the specified requirements.
    - The integrated access solution must continually and meaningfully advise users on how they should proceed further. The emphasis here is mainly on the word 'meaningful.'

    Vision
    - One possibility here is a real-time response for the specified requirements in the form of a continual generation of results. The number of results is not the key information for the user (e.g. it is irrelevant whether the user found 100 or 5 000 documents: in both cases, it is too much). Practice shows that working with high numbers of results has a dramatic impact on the performance of the retrieval systems. Moreover, it could be applied to the principle of 'get next set of results'.
    - The user assistance option may be represented for example by an intelligent suggestion, i.e. user assistance through real-time suggestions appearing below the search form during user guidance. This may seem common in all searches nowadays. However, it is an advanced implementation for all possible requirements of searches (WHAT/WHO/WHY/WHEN) here.
    - An alternative option could be a wizard that guides the user gradually, step-by-step through the document retrieval process.
    - These options are not mutually exclusive. Conversely, they can work in conjunction very well together.

6.3.5.3. Intuitive list of results

1. Refining of the results found
    - The user must have the permanent possibility to improve the search results by query modification or by application of facet filters working with the final vocabularies.

    Vision
    - The facet filters should have a tree organisation, i.e. they should be composed as taxonomies.
    - The user can then select the facet filter, which is available in a convenient way above the current results, and the results can be narrowed (improved) in a relevant way.
2. Contextual suggestions of the relevant results would be based on the already displayed set of results
    - The displayed results should have unimaginable contexts in the form of additional metadata which the user will not know about or will not work with.
    - The integrated access solution should offer such contexts by expanding the list of results, and adding a new list of results based on the common metadata groups. This would probably be substantiated particularly in the case of vocabularies topics, i.e. the thematic classification of the documents.

    Vision

- There is one highly used technique in the field of machine learning now and this is a 'recommender system'. Recommender systems seek to predict a list of items that would be preferable for a specific user. Such systems can recommend movies, songs, books, documents or any other content in general.
- The recommendations are usually based on collecting user specific data (search queries, viewed documents, time spent on specific pages/documents) and comparing such information with results of other similar users and/or by finding similarities by observing various metadata of examined items.
- Such techniques could be used, for example, to further increase the quality of the search results.
- There is an elementary requirement for the protection personal data, i.e. end-users should have a transparent and easy to use tools to manage what the integrated access solution records about their activities.

6.3.5.4. Intuitive detail of the document

1. Unified presentation
    - The user must be able to view the document in a uniform manner.
    - The documents must be adapted to a unified view and consequently, pre-prepared for this.
2. Sharing of the whole document or parts of the document
    - The user must be able to share the document not only as a whole, but also its parts, e.g. individual paragraphs, titles of chapters, etc.
    - The documents must be appropriately pre-processed for the sharing of its individual parts.
3. Relationships to similar documents
    - The document view must include an active context where the user has the documents directly related to the displayed document and with, respectively, similar documents based on their identical or similar metadata.

Vision

- Natural language processing also offers a feature which can find similar documents based on their content.
- This could potentially be used in a feature such as 'find similar documents'. Similar documents could be found solely by their content or, where possible, by their attributed metadata as well (metadata such as: document type, topics of document, document relations, etc.)

6.3.5.5. Personalised usage

1. 'Industry standards' of user authentication (i.e. Google, Facebook, Twitter)
2. Personal storage
    - The user must be able to save links of the list of results and to their own documents in folders, including advanced features. The advanced features include, for example:
        - Annotation of documents by their own metadata or text description
        - Notification of new or relevant documents in a common way (e.g. RSS channel, e-mail, etc.)
        - Sharing of saved folders with links to documents, respective to the list of results.

3. Personal notepad
    - The user must be able to comment on the document or its parts and to share their comments with other users. Additionally, the user must be able to use the comments in the documents shared with them by the other users.
4. History of user activities
    - The user must be able to view the history of his/her work within the integrated access solution based on the time of the activity or by the metadata of documents sought or otherwise worked with – e.g. via the search.

6.3.6. Mind map of the integrated access solution vision

Figure 50 shows a high-level mind map of the integrated access solution vision as described earlier in this chapter. From this point of view it represents 'the best-case solution' providing maximum benefit to the end-users. Each object of the mind map represents a standalone building block to be developed and deployed.

The complex implementation of all building blocks appears very labour intensive at first glance. Of course, it is not necessary to carry out the full set of these building blocks. Alternatively, they can be carried out gradually, in stages.

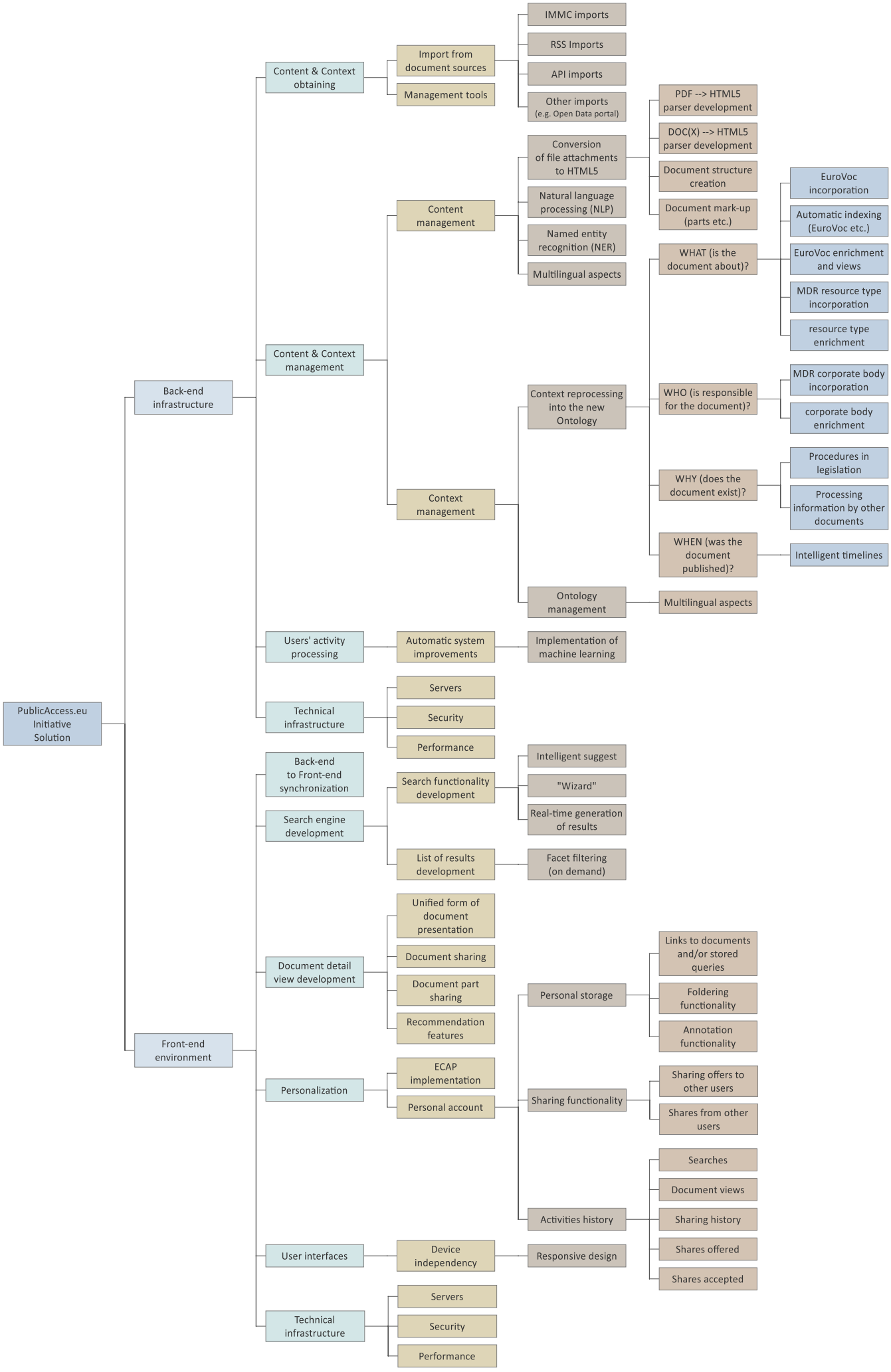


Figure 50: High-level mind map of the integrated access solution

6.3.7. Notes on associated risks

The mind map implies that each building block of the integrated access solution is important and represents an essential element in it's success. It cannot be said in any case that, if n-number of building blocks are omitted, the integrated access solution would achieve its goals adequately compared to the number of dropped blocks, i.e. from (100-100/n) percent (where n is the number of omitted building blocks).

The integrated access solution would constitute an independent and comprehensive information system. As already stated at the beginning of chapter 6.3, the integrated access solution would be operated in parallel with the existing systems of the EU institutions and agencies, who would also be contributors to the integrated access solution's content and context.

However, one of the obvious goals of the integrated access solution is to make the public documents more accessible at least for some groups of end-users (i.e. citizens) for which the current processes of document retrieval from the document sources are complicated. For those users, the integrated access solution would serve as a replacement to the current generic document sources. This is a very broad ambition generating a lot of risks, for example:

- Lack of support from the document sources' stakeholders, or a lack of high-level political support;
- Problematic transfer of content from file attachments to HTML;
- Poor quality of the generic context without the metadata of the original sources;
- Difficulties in maintaining the whole system;
- Problematic implementation of end-user features due to the wide variety of data sources.

The escalation of each of these risks can lead to the failure of the entire integrated access solution deployment project. In literature and in the web sources it is possible to find many case studies on project failures, which can be regarded as analogous to the integrated access solution in terms of scale and complexity. Refer to, for example, the series of articles; 'Lessons From a Decade of IT Failures - The takeaways from tracking the big IT debacles of the last 10 years.'[244]

Therefore, it is necessary to pay attention to all aspects of the integrated access solution feasibility.

[244] IEEE Spectrum – 'Lessons From a Decade of IT Failures - The takeaways from tracking the big IT debacles of the last 10 years' (http://spectrum.ieee.org/static/lessons-from-a-decade-of-it-failures).

6.4. Integrated access solution alternatives

The components of the integrated access solution were described in the previous chapter. Basically, it is the maximal scope of the potential integrated access solution.

The overall mind map of the solution architecture (see 6.3.6) shows that the creation of the integrated access solution has many alternatives and it can be considered as a set of several building blocks divided between back-end side, i.e. storage, content and context management tools and processes and front-end side, i.e. the web application with the relevant user interfaces and possibly applications (mobiles & tablets).

This chapter describes the selected alternatives. No technological details are mentioned (e.g. type of search engine, type of database or hardware/software specifications).

These alternatives must be considered as "big picture models" and of course they may be developed and elaborated in various ways. Nevertheless, key management decisions with primary attention paid to the feasibility aspects are necessary, for example if the integrated access solution would be focused only on the future documents, or also documents from the past and, if so, to what extent.

The description of each integrated access solution alternative is realised in the following simple methodological breakdown:

1. Basic characteristics of the solution

The description of the technical aspects of the proposed solution is based on the choice of their building blocks and the aspects of their implementation.

2. High-level architecture block scheme

The illustration of the essential parts of the solution architecture in the block scheme, visually compatible with the integrated access solution vision block scheme (see Figure 48).

3. Organisational aspects

This part of the solution description summarizes the impacts on the existing document sources' stakeholders and the future integrated access solution staff.

It is necessary to know that any innovation has practical implications for the future life of the organisation which introduces this innovation. Typically, this means some process of change. Ideally, the innovation is implemented seamlessly without affecting the organisation or work of the participants. However, it is obvious that such ideal cases very rarely happen in practice.

4. Implementation time and estimation of the overall cost

The overall assessment of the timing and cost aspects.

5. Conclusion in the form of a SWOT diagram

The SWOT diagram is used for an overview of the specific integrated access solution, i.e. 2*2 matrix, where

- Upper row attributes origins are **internal**;
- Lower row attributes origins are **external**;
- Left column attributes are **helpful** to achieving the objectives;
- Right column attributes are **harmful** to achieving the objectives.

The SWOT template used for an evaluation of the possible future integrated access solution alternatives is depicted in Figure 51.

Figure 51: SWOT template

6.4.1. Alternative No. 1: Web application built upon CELLAR

CELLAR is the general content and metadata repository of the Publications Office. It is equipped with the RESTful interface for both content and metadata.

The basic method of transferring the data (context and content) from document sources is IMMC Core Metadata Exchange Protocol as described in section 6.3.1.

New improvements are added to CELLAR functionality continuously. Nowadays the main branch of these improvements is focused on document sources' stakeholders. It should bring them new possibilities, tools and scenarios for the smooth integration of their document sources with CELLAR. In the end the document sources' stakeholders (institutions and agencies) should be able to transfer their documents' content and context seamlessly. Ideally without the necessity of changing their IT systems while the CELLAR team will prepare for them particular tools and methods.

Alternative No. 1 adds a new, independent integrated access solution web application to CELLAR. It would be placed at a new URL address with maximum use of the options provided by CELLAR.

6.4.1.1. Basic characteristics of the solution

The basic characteristics of the solutions are:

- Document context (compact set of metadata of the document) would be stored in CELLAR.
- Document content issues
    - Content harmonisation would be done at the level of the document sources:
        - Document sources' stakeholders could decide whether they invest in the conversion of their documents (mostly published as file attachments) in the structured form or leave them as they are.
        - Document sources' stakeholders can change their decision anytime in the future.
    - There are two basic options for storing the content of the document:
        - Content of the documents could be stored in CELLAR (preferred option).
        - Documents could be stored in their original repositories (document sources) while CELLAR would only handle the references to them (less preferred option).
- Obtaining the document sources' content and context and copying this to CELLAR would be processed:
    - Either via the IMMC Core Metadata Exchange Protocol (new implementations or improvements of the implementations already under operation);
    - Or via newly developed simplified methods focused on document sources for which full implementation of IMMC Core Metadata Exchange Protocol would be too complicated.
- Selected context and content would be synchronised from CELLAR into the integrated access solution database in a unidirectional way.
    - The information required by the integrated access solution would be defined by the common knowledge model (simplified ontology developed for presentation purposes)
    - The core component of this solution is the synchronisation process between CELLAR and the integrated access solution database, i.e. the replication of content in a structured format and the related subsets of the metadata that are needed to comply with the common knowledge model would be replicated from CELLAR to the integrated access solution database.

- Documents that are not physically stored in CELLAR but referred to by a link would have to be crawled by the integrated access solution during in the scope of the synchronisation process.

(General note: Improvements of CELLAR e.g. notification of changes to the context pieces (single metadata) would be welcome because they may make the synchronisation more robust.)

- A new front-end web-application providing content from the integrated access solution database to the end-users would be built from the following building blocks:
  - Common knowledge model (ontology) based intuitive search & document retrieval
  - Personalisation (wide possibilities)
  - Continual improvements of the integrated access solution based on the processing of users' activities
  - Adjustments for general purpose search engines
- A new, robust server infrastructure for the integrated access solution database and front-end web application will be necessary.

6.4.1.2. High-level architecture block scheme

Figure 52 shows the block scheme of the high-level architecture of alternative No. 1 of the integrated access solution – New front-end application built upon CELLAR.

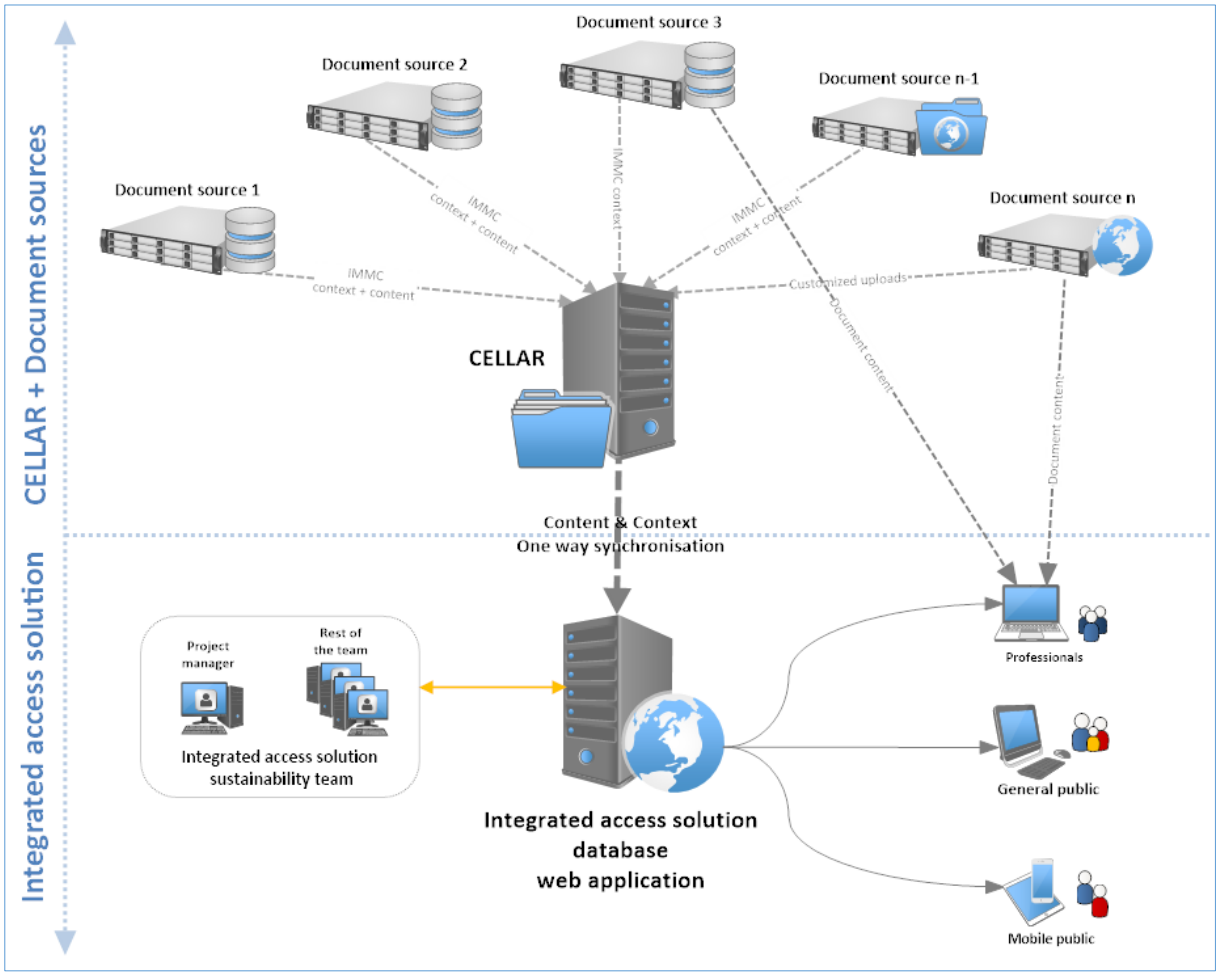


Figure 52: Alternative No. 1 – New front-end application built upon CELLAR

6.4.1.3. Organisational aspects

This alternative will be introduced by the detailed feasibility study defining mainly the scope and goals between relevant stakeholders/solution providers/suppliers and the detailed clarification of integration requirements.

Each institution/agency would establish the transfer channel for pushing both context and content from their document sources into CELLAR.

In some serious cases, the document content could remain in the document source's storage while only its reference would be pushed to CELLAR.

The full implementation of the IMMC Core Metadata Exchange Protocol would be ideal however and another simpler push transfer method could be negotiated with the CELLAR team. For instance, the institution/agency could hook its database through a user interface and the CELLAR team would take care of the rest (e.g. extension of the Common Data Model and the subsequent technical implementation like generating formal packages needed for input the document source context and content into CELLAR behind the scenes). It would be pragmatic to undertake each implementation of any institution/agency transfer channel at their own expense as part of the development of its information systems.

The synchronisation developer on the integrated access solution side will be provided with sufficient information to create the synchronisation routines for CELLAR to the integrated access solution, i.e. including the meanings of each metadata.

The initial development and implementation of the integrated access solution, including the development of unidirectional synchronisation between CELLAR and the integrated access solution database would be made by the selected supplier of the integrated access solution. Technical solutions of the document sources are subject to permanent development, and so it must be assumed that for permanent sustainability, a newly established integrated access solution sustainability team is required.

6.4.1.4. Estimations of the implementation time

All estimations of implementation times are based on experience with similar projects.

Overall estimations

Creation of the detailed feasibility study is necessary and will last ca 12 months.

For the purposes of this study eleven out of twelve document sources (which were evaluated in the analytical part of this study as re-usable for the future integrated access solution, see 4.4.5) will be involved in time and cost estimations. Additionally, CELLAR is considered to be a target data storage.

- Document sources have either implemented IMMC transfers partially or have not started with the implementation yet (see Table 12 in Chapter 4.2.10.1.1 and Chapter 4.4.5), that means 10 implementations should be carried out. EUR-Lex is in a special position because it is not an input for CELLAR but an output.
- The other document sources would decide either to implement IMMC transfers (or finish their implementation) or to implement new tools and methods transfer which are under development by the CELLAR team.

Document sources side

- Estimation of the implementation of each of 10 transfer channels "document source TO CELLAR":
  - 48 weeks including definitions, negotiations with the CELLAR team, necessary development and testing where 50% of the net time is supposed, that means 12 weeks * 2 team members (analysis + development) = 48 weeks = 240 man-days for all activities;
  - 8 weeks of time backup should be added, that means 40 man-days
  - Summary of manpower needed for one implementation: 56 weeks = 280 man-days
- The implementation will be planned and organized by the CELLAR team
- Summary of manpower necessary: 10 document sources * 280 man-days = 2 800 man-days

CELLAR side

- Three full time members are projected to be necessary for each transfer channel development team on the CELLAR side:
  - 1 project manager taking care of the communication with document sources' stakeholders
  - 1 developer
  - 1 tester
- Transfer channel development and execution projection on the CELLAR side:
  - It would be practical to plan the development of transfer channels in a sequential way rather than in parallel, with some overlap between the periods allocated for the development of specific channels. This would mean the development and implementation of a transfer channel from one document source to CELLAR every 8 weeks.
  - This is 80 weeks in total (10 document sources á 8 weeks) + 24 weeks (30% reserve) for solving exceptional and non-standard situations or unexpected delays = 104 weeks, or 520 man-days of implementation.
  - Summary of necessary man-days: 3 team members * 520 man-days = 1 560 man-days

Integrated access solution side

- Integrated access solution development timing projection:
  - Integrated access solution project definition, negotiations, tuning: 20 weeks
  - Integrated access solution analyses & development & tuning: 104 weeks
  - Integrated access solution testing: 104 weeks
  - Time backup for non-standard situations and risk processing: 20 weeks
  - Summary of timing: 144 weeks (~ 36 months)
- Integrated access solution development team work effort projection:
  - 7 team members, 144 weeks each, 1 008 weeks all
    - 1 project manager
    - 1 database developer
    - 1 user interface designer
    - 1 context & content manager
    - 3 developers
  - 1 team member (tester): 124 weeks
  - Summary of timing: 1132 weeks = 5 660 man-days

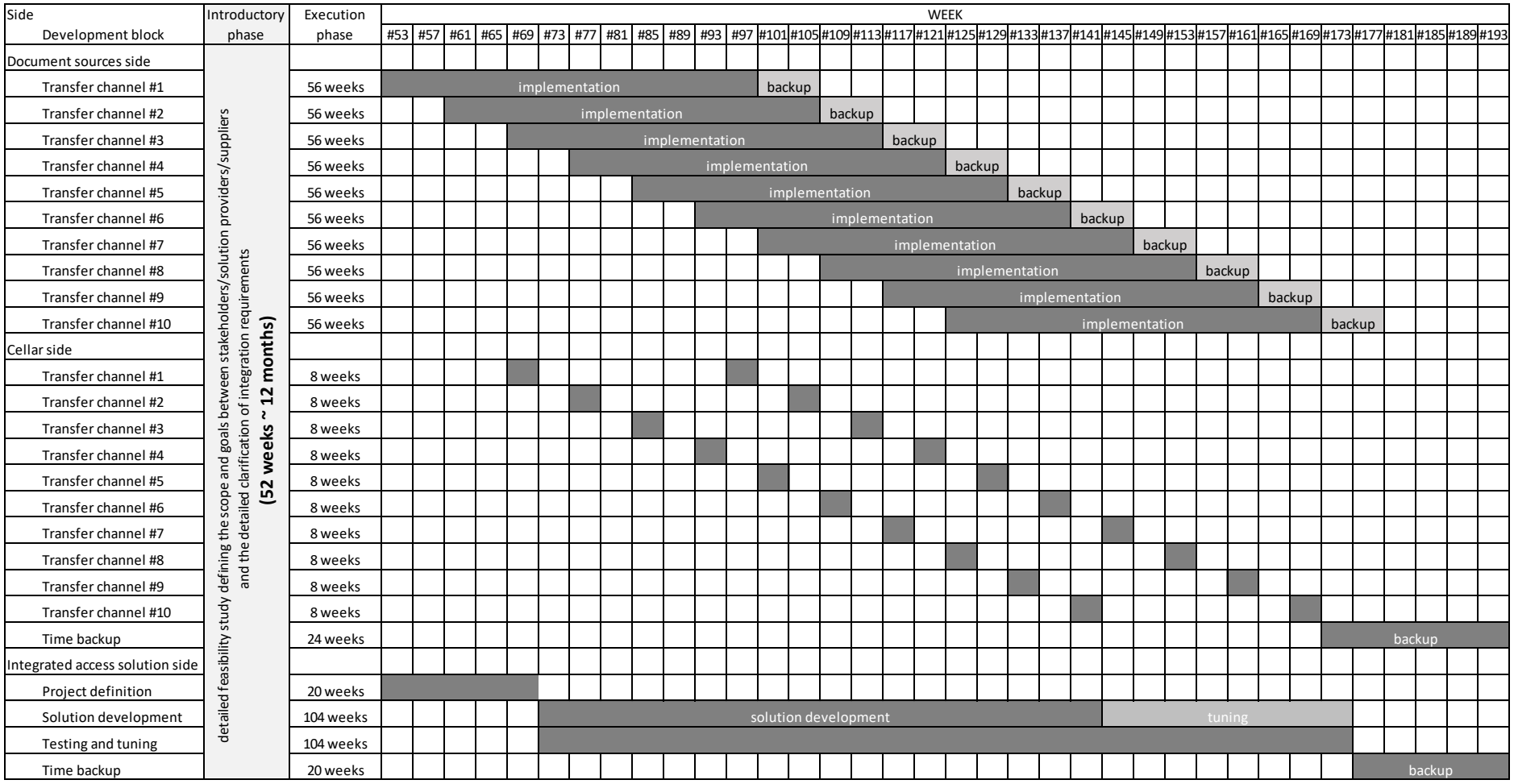


Figure 53: Integrated access solution deployment schedule projection – alternative No. 1

6.4.1.5. Estimations of the overall cost

The estimations of overall costs are based on the estimations of the implementation time multiplied by the estimation of the man-day cost.

Rough overall estimations of integrated access solution alternative No. 1 implementation costs:

- Document sources side:
  - Manpower summary estimations
    - Summary of man-days: 2 800
    - Estimation of the costs of one man-day: 600 €
    - Estimation of summary costs: **1 680 000 €**
- CELLAR side
  - Manpower summary estimations
    - Summary of man-days: 1 560
    - Estimation of the costs of one man-day: 600 €
    - Estimation of summary man-power costs: **936 000 €**
- Integrated access solution side:
  - Manpower summary estimations
    - Summary of necessary man-days: 5 660
    - Estimation of the costs of one man-day: 600 €
    - Estimation of summary man-power costs: **3 396 000 €**
  - Infrastructure estimation:
    - Hardware: 200 000 €
    - Software licences (e.g. operating systems and databases in the case of not using open source software): 200 000 €
    - Deployment (10%): 40 000 €
    - Estimation of summary costs: **440 000 €**
- **Overall cost estimations: 6 452 000 €**

Operational aspects

An integrated access solution sustainability team will be required, composed of the following roles:

- Project manager
- Database & ontology administrator
- Web application administrator
- User feedback operator

6.4.1.6. Conclusion (SWOT diagram)

Conclusions on the integrated access solution alternative No. 1 - New front-end application built upon CELLAR is shown in Figure 54 in the form of a SWOT diagram

Figure 54: Alternative No. 1 SWOT diagram - New front-end application built upon CELLAR

6.4.2. Alternative No. 2: Decentralized aggregated searches

Alternative No. 2 adds a new independent web application, which would provide the users with direct searching in the document sources using a sequential querying of the set of search engine gateways and consequent processing of their responses.

6.4.2.1. Basic characteristics of the solution

- The integrated access solution as per this alternative would consist of:
  - A set of decentralised search engine gateways implemented on the top of each document source (including CELLAR) where interoperability of the search engine gateways should be guaranteed (nowadays such search engines are available as an open source and they are approved in many implementations while a few years ago they were not).
  - A PA Web application which would pass a series of queries composed according to the end-user's requirement to particular search engine gateway(s), process their responses and then present the results in a unified form to the device on the end-user's side.
- All document data (context and content) would remain on the document sources' side while the integrated access solution would serve as the presentation layer of the data which document sources' stakeholders decide to provide to the public via the integrated access solution:
  - No major infrastructure improvement is necessary on the integrated access solution side except of one central PA Web application server.
  - However, this solution requires significant technical improvements on the document sources side (newly established search engine gateways) to be able to process end-users' requests; API both for accepting the requests and providing the responses (JSON or XML) within each document source will be on the part of the particular search engine.
- The user interface of the new front-end web-application could be composed as a wizard which guides the users in a few steps through the available options and allows them to precisely define their requirements:
  - A wizard would be built on the simplified ontology composed according to the previously defined WHAT/WHO/WHEN/WHY paradigm.
- A new front-end web-application, providing content from the integrated access solution database to the end-users, would be built from the following building blocks:
  - A component for querying the document sources' search engine gateways and processing their responses is the most important part of this alternative on the integrated access solution side;
  - Basic personalisation (such as stored favourite searches or notification of new results preselected by the users);
  - Processing of users' activities.

6.4.2.2. High-level architecture block scheme

Figure 55 shows the block scheme of the high-level architecture of the alternative No. 2 – Decentralised aggregated search.

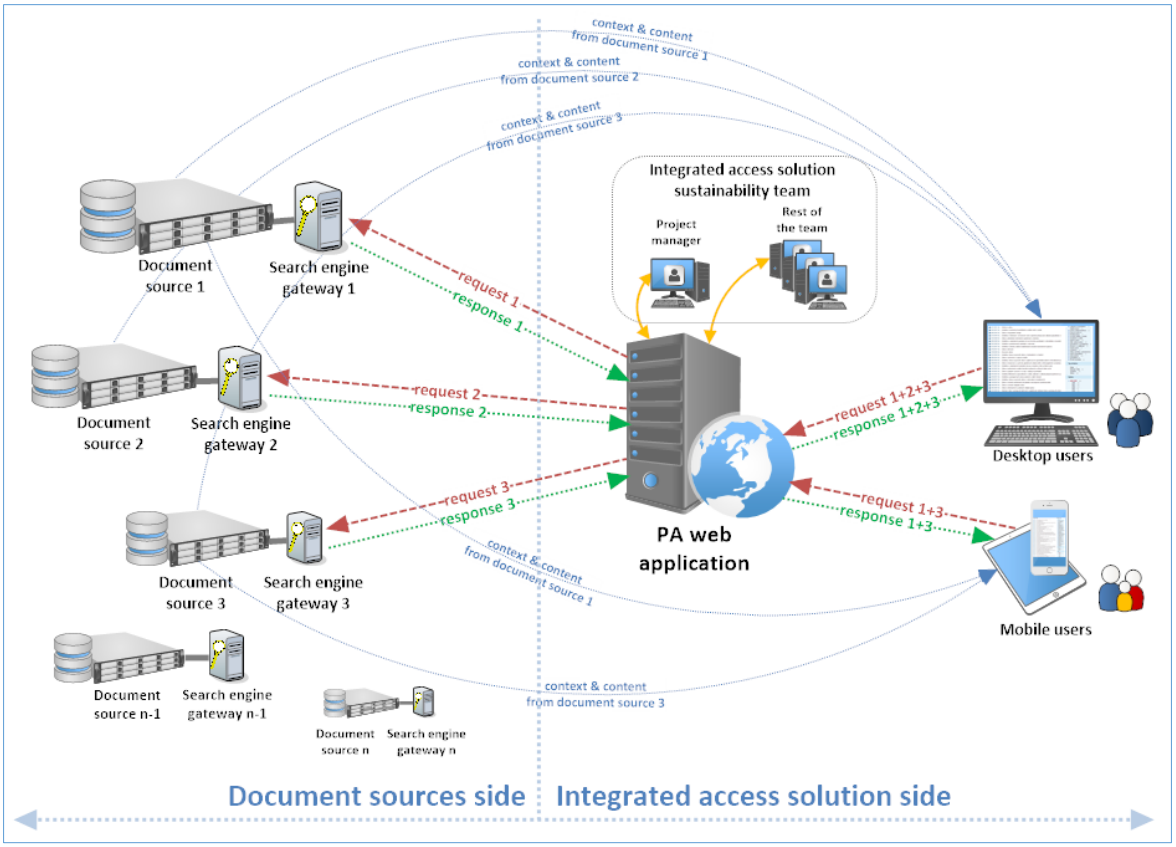


Figure 55: Alternative No. 2 – Decentralised aggregated search

6.4.2.3. Organisational aspects

Alternative No. 2 moves a significantly large part of the implementation of the integrated access solution to the document sources' stakeholders (institution/agencies). It would be only feasible under the condition that each document source would be equipped with a search engine gateway providing an API interface, which allows one to send the requests and to obtain the results in a form which will then be converted to provide the end-users with these results in a more user-friendly form. It would be pragmatic to undertake each implementation of any institution/agency at their own expense as part of the development of its information systems.

It is necessary to introduce this alternative by the detailed feasibility study defining mainly the scope and goals between relevant stakeholders/solution providers/suppliers and the detailed clarification of integration requirements.

From the point of view of the integrated access solution, the necessary development would be carried out - the creation of the user interfaces and managing the data exchange between the integrated access solution and the document sources' search engine gateways.

The initial development and implementation of the integrated access solution including the cooperation with the document sources' stakeholders by implementation of the search engine gateways would be carried out by the selected supplier of the integrated access solution.

Both the data and IT solutions on the document sources' side are subject to permanent development, and so for the care of permanent sustainability, it must be assumed that a dedicated integrated access solution sustainability team would be needed.

6.4.2.4. Estimations of the implementation time

Overall estimations

Creation of the detailed feasibility study is necessary and will last ca 6 months.

12 document sources (which were evaluated in the analytical part of this Study as re-usable for the integrated access solution, see Chapter 4.4.5) would be involved:

- Each document source has to implement search engine gateway;
- In the Publications Office the search engine gateway will be implemented upon CELLAR instead of EUR-Lex.

Document sources side

- Each document source is unique, therefore, the implementation of the search engine gateway would be a specific project as well, however the unified interface of the gateway will be necessary.
- It may be assumed that, at least in some document sources, it would be necessary to change the IT infrastructure to achieve the ability to implement the search engine gateway as required by the interfaces.
- This could lead to a large and very diverse time effort.
- Estimated time required: up to 18 months with sufficient support (incl. political support and tender organisation support) should be enough for most document sources for the implementation of search engine gateways:
    - Implementations will be carried out in parallel.
- Net time and work force necessary for each implementation would vary so it could be estimated only very roughly:
    - 1 project manager: 52 weeks
    - 1 developer of the data feeders and parsers: 26 weeks
    - 1 search engine gateway developer: 26 weeks
    - Summary of manpower needed for one implementation: 104 weeks = 520 man-days
    - Summary of manpower needed for twelve search engine gateways implementations: 6 240 man-days
- Infrastructure aspects:
    - It is assumed that the search engine gateway will be built on open source software which does not require any additional investment to purchase.
    - Additional investments into hardware equipment is needed where rough average estimation per search engine gateway is 20 000 €.
    - Summary of hardware investment for twelve search engine gateways: 240 000 €.

Integrated access solution side

- Integrated access solution development team projection:
  - Integrated access solution project definition, negotiations, tuning: 20 weeks
  - Integrated access solution development: 52 weeks
    - Within this time frame the search engine gateway implementation at the document sources' side specified above will be carried out
  - Integrated access solution testing and tuning: 26 weeks
  - Time backup for non-standard situations and risks processing: 12 weeks
  - Summary of time necessary for one implementation: 92 weeks (~ 23 months)
- Integrated access solution development team work effort projection:
  - 1 project manager: 92 weeks
  - 1 database developer: 26 weeks
  - 1 user interface designer: 26 weeks
  - 3 web application developers: 92 weeks each, 276 weeks all
  - 1 tester: 26 weeks
  - Time backup 12 weeks for all 7 team members: 84 weeks
  - Summary of timing: 530 weeks = 2 650 man-days

Side / Development block	Introductory phase	Execution phase	WEEK #27	#31	#35	#39	#43	#47	#51	#55	#59	#63	#67	#71	#75	#79	#83	#87	#91	#95	#99	#103	#107	#111	#115
Document sources	detailed feasibility study defining the scope and goals between stakeholders/solution providers/suppliers and the detailed clarification of integration requirements **(26 weeks ~ 6 months)**																								
Search engine gateway #1		52 weeks					development													time backup					
Search engine gateway #2		52 weeks					development													time backup					
Search engine gateway #3		52 weeks					development													time backup					
Search engine gateway #4		52 weeks					development													time backup					
Search engine gateway #5		52 weeks					development													time backup					
Search engine gateway #6		52 weeks					development													time backup					
Search engine gateway #7		52 weeks					development													time backup					
Search engine gateway #8		52 weeks					development													time backup					
Search engine gateway #9		52 weeks					development													time backup					
Search engine gateway #10		52 weeks					development													time backup					
Search engine gateway #11		52 weeks					development													time backup					
Search engine gateway #12		52 weeks					development													time backup					
Integrated access solution side																									
Project definition		20 weeks																							
Solution development		52 weeks						solution development																	
Testing and tuning		26 weeks																							
Time backup		12 weeks																							

Figure 56: Integrated access solution deployment schedule projection – alternative No. 2

6.4.2.5. Estimations of the overall cost

The estimations of the overall costs are based on the estimations of the implementation time multiplied by the estimation of the man-day cost.

Rough overall estimations of integrated access solution alternative No. 2 implementation costs:

- Document sources side:
  - Manpower summary estimations:
    - Summary of manpower needed for twelve search engine gateway implementations: 6 240 man-days
    - Estimation of the costs of one man-day: 600 €
    - Estimation of summary costs: **3 744 000 €**
  - Infrastructure estimation:
    - Overall hardware investment for twelve search engine gateways: **240 000 €**
- Integrated access solution side:
  - Manpower summary estimations:
    - Summary of necessary man-days: 2 650
    - Estimation of the costs of one man-day: 600 €
    - Estimation of summary costs: **1 590 000 €**
  - Infrastructure estimation:
    - Hardware: 100 000 €
    - Software licences (e.g. operating systems and databases in the case of not using open source software): 100 000 €
    - Deployment (10%): 20 000 €
    - Estimation of summary costs: **220 000 €**
- **Overall cost estimations: 5 794 000 €**

Operational aspects

An integrated access solution sustainability team will be required, composed of the following roles:

- Project manager
- Database & ontology administrator
- Web application administrator
- User feedback operator

6.4.2.6. Conclusion (SWOT diagram)

Conclusions on integrated access solution alternative No. 3 – Decentralised aggregated search is shown in Figure 57 in the form of a SWOT diagram.

Figure 57: Alternative No. 2 SWOT diagram - Decentralised aggregated search

6.4.3. Alternative No. 3: Centralised aggregated search

Alternative No. 3 adds a new independent web application, which would provide the users with direct document source searching provided by a newly established central search engine, in this alternative of the integrated access solution.

6.4.3.1. Basic characteristics of the solution

- The integrated access solution per this alternative would consist of:
  - A newly established central search engine which indexes' would be fed by the document sources' APIs implemented on top of each document source.
  - A PA Web application, which would
    - pass a query composed according to end-user's requirements to the central search engine
    - process its response and push them to the devices at the end-user's side in a unified form.
- All content would remain on the document sources' side while the integrated access solution would serve as the presentation layer of the data, which document sources' stakeholders decide to provide to the public via the integrated access solution:
  - Improvement of the technical infrastructure would be necessary on the integrated access solution side.
  - Significant technical improvements on the document sources side would be also necessary to feed the central search engine with data:
    - Establishing API both for accepting the requests and providing the answers (JSON or XML) within each document source would be the best possibility;
    - Document sources' stakeholders would control these APIs and decide which data will be provided to the central search engine;
    - As stated in the conclusion of the analytical section of the study, only EUR-Lex is currently equipped with such API.
- The user interface of the new front-end web-application could be composed as a wizard which guides the users in a few steps through the available options and allows them to precisely define their requirements:
  - This wizard will be built on the simplified ontology composed according to the previously defined WHAT/WHO/WHEN/WHY paradigm.
  - Because of the use of one central search engine the development of such a wizard will be easier than in alternative No. 2.
- A new front-end web-application providing content from the integrated access solution database to the end-users will be built from the following building blocks:
  - Central search engine, indexing all the document sources' context and content;
  - Basic personalisation (such as stored favourite searches or notification of new results preselected by the users)
  - Processing of users' activities

6.4.3.2. High-level architecture block scheme

Figure 58 shows the block scheme of the high-level architecture of the alternative No. 3 – Centralised aggregated search.

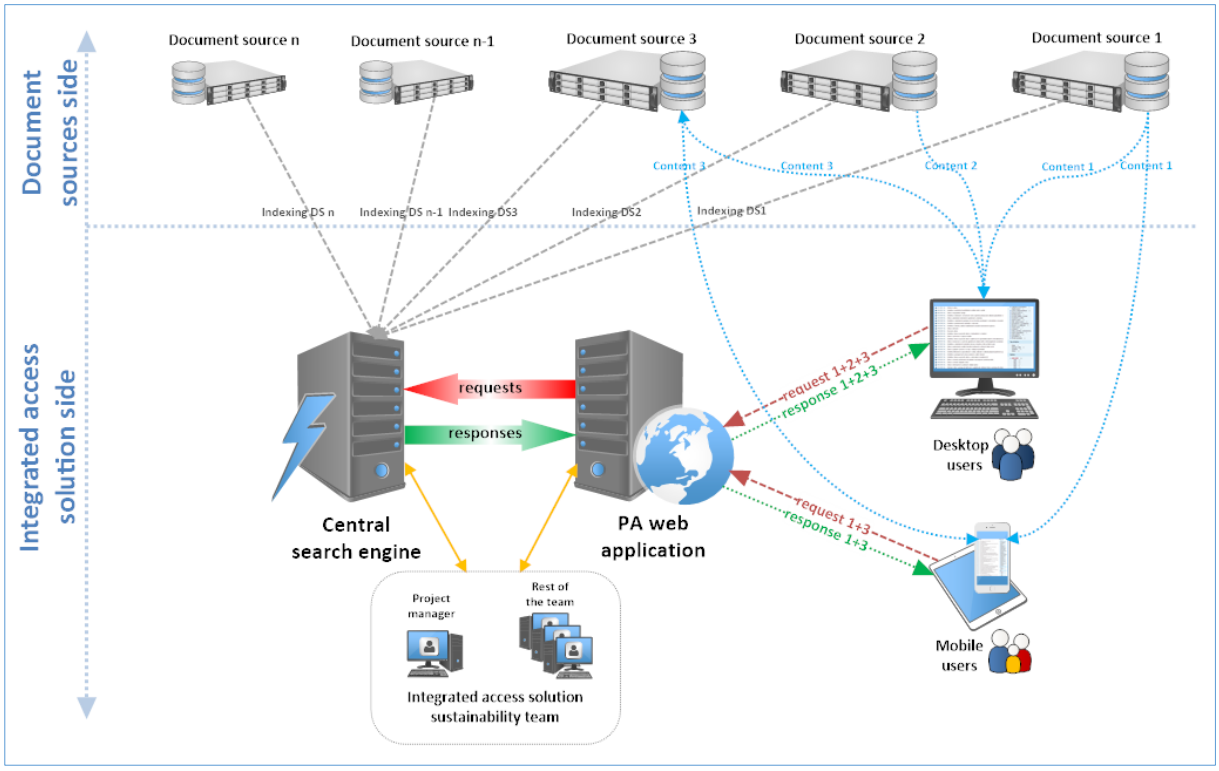


Figure 58: Alternative No. 3 – Centralised aggregated search

6.4.3.3. Organisational aspects

Alternative No. 3 would pass a large part of the implementation of the integrated access solution to the document sources' stakeholders (institution/agencies). It would only feasible under the condition that each document source will be equipped with an API interface able to feed the central search engine indexes. It would be again pragmatic to undertake each implementation of any institution/agency at their own expense as part of the development of its information systems.

It is necessary to introduce this alternative by the detailed feasibility study defining mainly the scope and goals between relevant stakeholders/solution providers/suppliers and the detailed clarification of integration requirements.

From the point of view of the integrated access solution development, the intensive coordination and cooperation with the document sources' stakeholders by implementation of their APIs would be essential. This coordination and cooperation would be led by the selected supplier of the integrated access solution.

Both the data and IT solutions on the document sources' side are subject to permanent development, and so for the care of permanent sustainability, it must be assumed that a dedicated integrated access solution sustainability team would be needed.

6.4.3.4. Estimations of the implementation time

Overall estimations

Creation of the detailed feasibility study is necessary and will last ca 6 months.

12 document sources (which were evaluated in the analytical part of this Study as re-usable for the future integrated access solution, see Chapter 4.4.5) would be involved:

- Each document source should implement the API providing both context and content to the central search engine indexing processes.
- In the Publication office, the necessary APIs both for TED and Eur-Lex are working and can be used by the integrated access solution central search engine, which means that 10 implementations should be carried out.

Document sources side

- Each document source is unique, therefore, the creation of API for each document source would be a specific project as well.
- It may be assumed that, at least in some document sources, it will be necessary to change the IT infrastructure to gain the ability to implement the API.
- This could lead to a large and very diverse time effort.
- Estimated time required: up to 12 months with sufficient support (incl. political support and tender organisation support) should be enough for most document sources to build an API interface:
    - Implementations would be carried out in parallel.
- Net time and work force necessary for each implementation would vary and it could be estimated only very roughly:
    - 1 project manager: 32 weeks
    - 2 API developers: 16 weeks each, 32 weeks both
    - Summary of manpower needed for one implementation: 64 weeks = 320 man-days
    - Summary of manpower needed for 10 API implementations: 3 200 man-days
- Infrastructure aspects:
    - It could be estimated that half of the document sources will need to purchase new hardware equipment for the API implementations needs – on average 20 000 €
    - Summary of hardware investment for 5 document sources: 100 000 €

Integrated access solution side

The necessary development for this alternative could be estimated as a 10% less time demanding than in the previous alternative No. 2.

- Integrated access solution development team projection:
    - Integrated access solution project definition, negotiations, tuning: 20 weeks
    - Integrated access solution development: 52 weeks
        - Within this time frame search engine gateway implementations at the document sources' side, as specified above will be carried out
    - Integrated access solution testing and tuning: 16 weeks
    - Time backup for non-standard situations and risks processing: 12 weeks
    - Summary of timing: 84 weeks (~ 21 months)
- Integrated access solution development team work effort projection:

- 1 project manager: 84 weeks
- 1 database developer: 26 weeks
- 1 user interface designer: 26 weeks
- 3 developers (central search engine + web application):
  84 weeks each, 252 weeks all
- 1 tester: 16 weeks
- Time backup 12 weeks for all 7 team members: 84 weeks
- Summary of timing: 488 weeks = 2 480 man-days

Side Development block	Introductory phase	Execution phase	WEEK #27	#31	#35	#39	#43	#47	#51	#55	#59	#63	#67	#71	#75	#79	#83	#87	#91	#95	#99	#103	#107
Document sources	detailed feasibility study defining the scope and goals between stakeholders, solution providers, suppliers and the detailed clarification of integration requirements **(26 weeks ~ 6 months)**																						
API interface #1		32 weeks					development								time backup								
API interface #2		32 weeks					development								time backup								
API interface #3		32 weeks					development								time backup								
API interface #4		32 weeks					development								time backup								
API interface #5		32 weeks					development								time backup								
API interface #6		32 weeks					development								time backup								
API interface #7		32 weeks					development								time backup								
API interface #8		32 weeks					development								time backup								
API interface #9		32 weeks					development								time backup								
API interface #10		32 weeks					development								time backup								
Integrated access solution sid																							
Project definition		20 weeks																					
Solution development		52 weeks						solution development															
Testing and tuning		16 weeks																					
Time backup		12 weeks																					

Figure 59: Integrated access solution deployment schedule projection – alternative No. 3

6.4.3.5. Estimations of the overall cost

The estimations of the overall costs are based on the estimations of the implementation time multiplied by the estimation of the man-day cost.

Rough overall estimations of integrated access solution alternative No. 3 implementation costs:

- Document sources side:
  - Manpower summary estimations:
    - Summary of manpower needed for ten API implementations: 3 200 man-days
    - Estimation of the costs of one man-day: 600 €
    - Estimation of summary costs: **1 920 000 €**
  - Infrastructure estimation:
    - Overall hardware investment for five document sources: **100 000 €**
- Integrated access solution side:
  - Manpower summary estimations:
    - Summary of necessary man-days: 2 480
    - Estimation of the costs of one man-day: 600 €
    - Estimation of summary costs: **1 488 000 €**
  - Infrastructure estimation:
    - Hardware: 200 000 €
    - Software licences (e.g. operating systems and databases in the case of not using open source software): 200 000 €
    - Deployment (10%): 40 000 €
    - Estimation of summary costs: **440 000 €**
- **Overall cost estimations: 3 948 000 €**

Operational aspects

An integrated access solution sustainability team will be required, composed of the following roles:

- Project manager
- Database & ontology administrator
- Web application administrator
- User feedback operator

6.4.3.6. Conclusion (SWOT diagram)

Conclusions on integrated access solution alternative No. 3 – Centralised aggregated search is shown in Figure 60 in the form of a SWOT diagram.

Figure 60: Alternative No. 3 SWOT diagram - Centralised aggregated search

6.4.4. Alternative No. 4: Content harmonisation on a central level

In fact, alternative No. 4 is based on alternative No. 1 however provides improvements. The main differences are:

1. The integrated access solution in this alternative would take care of the content harmonisation – content typically provided by the document sources in the form of the file attachments would be converted to the unified structured form (HTML5) by the tools newly developed for this alternative.
2. The advantages of the structured form would then be fully utilised in the user interfaces of the PA Web application – this would ensure maximal benefits for the end-users, providing them with new possibilities for displaying and further working with the documents brought to them by integrated access solution.

The main source of the context and content would be the CELLAR as in alternative No. 1. Transferring the document sources' data to CELLAR would depend highly upon the document source stakeholders' willingness to cooperate. In case of technical obstacles which may block cooperation between a particular document source and CELLAR, this alternative would try to establish its own methods on how to eliminate these obstacles – especially in developing particular transfer channels based on the PULL transfer method (see 6.3.1.2).

6.4.4.1. Basic characteristics of the solution

- The integrated access solution per this alternative would consist of:
  - A robust PA database:
    - for storing all context and content (harmonised into the unified form) obtained from the document sources;
    - for storing all data concerning the usage of the integrated access solution by users.
  - Content and context management tools (ontology management) for ensuring seamless integrated access solution operation:
    - Details described in the chapter *Intuitive management of both the content and the context* (see 6.3.4).
  - A PA Web application which would pass a query, composed according to end-user's requirements, to the PA database, process its response and push them to the devices at the end-user's side in a unified form.
- Document content issues:
  - All data from the document sources would be stored in the PA database organised according to integrated access solution ontology:
    - The principles of the ontology would be based on the previously defined WHAT/WHO/WHEN/WHY paradigm based on *PA ontology vision* (see 6.3.2.1)
      - It may appear that this new ontology duplicates either CELLAR or the ontology in the alternative No. 1.
      - In fact this is true only partially because this ontology is intended for presentation purposes and assumes wide usage of natural language processing only while the CELLAR approaches are much wider.
  - The tools and methods for harmonisation of the content on the central level (**the main difference and main advantage to the previous alternatives**) would be implemented:

        - This means conversion of the content currently existing in the file attachment form into re-usable formats (HTML5 preferred) enabling additional benefits to the end-users.
        - Details described in the chapter *Automated content parsing* (see 6.3.3).
- PA database would be fed:
    - Preferably from CELLAR where data would be pushed by document sources similar to in alternative No. 1;
    - In exceptional cases, directly from document sources by establishing dedicated PULL transfer channels; this method is not preferred and would be used in case of technical obstacles on the document sources' side by implementing transfer channels offered by CELLAR.
- A PA web application providing the results to end-users, fulfilling requirements, and applying all useful innovations specified in the chapter *Web and mobile application* (see 6.3.5), namely:
    - Maximal ergonomics and usability for the end-users
    - Intuitive search process and list of results
    - Intuitive detail of the document
    - Personalisation (wide possibilities)
    - Continual adjustments of the integrated access solution based on processing of users' activities
    - Adjustments for general purpose search engines
- A new robust server infrastructure both for PA database and the PA web application would be necessary

6.4.4.2. High-level architecture block scheme

Figure 61 shows the block scheme of the high-level architecture of alternative No. 4 of the integrated access solution, which follows the architecture outlined in chapter 6.3.

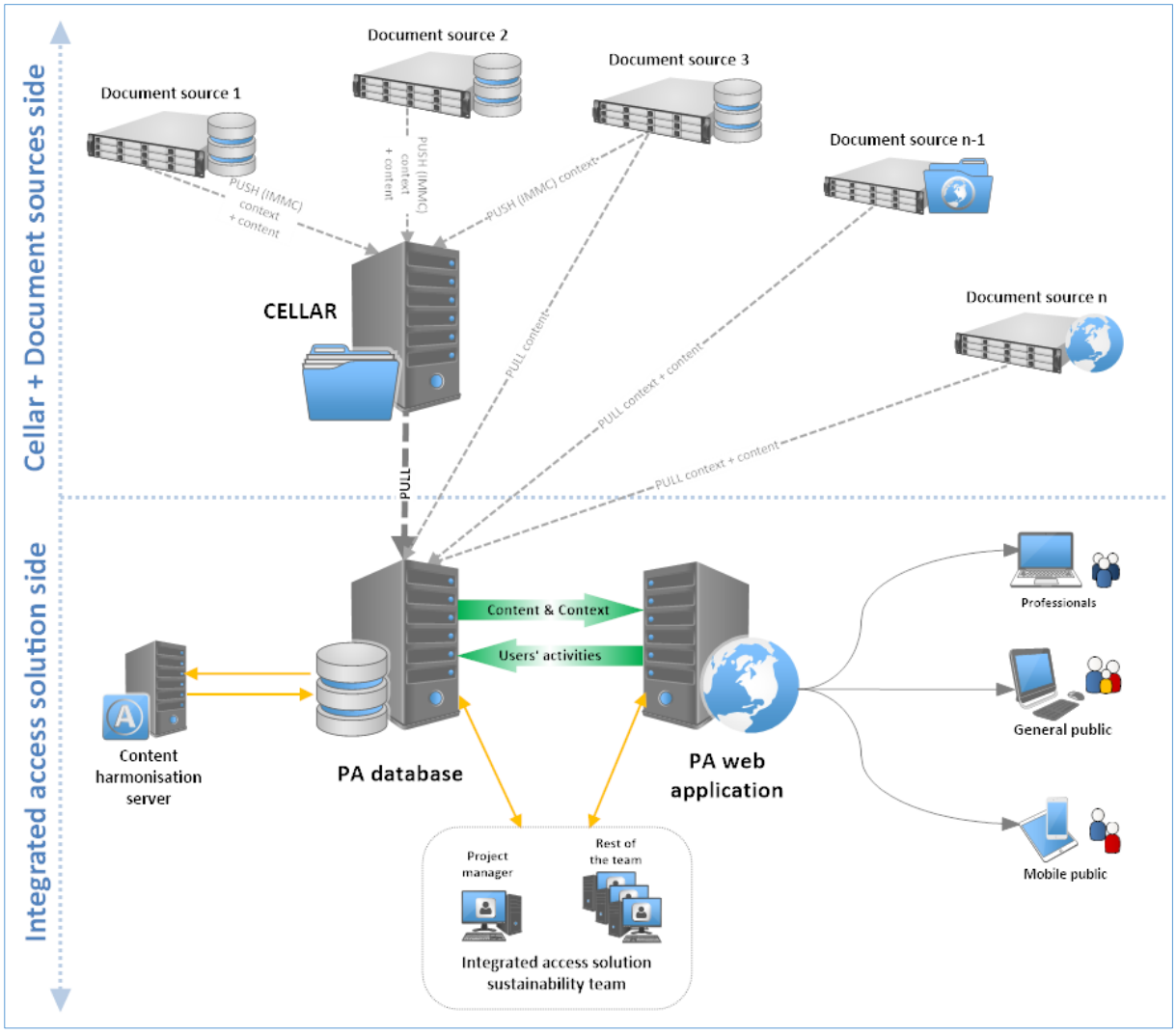


Figure 61: Alternative No. 4 – Content harmonisation on a central level

6.4.4.3. Organisational aspects

Alternative No. 4 is the most challenging of all the previous alternatives in terms of the work effort with both the content and context. It will be introduced by the detailed feasibility study defining mainly the scope and goals between relevant stakeholders/solution providers/suppliers and the detailed clarification of integration requirements.

It is obvious that this alternative of integrated access solution would have a gradual development, proportional to the progress of developments on the content harmonisation methods. In fact, these methods and their utilisation represent the main benefits of this alternative. However, successful deployment of these methods is a “big issue” and it should be carefully decided what types of documents they would cover and how far into the past they should go.

It is again pragmatic to undertake each implementation of any institution/agency at their own expense as part of the development of its information systems.

This alternative would assume the creation of a strong integrated access solution sustainability team, who should take care of:

- Cooperation in resolution of the operational and development issues with the CELLAR team during the obtaining the content and context of the document sources;
- Cooperation with the document source stakeholders in selected cases where their data will not be pushed to CELLAR;
- Content harmonisation maintenance and continual improvements;
- Ontology maintenance and continual improvements;
- Improvements arising from analyses of users' activities.

The development and implementation of the integrated access solution, including the conversion of the content, will be carried out by the selected supplier of the integrated access solution.

6.4.4.4. Estimations of the implementation time

Overall estimations

Creation of the detailed feasibility study is necessary and will last ca 18 months.

For the purposes of this study eleven out of twelve document sources (which were evaluated in the analytical part of this study as re-usable for the future integrated access solution, see 4.4.5) will be involved in time and cost estimations. Additionally, CELLAR is considered to be a target data storage and EUR-Lex is in a special position because it is not an input for CELLAR but an output.

For the rough time and cost estimation purposes the following assumptions were made for 10 document sources:

- Data from 8 document sources will be obtained from CELLAR;
- Data from 2 document sources will be obtained directly from the document sources.

Because the grounds of this alternative are based on alternative No. 1, the basic time and cost estimations in the CELLAR part are identical to alternative No. 1 for each one of 8 document sources considered. Obtaining the data from the rest 2 document sources will be solved as a part of integrated access solution development.

It is necessary to note that the estimated distribution (i.e. for 8 document sources the data are obtained from CELLAR while for the other 2 document sources the data are obtained directly) is only an assumption. Nevertheless, in some cases obtaining content directly from the document sources could be faster and more feasible than the unification via CELLAR. The solution will be clearly defined in the feasibility study.

Document sources side

- Estimation of the implementation of each of 8 transfer channels "document source TO CELLAR":
  - 48 weeks including definitions, negotiations with the CELLAR team, necessary development and testing where 50% of the net time is supposed, that means 24 weeks * 2 team members (analysis + development) = 48 weeks = 240 man-days for all activities;
  - 8 weeks of time backup should be added, that means 40 man-days
  - Summary of manpower needed for one implementation: 56 weeks = 280 man-days
- The implementation will be planned and organized by the CELLAR team side

- Summary of manpower necessary: 8 document sources * 280 man-days = 2 240 man-days

CELLAR side

- Three full time members are projected to be necessary for each transfer channel development team on the CELLAR side:
  - 1 project manager taking care of communication with the document sources' stakeholders
  - 1 developer
  - 1 tester
- Transfer channel development and execution projection on the CELLAR side:
  - It would be practical to plan the development of transfer channels in a sequential way rather than in parallel, with some overlap between the periods allocated for the development of specific channels. This would mean the development and implementation of a transfer channel from one document source to the CELLAR every 8 weeks.
  - It is 64 weeks in total (8 document sources and 8 weeks) + 20 weeks (30% reserve) for solving exceptional and non-standard situations or unexpected delays = 84 weeks, or 420 man-days of implementation.
  - Summary of necessary man-days: 3 team members * 420 man-days = 1 260 man-days

Integrated access solution development team

- Integrated access solution development timing projection (incl. Obtaining the data from 2 document sources, that means development of 2 dedicated PULL transfer channels):
  - Integrated access solution project definition, negotiations, tuning: 24 weeks
  - Integrated access solution development: 120 weeks
    - with 8 weeks overlap to the previous project definition stage because the elementary development of the building blocks concerning the content harmonisation tools could be performed even without having project definition stage finished
  - Integrated access solution testing and tuning: 120 weeks
    - for both web site and mobile applications
  - Time backup for non-standard situations and risks processing: 20 weeks
  - Summary of timing: 156 weeks (~ 39 months)
- Integrated access solution development team work effort projection:
  - 10 team members: 156 weeks each, 1 560 weeks all
    - 1 project manager
    - 1 database developer
    - 1 user interface designer
    - 1 context & content manager
    - 6 developers
  - 2 team members (testers): 140 weeks each, 280 weeks both
  - Summary of timing: 1 840 weeks = 9 200 man-days

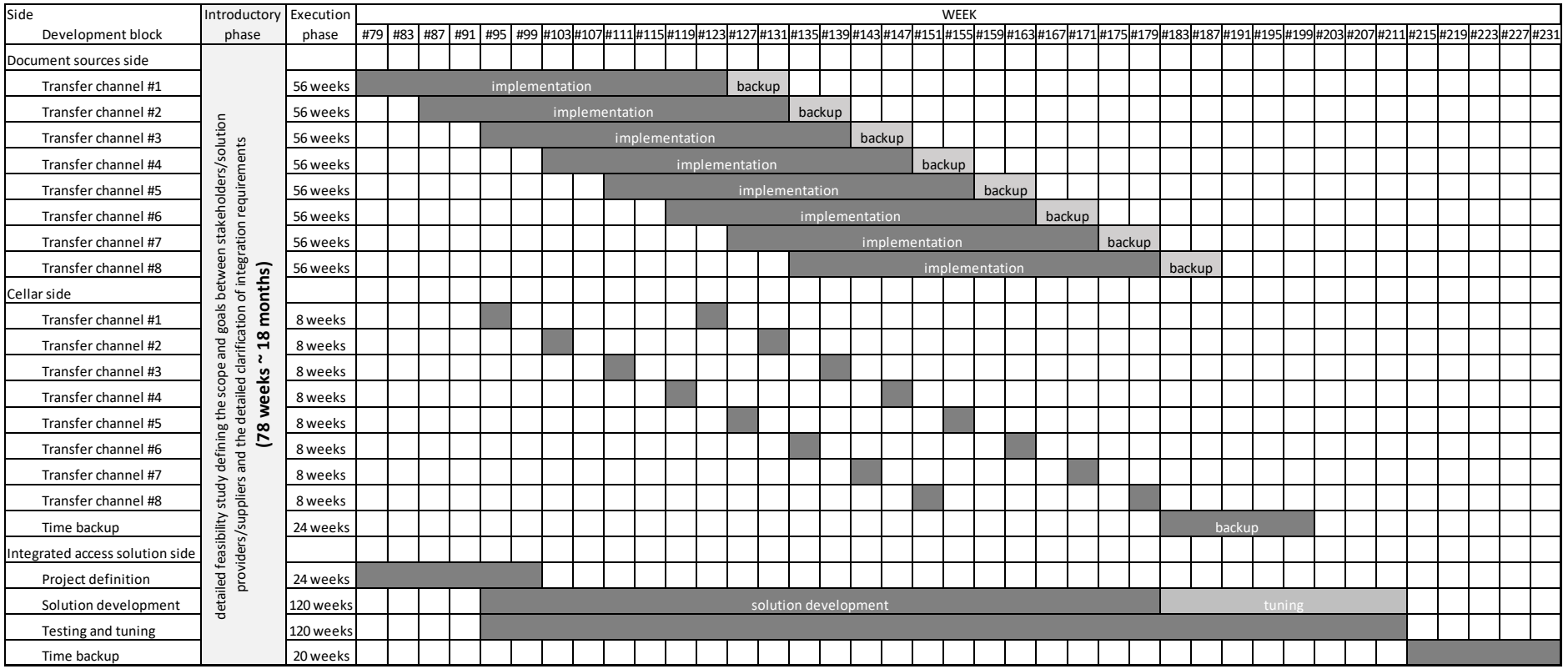


Figure 62: Integrated access solution deployment schedule projection – alternative No. 4

6.4.4.5. Estimations of the overall cost

The estimations of the overall costs are based on the estimations of the implementation time multiplied by the estimation of the man-day cost.

Rough overall estimations of integrated access solution Alternative No. 4 implementation costs:

- Document sources side:
  - Manpower summary estimations:
    - Summary of man-days: 2 240 man-days
    - Estimation of the costs of one man-day: 600 €
    - Estimation of summary costs: **1 344 000 €**
- CELLAR side
  - Manpower summary estimations
    - Summary of man-days: 1 260
    - Estimation of the costs of one man-day: 600 €
    - Estimation of summary man-power costs: **756 000 €**
- Integrated access solution side:
  - Manpower summary estimations:
    - Summary of necessary man-days: 9 200
    - Estimation of the costs of one man-day: 600 €
    - Estimation of summary man-power costs: **5 520 000 €**
- PA Infrastructure estimation:
  - Estimations are doubled with respect to alternative No.1 because of the much larger storage and server capacity needed (storage of many millions of converted documents is expected)
  - Hardware: 400 000 €
  - Software licences (databases): 400 000 €
  - Deployment (10%): 80 000 €
  - Estimation of PA infrastructure summary costs: **880 000** €
- **Overall cost estimations: 8 500 000 €**

Operational aspects

An integrated access solution sustainability team will be required, composed of the following roles:

- Project manager
- Database & ontology administrator
- Content harmonisation administrator
- Web application administrator
- User feedback operator

6.4.4.6. Conclusion (SWOT diagram)

Conclusions on integrated access solution alternative No. 4 – Content harmonisation on a central level is shown in Figure 63 in the form of a SWOT diagram.

Figure 63: Alternative No. 4 SWOT diagram - Content harmonisation on a central level

6.5. Final conclusions

The alternative solutions described in chapter 6.4 should be considered as high-level examples of the possible solutions. It should be clear from the descriptions that each alternative uses common components. Of course, it would be possible to combine the various elements of each alternative.

It is again necessary to refer to the chapter Mind map of the integrated access solution vision (see 6.3.6). The mind map contains many building blocks that could be added to or removed from the various modified integrated access solution alternatives. Also, these blocks may represent possible options for an enhancement of the integrated access solution in the future.

In general, it is possible to assert that the main differences between the alternatives include:

- The choice of the way of obtaining the content and context further provided to the end-users;
- The method for re-processing the context into the ontology, which will then serve as a basis for the web application features and which is contained in all alternatives in the front-end part;
- Whether the content in the file attachments will be re-processed into more flexible formats or not.

The resulting web application, i.e. the user interface of the integrated access solution, is always the presentation layer of the obtained data (content and context). Therefore, it could be deployed in a similar or slightly modified form to any alternative.

The description of the experiment (see 5.2) also clarifies the need for a common and unique methodology for processing documents (content and context) that would be provided within the integrated access solution. It is clear from the description of each alternative that the decision on how this methodology would be applied is in the responsibility of the document sources stakeholders. They would decide which data and in which way data from their document sources would be published for further use in the integrated access solution.

6.5.1. Solution alternatives comparison

This chapter summarizes mainly the timing and cost aspects of the alternatives described in chapter 6.4 in relation to the expected user experience (overall benefits for the end-users) and the estimated risks during the development and deployment of such an alternative.

		Rough estimations				
#	Alternative	Overall costs (PA web app / data / infrastructure)	Timing	User experience	Risk level	Potential to growth
1	**Web application built upon CELLAR** (see 6.4.1)	6 452 000 € (3 396 000 € / 2 616 000 € / 440 000 €)	48 months	★★★	★★★	★★★★★
2	**Decentralised aggregated search** (see 6.4.2)	5 794 000 € (1 590 000 € / 3 744 000 € / 460 000 €)	29 months	★★★	★★★	★★
3	**Centralised aggregated search** (see 6.4.3)	3 948 000 € (1 488 000 € / 1 920 000 € / 540 000 €)	27 months	★★	★★	★★
4	**Content harmonisation on a central level** (see 6.4.4)	8 500 000 € (5 520 000 € / 2 100 000 € / 880 000 €)	57 months	★★★★★	★★★★★	★★★★★

Table 30: Comparison of Timing / Costs / User experience / Risk level / Potential to growth of the alternatives No. 1 - 4

Table legend

All entries in the table must be considered as rough estimations based on the experience of the authors of the study with similar projects from the past.

The column *Overall cost* (PA web application /data /infrastructure)

- The upper row includes the total cost of the given alternative;
- The lower row includes the calculation of the total price divided into three parts:
  - The cost of the PA web application development
  - The cost of the data – necessary development and deployment at the document sources' side (incl. developments at the CELLAR side where relevant)
  - The cost of the necessary hardware + software infrastructure
- A brief clarification of the cost estimation is always included in the subchapter Estimations of the overall cost of the relevant alternative (the reference to subchapter is specified in parentheses in the column Alternative).

The column *Timing*

- Includes an estimate of the duration of the development, deployment and testing of the given alternative;
- A brief clarification of the time estimate is always included in the subchapter Estimations of the implementation time of the relevant alternative (reference to the subchapter is specified in parentheses in the column Alternative).

The column *User experience*

- The general evaluation of Strengths and Weaknesses part of a SWOT diagram, where
  - one ✶ is a minimum
  - five ✶✶✶✶✶ is a maximum

The column *Risk level*

- The general evaluation of Threats and Opportunities part of a SWOT diagram, where:
  - one ✶ is a very low risk
  - five ✶✶✶✶✶ is a very high risk

The column *Potential to growth*

- The estimation of the future potential of the given alternative, meaning how many benefits for the end-users could be added in the future after implementation of the given integrated access solution:
  - one ✶ is a very low potential to growth
  - five ✶✶✶✶✶ is a very high potential to growth

Notes on alternatives comparison

The following conclusions can be made based on the existing information:

- Costs and implementation time in general:
  - It is difficult to determine costs and implementation time accurately without direct and detailed consultations with document source stakeholders, who would provide more technical information about the technical modifications which must be implemented on the document sources' side in order to be able to integrate with the integrated access solution.
  - For more detailed cost and timing estimations a dedicated feasibility study should be carried out.
- Overall costs
  - There are no significant differences in the overall costs in the estimations.
  - The lowest cost estimation, for alternative No. 3, is approximately 46% of the highest cost estimation, that of alternative No. 4.

- Overall timings
  - Differences in implementation timings of the alternatives are almost identical.
  - The lowest timing estimation, for alternative No. 3, is only approximately 47% of the highest timing estimate, that of alternative No. 4.
- User experience
  - The estimated user experience is identical for the alternatives Nos. 1 and 3. The reason is that the document content remains in the form of file attachments in these alternatives; further differentiation of the user experience would depend on the design and implementation approach taken by the selected supplier of the integrated access solution.
  - Alternative No. 2 has the lowest estimated user experience rating. The reason is in many external dependencies – sequential processing of the results provided generated by distributed search engines.
  - The estimated user experience is the highest for alternative No. 4 because it offers maximal benefits to the end-users.
- Risk level
  - The estimated risk level is identical for alternative Nos. 1 and 2 however the reasons are different. The risk level for alternative No. 1 is because of possible difficulties in IMMC implementations while the risk level for the alternative No. 2 is because of possible complications in implementations on the distributed search engines which can reasonably be foreseen.
  - Alternative No. 3 has the lowest risk level. This is because indexing the document source by an external search engine is a proven method, used in similar projects.
  - Alternative No. 4 has the highest estimated risk level. This is because of the complicated development of the re-arrangement of the content stored in file attachments into more flexible structured formats.
- Potential for growth
  - Potential for growth is at its highest possible level at alternatives Nos. 1 and 4. That is because alternative No. 1 can first grow into alternative No. 4 which can then grow even further.
  - On the contrary the potential to growth is at its lowest level in alternatives Nos. 2 and 3. The reason lies in the very limited possibilities for future improvements because of the high number of external dependencies.

6.5.2. Final recommendation

Based on the evaluations above - implementation time, satisfactory user experience, acceptable risk level and excellent potential to growth - alternative No. 1 - Web application built upon CELLAR – is recommended.

Alternatives No. 2 and 3 also appear interesting, in particular due to their short implementation time, and for this reason it would be useful to test them by means of a proof of concept exercise in order to see whether they are capable of bringing the expected benefits.

Alternative No. 4, despite its progressiveness and maximal user benefits and usability, is not recommended at this stage due to its high risks.

When making a decision on the choice of the solution alternative, it is important to remember that the integrated access solution is a project which aims not to address the short-term needs of the users, but one that should bring long-term benefits and a completely new way of working with the documents, and a new, far higher level of quality in comparison with the existing diversified document ecosystem (*.europa.eu)

PUBLICACCESS.EU STUDY

ANNEXES

Annex 1: Study database

This Annex describes the study database created to simplify both the gathering and presentation of the information needed for the analytical section of the study.

In the first part of this Annex the basic principles of the study database ontology are described.

In the second part, the study database environment and user interface are described.

1. Description of the study database ontology

The primary aim of the analysis of the document sources was to describe them in a structured, verifiable and unified way in order to design the alternatives for the overall architecture for the future integrated access solution.

An ontological approach was chosen and the study database ontology follows the ISO/IEC 13250:2003 standard (Topic Maps.[245])

The principles of the ontology of the study database are shown in Figure 64 with the general description included and described onwards.

In fact, the ontology of the study database consists of three basic parts:

i. **Study investigations** containing document sources, vocabularies, their entries and attributes,[246]
ii. **Sample documents** containing information used to model the structure of the unified document later presented in some of the alternatives of the integrated access solution,[247]
iii. **EuroVoc** reworked into the Topic Maps ontology used in both of the aforementioned ontology parts.[248]

Information on how to log in to the study database can be found in the section 2.2 of this Annex.

[245] Basic information about the Topic Maps information modelling approach: https://en.wikipedia.org/wiki/Topic_Maps.
[246] Part of the study database ontology used in analyses of document sources http://atom.ts-publicaccess.eu/form/space?SpaceId=100031.
[247] Part of the study database ontology used for modelling the unified document structure: http://atom.ts-publicaccess.eu/form/space?SpaceId=100407.
[248] EuroVoc reworked into the Topic Maps ontology: http://atom.ts-publicaccess.eu/form/space?SpaceId=100082.

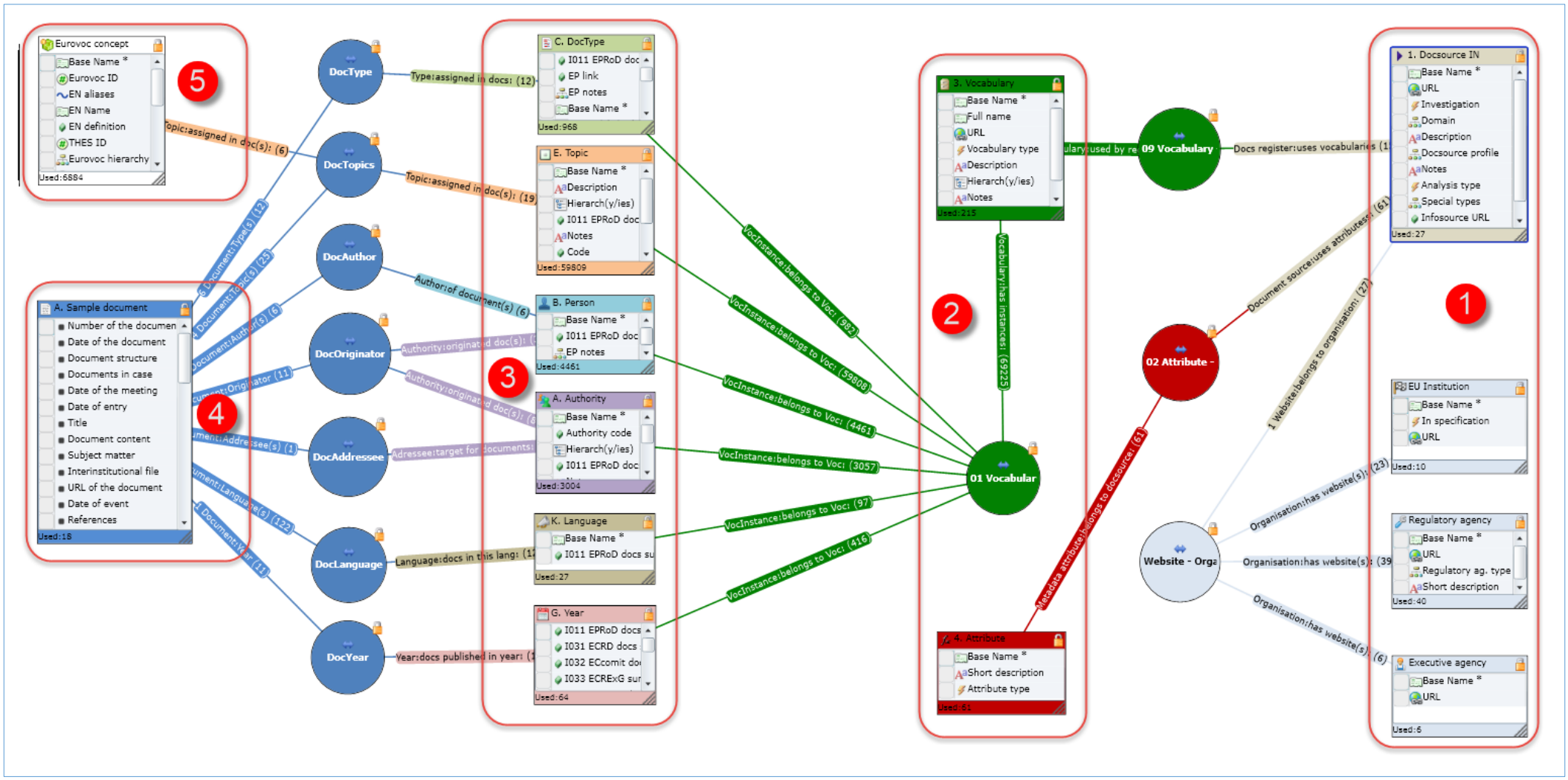


Figure 64: Principles of the ontology of the study database

Basic legend:

Rectangles in Figure 64 represent Classes which capture the basic categorization of the problem domain.
Circles in Figure 64 represent Associations linking the specific Classes with the named and directed relationships.
Class attributes are represented by rows in the rectangles of Classes, where each row represents one attribute

Description of the ontology (according to the numbered legend in Figure 64)

1	Document source A document source is a web application operated by a specific subject – either an EU institution or an agency which provides access to documents to its users in a specific way. Institutions or agencies call document sources differently, e.g. document registers. *EU institutions operate one or more of the document sources.*
2	The document source uses metadata for the description of its documents. There are two major types of metadata recognized: vocabularies and attributes. Vocabulary A vocabulary is a file/list of metadata entries created for the specific purpose, organised and named, and existing independently on the document. Its role is to capture the metadata, i.e. specific document properties in the form of a relationship between a document and its entry/entries in it. It should be possible to describe the individual vocabulary entries. *Attribute* The attribute is a meta-information existing only in the view of the document, for example in the form of a character string or numbers or a date. The attribute is typically displayed as the value of a variable in the user interface.
3	Individual *Document sources* use different *Vocabularies* (uncontrolled/controlled, taxonomy, thesaurus). *Vocabularies* contain instances, which are referenced by the documents published in the document sources. They are divided into logical blocks, which are supposed to be part of the final future integrated access solution. Following basic types of vocabularies were used in the analysis: • *DocType* – represents *Document Type* ○ the types of documents are grouped into the logical units in some of the document sources, we track it separately because of its applicability in the design of the future solution • *Topic* – represents the document theme • *Person* – represents Author, Rapporteur or another physical person • *Authority* – represents the document author who is not a physical person but the organisation structure of the institution • *Language* – specifies the language versions, in which the document is available • *Year* – is the shorter expression of the document creation date
4	A Document is the basic entity in each of the document sources. It is described by a set of metadata, i.e. relationships to vocabulary entries or attribute values (while maintaining the relationship property ←→ document register).
5	EuroVoc is a Thesaurus type vocabulary. In the study database it is handled separately because of its potential for the future integrated access solution.

2. Description of the study database environment

The study database is equipped with an intranet style web application for the execution of the analysis in accordance with the ontology outlined in the previous part of this Annex and for the presentation of the analysis' results.

The following chapters contain basic information on the usage of the study database.

2.1. Used terms

Ontology

Ontology is an organised scheme of data collections.
Ontology consists of classes, their attributes and associations.

Class, Class instance

The Class is a basic ontology component, a container for storing data.
In the study database the Class represents a form for data input.
After filling out this form and saving it to the study database, a class instance is created.

Class attribute, Attribute instance

The Class Attribute is a property of the class which is set in the Class instance form. It can have various types like strings, names, selections, taxonomy or images and files attachments.
If the entry is set and has a value, it is an Attribute instance.
Attributes can be labeled as Class Features.

Association, Association instance

Association is a relationship between Classes. The association is represented by the entry in the form of a Class.
Association instance is a relationship between the individual Class instances.

Base Name*

The 'Base Name*' Attribute is the main Attribute of Class instance used for Class identification and its value should conclusively define just one Class instance.

Click/short click, double click

In the following text, these terms represent the names for ways of working with the computer mouse. After placing the cursor on a particular spot on the screen, a mouse button is pressed shortly. This is the click/short click. This can cause an event, such as highlighting an object under the cursor, executing a certain action, as far as it is programmed for the object. Double click is a name for two quick clicks following one another.

Drop-down menu

This is a component that allows the user to choose one item (sometimes more items as well) from a list of predefined options and display it in the text box. Expanding and collapsing a menu can be performed mostly by clicking a button on the right hand side of the box. Selection from the list replaces the text in the box.

2.2. Entering the study database

Study database only requires an internet browser for its operation. After entering the URL http://atom.ts-publicaccess.eu/ into the address bar in the browser and sending the request, the main screen is displayed.

Access into the study database is protected against unauthorized access by a System login.

In the left control panel navigation a 'display name' of the user [1] can be seen along with a figurine icon, underneath is the 'Login' link [2]. By clicking on one of these links, the system redirects the user to the login screen.

The page contains a login form where identification is verified through Username and Password.

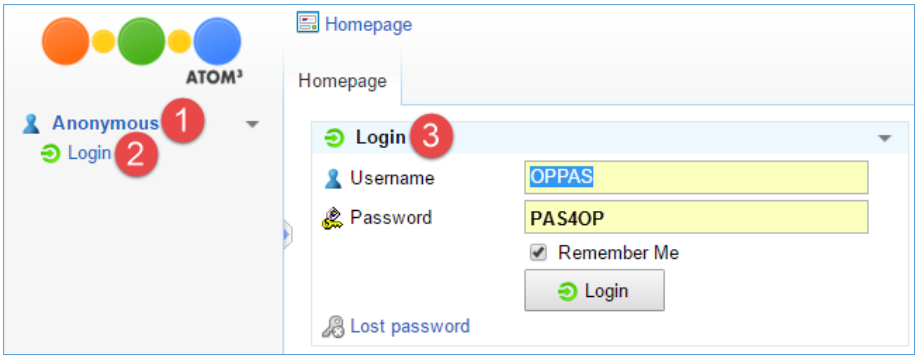


Figure 65: Study database - Login

1. *displayed username*
2. *login access*
3. *login panel*

If the 'Login' panel [3] is displayed on the main screen, login can be performed through this panel.

There is a default read-only access for the study database:

Login: OPPAS
Password: PAS4OP

Read-write access will be granted after a request is sent to info@aion.cz and the user is authorized.

Logout from the study database:

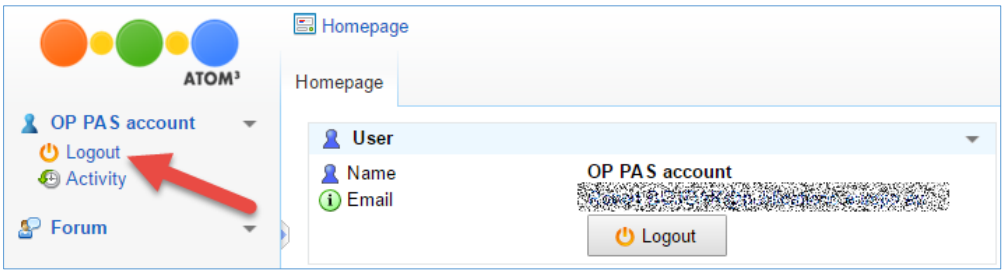


Figure 66: Study database - Logout

2.3. Study database editing environment

2.3.1. User interface interface

The study database environment is composed of several basic parts.

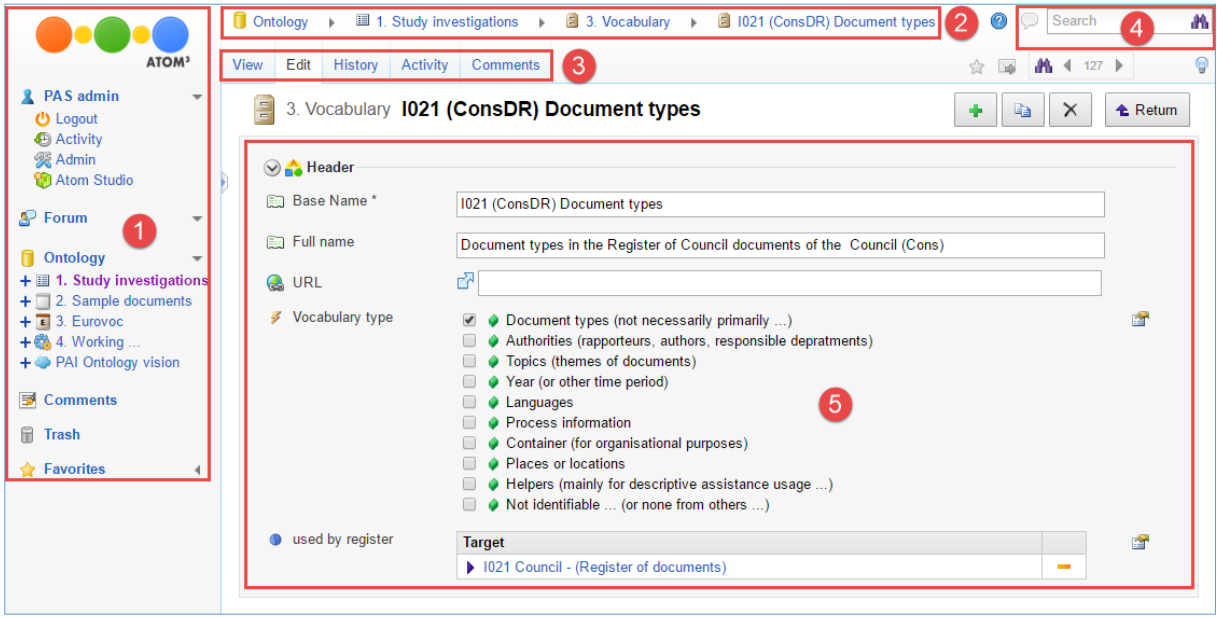


Figure 67: Study database - User interface basics

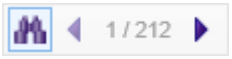
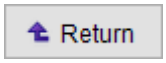
1. *Navigation panel - left navigation panel with links for quick switching between pages. The width of the panel can be changed in a few steps.*
2. *Breadcrumb bar - display of links according to the web pages' hierarchy, or else according to the ontology.*
3. *Tabs - They enable easy switching between related pages, related content categorization, or switching between different data views or available functions.*
4. *Set of tools and indicator icons. Searching etc. can be performed here.*
5. *Page content - depends on the user's particular activity and displays variable content according to the particular use case.*

2.3.2. Set of tools and indicators

The area with a set of tools and indicators is situated at the top on the right and contains a set of icons which represents various utility functions and search feature. The composition of this set can differ while browsing different parts of the application.

Tools and indicators overview:

ON OFF	Tooltip **This switches the floating panel mode ON/OFF. Floating panel displays selected info about Class instance.** **ON and OFF state is indicated by colour highlight.**
Search	**Search** Searching directly from the current page. User enters the desired piece of text into the 'Search' field and by clicking on the icon on the right, a page with full text search is displayed showing the search results.
	Favourites

ON OFF	This function regulates whether the currently displayed page should be marked as a favourite or not. The state of the page's inclusion to favourites is indicated by a change in icon colour.
 **ON OFF**	**Subscriptions** With this feature user can select which pages are subscribed, meaning that when their content is edited, users receive an e-mail notification. The state of the page's inclusion to favourites is indicated by a change in icon colour.
 **ON OFF**	**Search results pin** This activates tools for browsing search results. A control feature for browsing search results is added to the set of tools on the page.
	**Search results navigation** This enables returning to a page with the previous search results (settings and result), and enables browsing data instances of search results on the instance page.
	**Bulb** Overload indicator.
	**Return** Return button for returning to the previous web hierarchy level, or else back to the previous superior node (page). It is not just a return to the previous page. The user must get used to this when navigating through the application.

2.4. Study database general control methods

In various parts of the application similar components are used and their control methods are the same, or very similar. Below are described some frequently used components and their controls.

2.4.1. Chart controls

Charts are used in the application for displaying instance lists. The list is paged.

The number of displayed records is displayed on the right above the chart [1]. Next to it there is a field with expandable list of several varieties of number of records per page [2]. The number of records to be displayed is set here.

The user can switch between individual pages with navigation arrows [3]. A sequence of page numbers is displayed between the navigation arrows at the bottom of the list with the option to directly jump to a page [4].

Figure 68: Study database - Chart controls

1. *number of displayed records*
2. *number of records per page*
3. *navigation arrows for scrolling through pages.*
4. *link to particular page.*
5. *field for entering the number of desired page*
6. *number of pages*
7. *field for entering filtering text - masks, search is performed in 'Base Name*' instance*

Another navigation option is typing the page number into field [5] and confirming with 'Enter'. The overall number of pages is displayed next to it [6].

Records can be filtered in the chart through Mask [7], which then displays only records containing the desired text fragment in their Base Name*. Character size and diacritic are ignored when filtering. Compared to full text search, similar or derived words are not counted, filtering is done only according to 'Base Name*' attribute. Filter is set by typing text into the field [7] and confirming with 'Enter'.

2.4.2. Full text search settings

Full text search component is not only part of the 'Search in ontology' page, but is also used on many pages containing a list (mostly a chart), or as an assistance component when selecting from table lists.

The component contains a search method switch. User can choose from:

[1] – a search which is limited to coded and named items through the CONTAINS method, which searches only the text where words start with expression typed into the filtering field.

[2] – a search of all items by the FREETEXT method, which searches the occurrence of text that either contains the expression typed in the filtering field itself or other forms of the expression.

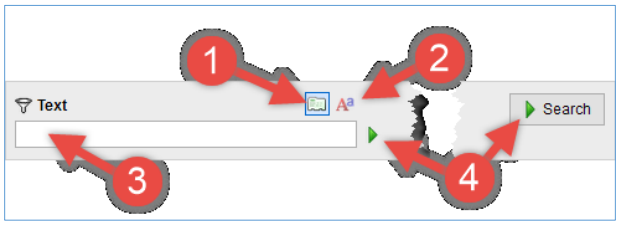


Figure 69: Study database - Full text search component

1. *names and codes search settings*
2. *all text types search settings*

3. *field for typing the search text*
4. search is performed by clicking on the triangle (button)

Searched text is typed into text field [3] and the search process itself is executed by clicking on the triangle icon [4], or else by clicking on the 'Search' button, which is displayed in some cases as well.

The search process applies the set filter on the chart data linked with the search component.

If the search field is empty (empty field [3]), the full text filter is cancelled and a list of all instances is displayed.

2.4.3. Tooltip

The Tooltip button is placed in the set of tools and indicators area at the top right corner of the browser.

ON OFF

This turns the display of the floating panel with selected Class instance information on and off. ON and OFF state is indicated by colour highlighting. When moving the cursor over the Class instance identifier (for example in the tabular list of Class instances with links to Class instances) a panel with Class instance data preset for this type of view shows up after a while when the cursor is placed over a link to Class instance (Base Name*). Attributes data are displayed only if some of their instances exist, or else if the items have some preset value.

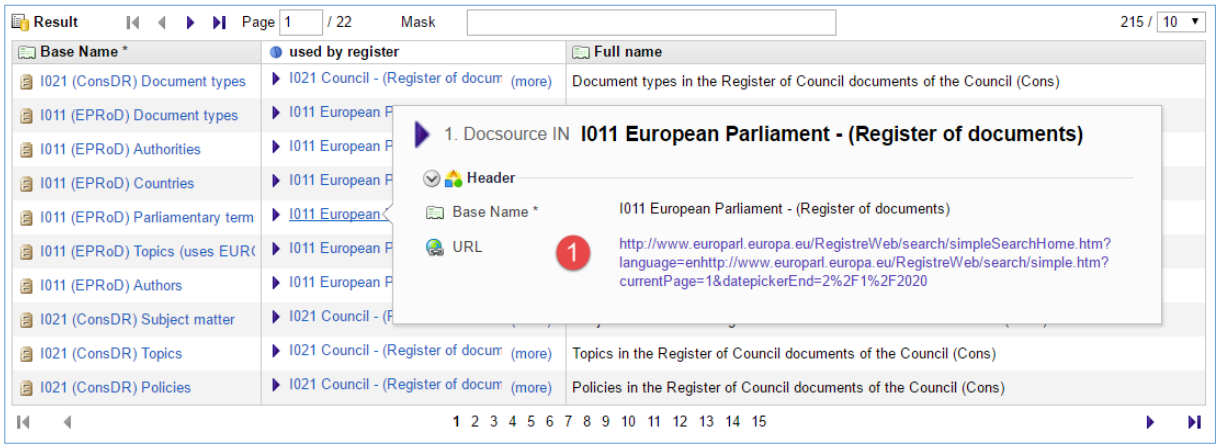


Figure 70: Study database - Tooltip - floating panel with Class instance data

1. *The tooltip panel is always displayed above all other content*

2.4.4. Search in ontology

Full text search page shows up after clicking on the Search button (telescope icon), which is situated in the set of tools and indicators area at the top right corner of the browser.

In the top left section there is a component for setting full text search [1]. Control is also described in the chapter about the full text search settings.

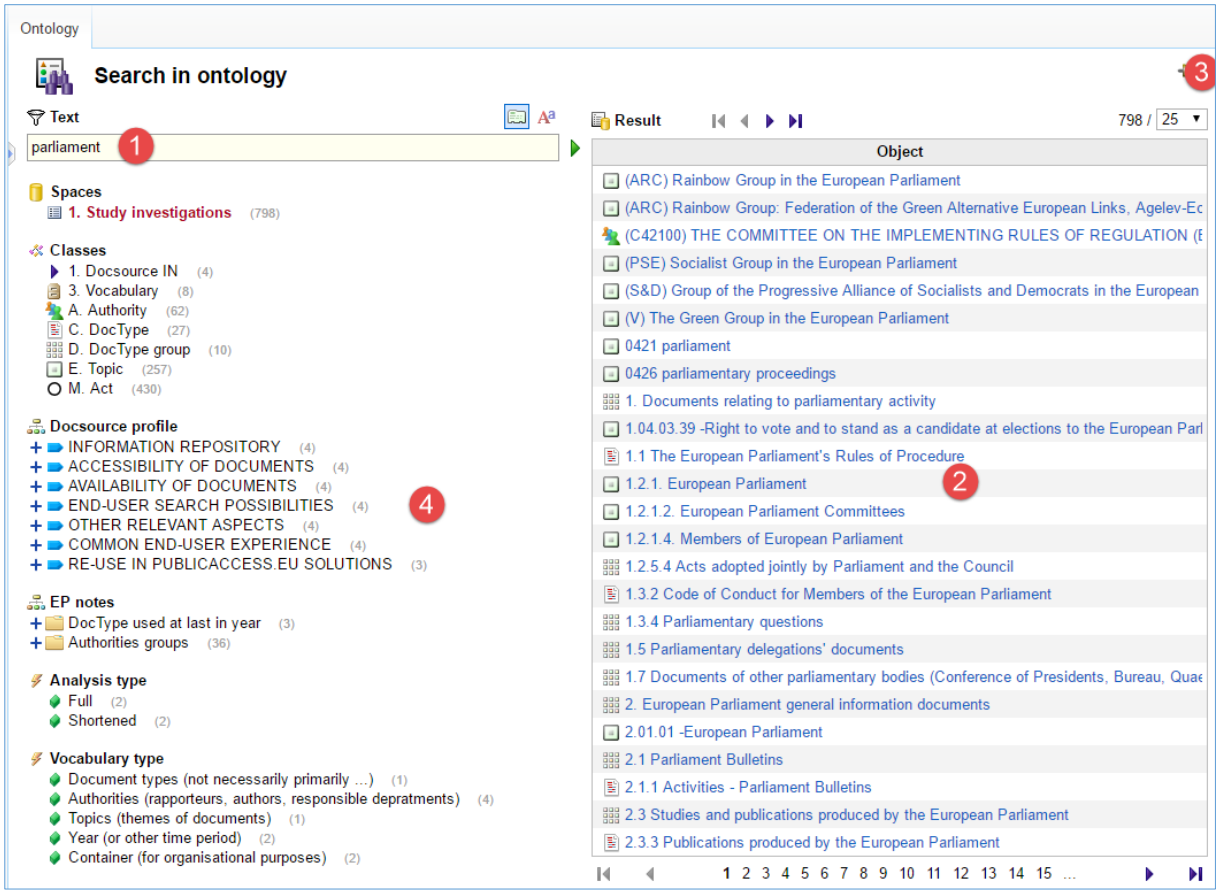


Figure 71: Study database - Search, List of results, Facet filters

1. *search settings component*
2. *list of search results*
3. *results pin - preservation of the list of searched results*
4. *list of results filters*

In the right hand section of the search page, a list of search results in the form of links shows up [2] after the search execution. The link redirects the user to the Class instance page. By clicking on the link, Class instance page is displayed.

The search result can be pinned to the tools by clicking on the pin icon [3].

In the left hand section there is a set of filters [4], which can be used to specify the search results. Number of results is displayed in the brackets next to each filter. The set of filters copy the application data ontology.

2.5. Class instance editor and its components

The page with the Class instance list [1] is displayed after clicking on the link with the Class name. This way the study database interface is capable of displaying records for Vocabularies, Document sources etc.

On the Class list page there is a chart with links to individual Class instances. Links are situated in the 'Base Name*' column. Other columns can also be visible showing selected Class attributes. The chart can be sorted according to these values by clicking on the column header.

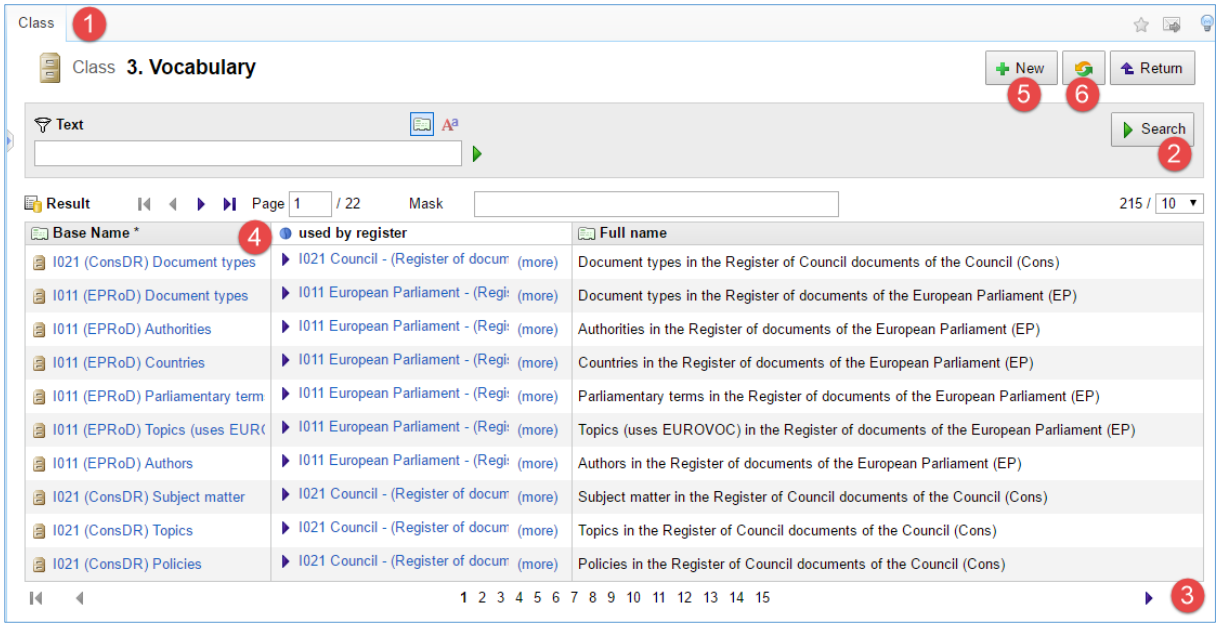


Figure 72: Study database - List of the class instances

1. *Class list page tab*
2. *full text search component*
3. *chart with Class instance list*
4. *sorting method switch*
5. *button for creating new instance*
6. *list display settings into predefined form*

A full text search component is implemented on page [2]. The control method is described in the chapter about full text search settings.

The list of Class instances is displayed in chart [3]. The control method is described in the chapter about chart controls.

By clicking on the chart column header, sorting according to the values can be either set or removed, or the sorting mode can be changed from ascending to descending or the other way, which is indicated by the triangle icon [4].

New Class instances can be added by the 'New' button [5]. Text for instance identification is typed into the Base Name* field [7] and confirmed by tapping the 'Create' button [8]

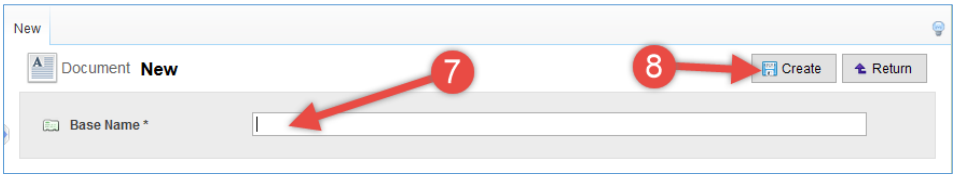


Figure 73: study database - Adding the new Class instance

7. *field for entering instance identifier (Base Name*)*
8. *button for creating instance*

Clicking 'Reload' [6] removes all user-defined settings and displays the chart in predefined mode.

By clicking on the link in the chart, the Class instance editor shows up and offers the option to edit relevant Attributes (items).

2.5.1. Class instance attributes editing

After clicking on the Class instance link a page for viewing or editing the Class instance appears [1].

Instance editor tabs:

i. **View** – intended for simple viewing of the Class instance (without the option to edit)
ii. **Edit** – intended for editing the Class instance
iii. **History** – providing the complete history of the Class instance
iv. **Activity** – overview information about who modified the displayed Class instance and when
v. **Comments** – intended for team collaboration

On the 'Edit' tab [1] (similar to View tab) there is a Base Name* [2], which is a text that identifies the Class instance. The page then contains a group of Class Attributes. Individual Attributes are sorted under one another. The left side of the Attribute [4] contains a name, the right side contains the value of the Attribute. Some Attributes can be edited, some are set automatically and their value cannot be changed.

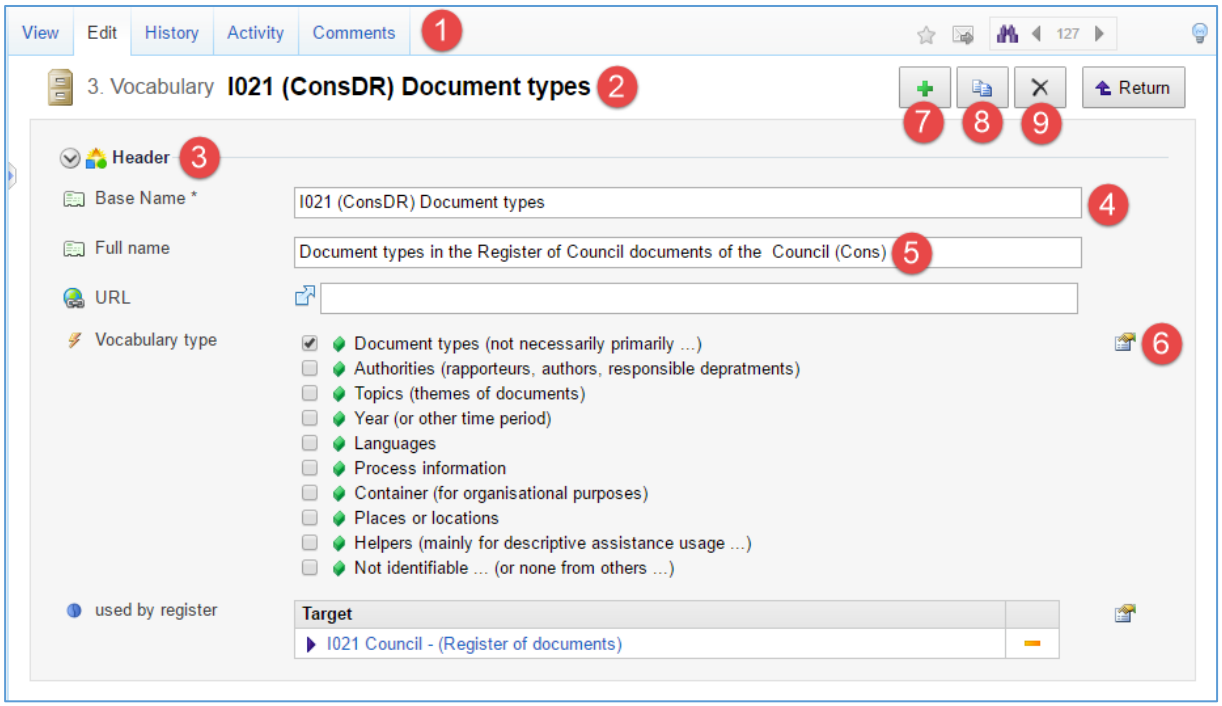


Figure 74: Study database - Class instance editor

1. *Class instance editing tab*
2. *Base Name* - identifies Class instances*
3. *folder header*
4. *Attribute (Feature, value)*
5. *Attribute instance editing field*
6. *link to the Attribute external editor page*
7. *button for creating new Class instance*
8. *button for creating a copy of currently displayed Class instance*
9. *button for deleting Class instance*

Attributes are organised into folders [3]. By clicking on the catch icon (legs of a triangle) the whole folder content is hidden or expanded. Icons enable streamlining the user interface and enable grouping attributes into categories.

Items (Attributes) can be edited directly in the editing field [5] or in Attribute editor for the particular item which is displayed after clicking on the edit icon [6]. Some types of Attributes support both editing methods.

With the 'Plus' button [7] a new Class instance can be created. The button with two pages icon [8] creates a new Class instance as a copy of the existing one (Clone).

Class instance is removed by clicking on a button with 'x' symbol [9].

Items without set value can contain an empty space.

If the Attributes can be edited directly by typing value into the field, or by other means of direct editing (for example by pasting with Ctrl+V), a pen symbol [1] is displayed at the right corner of the editing field, which is replaced by a floppy disc icon upon saving [2]. Saves are performed automatically when the editing field is closed. If the value cannot be saved into the database, a warning pops up [3]. By clicking on the 'x' symbol [1] inside the field, its content is deleted. Before saving the item to a database, the original content prior to editing the text can be restored by the 'Esc', or 'Backspace' key.

Multiple attributes contain additional 'Plus' and 'Minus' icons [4] at the right corner of the Attribute field [4], which are used for creating other item values, or deleting a certain value.

Figure 75: Study database - Attributes editing

1. *Attribute in editing mode - value changes*
2. *Attribute value saving*
3. *Attribute's value cannot be saved*
4. *Icons for adding value to multiple Attribute*
5. *Icon with link to relevant Attribute editor*

'Plus' and 'Minus' add or remove values of 'multiple items, such as file list. By clicking on 'Minus' the field with the particular value is removed, after clicking on 'Plus', a new field for a value is added.

Item value can also be filled by copying to the clipboard (Ctrl+C, Ctrl+V), or by moving the value with the mouse directly in the application. By setting the cursor on a highlighted part of the item value text and holding left mouse button, the value is sort of separated and it is possible to move it somewhere else, while holding the left mouse button during the whole process. The appearance of the cursor indicates which place is suitable for dropping the value and which is not. After placing the cursor over a suitable place and releasing the left mouse button, the moved text will be pasted onto the cursor's position. It will be removed from the original place. This is how value moving is performed. If the Ctrl key is pressed while releasing left mouse button, a copy of the text will be made and original text stays in place without being deleted.

At the right hand corner of the Attribute field an icon for link to Attribute editor page can be displayed [5]. After clicking on the icon, a page appears, enabling Attribute value edits depending on particular Attribute type. For example, a page with the Class instances selection for the setting 'Hierarchy' appears.

When placing the mouse cursor above the icon, the Attribute can be highlighted if there is no optical connection between the name and Editor icon.

2.5.2. Associations

Association instances can be edited in Association Editor environment (appears after clicking on icon [6]), but deleting Association instance can be done directly from the Class instance editing page as well by clicking on 'Minus' icon [7] and confirming the verification request.

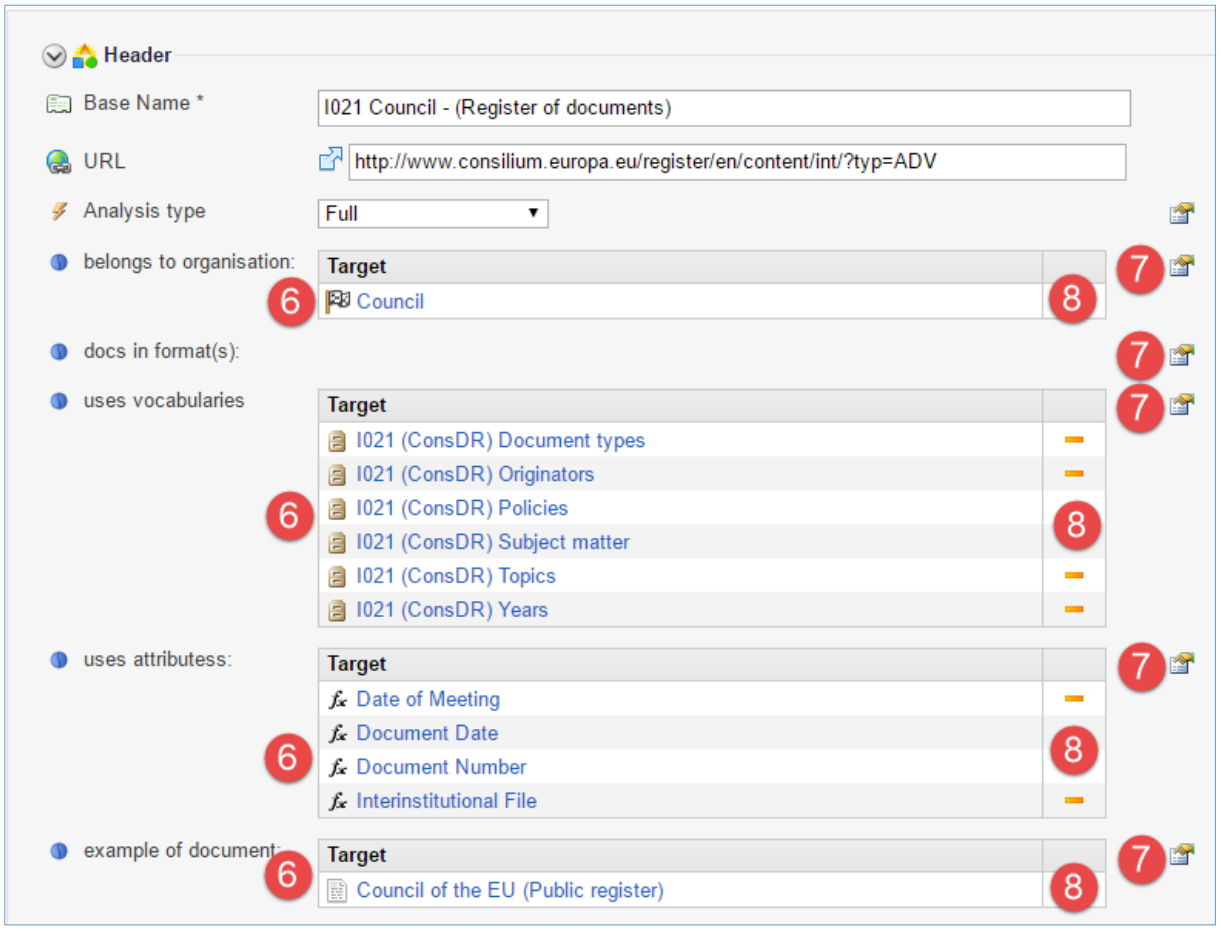


Figure 76: Study database - Associations overview

1. *List of linked instances of the target Class*
2. *Class instance attribute editor (Association instance)*
3. *Icon for removing Association instance*

2.5.3. Image Attribute

Editing the value of an 'Image' type item consists of assigning or inserting existing image to the Attribute. This is possible by clicking on the 'Plus' icon [1], which opens a dialog window with the option to choose a file from a disc representing the image. The file with the image can be moved from the file manager interface to the container [2] using a mouse. Another option is moving the displayed image from the web browser interface into the container [2] using a mouse.

Name and other features of the image are filled automatically. Name [3] and note can be edited. By clicking the 'Download' button [4], which is displayed in view mode, the image can be downloaded to the computer disc, or viewed.

For viewing the picture in a web browser interface, click on the displayed illustration [5]. A box with an image appears. In the box there is an image gallery of a relevant Attribute that can be browsed.

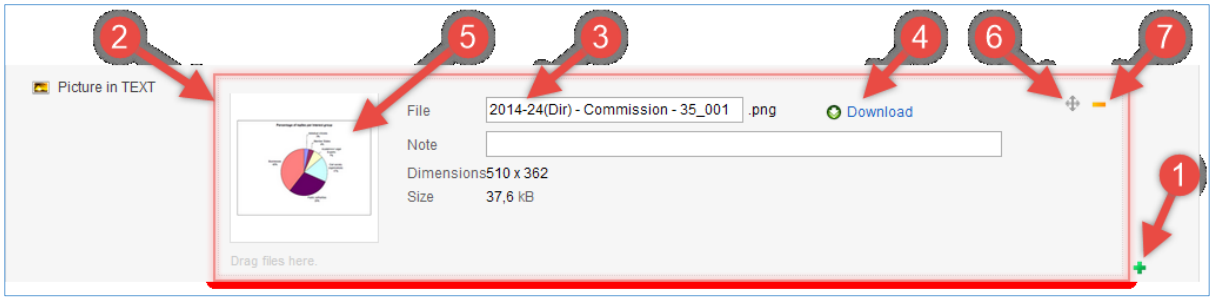


Figure 77: Study database - Image attribute editing

1. *launching the disc explorer, selecting and inserting image*
2. *image container*
3. *image file name*
4. *downloading the image - export to disc*
5. *illustration of saved image in unified size*
6. *icon for moving the image and change of order in the list*
7. *removing image from the list of images*

Using the icon of a four-legged arrow [6], the image position in the Attribute container [2] can be adjusted. By clicking on 'Minus' icon [7] linked image is removed.

2.5.4. File Attribute

Editing the value of a 'File' type item consists of assigning or inserting an existing file to the Attribute. This is possible by clicking on the 'Plus' icon [1], which opens a dialog window with the option to choose a file from the disc. The file can be moved from the file manager interface to the container [2] using mouse.

Name and other features of the file are filled automatically. Name [3] and note can be edited. By clicking 'Download' button [4], which is displayed in a view mode, the file can be downloaded to the computer disc, or viewed.

For viewing a file in web browser interface, click on the file icon [5]. A new browser window opens. Using this method, files for which the implemented display support exists can be viewed. Not all files can be viewed like this.

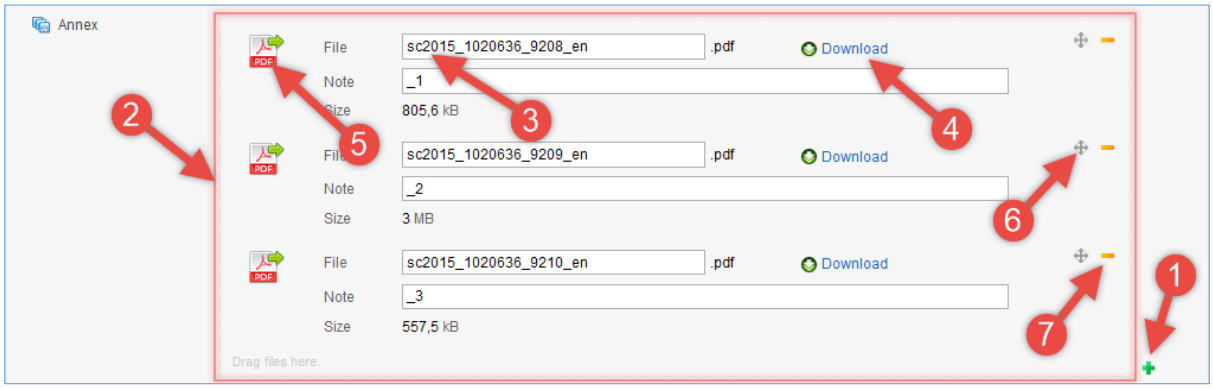


Figure 78: Study database - File attribute editing

1. *launching the disc explorer, selecting and inserting file*
2. *file container*
3. *file name on the disc*
4. *downloading the file - export to disc*
5. *illustration of saved file in file viewer.*

6. *icon for moving file and changing order in the list.*
7. *removing file from Attribute files list*

Using the icon of the four-legged arrow [6], the file position in the Attribute container [2] can be adjusted. By clicking on 'Minus' icon [7], the linked file is removed.

2.5.5. Text Attribute

'Text' type item is edited in integrated text editor. After clicking on the editing icon, the editing page opens with the option to insert text, multimedia, external links etc. with possible formatting and adjusting the look based on HTML.

For text formatting the editor features a set of tools [1] with a tooltip. The field for inserting text [2] displays the formatted text, images, etc. Using the clipboard (Ctrl+C), HTML content can be inserted, maintaining its formatting and overall look while automatically being integrated.

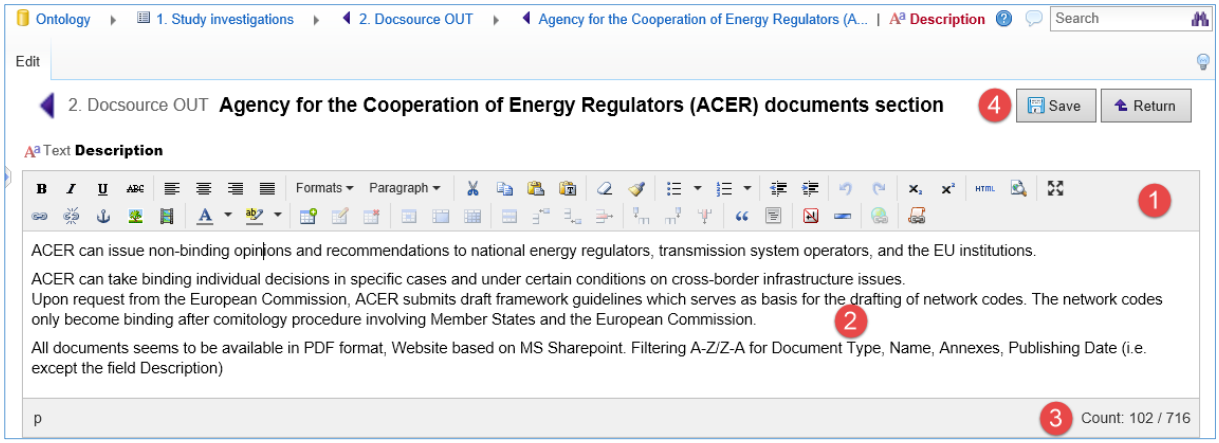


Figure 79: Study database - text editing

1. *Text editor tool set*
2. *Field for inserting text*
3. *Number of words/characters*
4. *Button for saving performed changes*

At the bottom right corner, a number of words/characters are shown [3].

By clicking on the 'Save' button [4], content is saved and the editor closes. By clicking on 'Return', the changes are not saved.

Text editor offers a set of tools for formatting and other adjusting of edited text, inserting links, charts, images and similar. By clicking on a tool icon, formatting applies, an object is inserted or a dialog window for specifying the inserted object appears, or another action or adjustment is performed.

Text editor offers these sets of tools:

Figure 80: Study database - Text editing tools

1. *character formatting: bold, italics, underlined, strikethrough*

2. *text alignment: left, middle, right, into block*
3. *block formatting: chart or image settings, predefined paragraph*
4. *copying via clipboard: remove, copy, paste, switching between different types of inserted text formatting*
5. *formatting removal, formatting check*
6. *block formatting: numbered list*
7. *block indentation: more left, more right*
8. *return to adjustments: previous, following*
9. *setting the text as index: lower, upper*
10. *HTML code editor display (F4)*
11. *switch between full screen editor view and tab view (F10)*
12. *inserting and removing hypertext link (Ctrl+K), removing link, inserting tab, inserting image file (Ctrl+M) inserting video file*
13. *font formatting: font color, font background color*
14. *chart inserting, adjusting and editing chart, cell, column, row.*
15. *highlighting paragraph into block, inserting framed 'code' block*
16. *inserting nonbreaking space, horizontal line, link to object, note into text*
17. *final look preview (F3)*

2.5.6. Group tree and Selection Attributes

On the Class instance editor page, Selection and Group type items can be edited as well.

Editors, which are used for editing Group or Selection type items, are displayed by clicking on the external editor icon.

Editor contains 'Edit' tab [1], where values are selected from displayed values, and a 'Design' tab, where the values can be edited or changed, new elements can be added or removed, and elements order and tree structure can be changed.

2.5.6.1. Selection

The Selection type attribute can be set directly on the Class instance editor page by clicking on the drop-down Attribute editing field along with a subsequent click on the required item from the list.

On the Selection editor page, which is only used for selecting one element from a list, a list of the elements is displayed with a ring [2] in front of the element. By clicking on the ring, the element is selected (or its selection changes). A right mouse click on selected element [3] cancels the selection and it is possible to delete existing Attribute value settings.

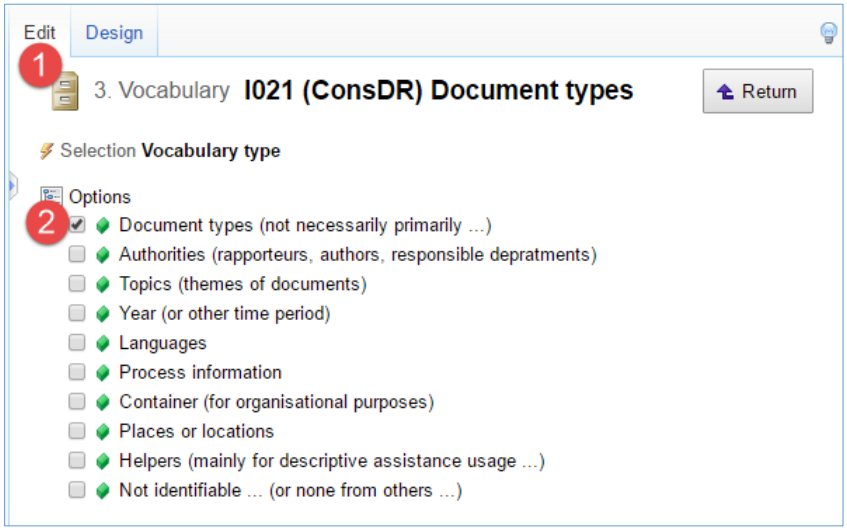


Figure 81: Study database - Choosing selection type elements

1. *page tab for element selection*
2. *checkbox type element is selected by a click*

2.5.6.2. Group tree (embedded taxonomy)

On the Group tree editor page, a list of elements is displayed with a square [2] in front of the element's name. By clicking on the square, the element is selected (included into set). By clicking on a selected element [3] the selection is cancelled and the existing Attribute value settings can be removed. The element hierarchy view can be gradually expanded or collapsed. By clicking on minus [4], sub-elements are displayed, by clicking on plus [5], sub-links are expanded.

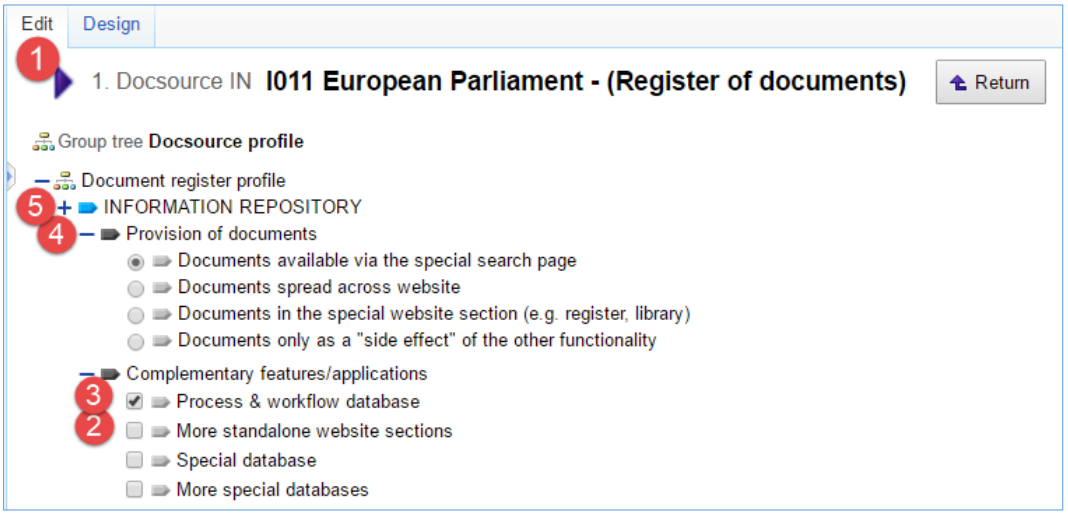


Figure 82: Study database - Group tree elements type selection

1. *element selection page tab*
2. *by clicking, check type element is selected (one or more)*
3. *by clicking on selected element, its selection is canceled*
4. *by clicking on minus, the tree under an element collapses.*
5. *by clicking on plus, the tree under an element expands.*

2.5.7. Association - link between Class instances

Association is a relationship between Class instances, where each side of the relationship plays different role. When editing Class instance, a relevant role of the Association is preset.

The association instances editor page is displayed by clicking on the edit icon in the Class instance editor, and this Class is set as one part of the Association, i.e. one role. This role can be firmly set in the Association editor and is represented by a link to relevant Class instance [1].

The other part of the relation to another Class, or else the second role of an Association, is set in the Associations editor. Based on a particular Association, a list of instances of a relevant Class (Classes) is displayed in a chart [2]. The method of control is described in the chapter about chart controls.

The chart contains a column with a selective element [3] for Class instance selection, a column with the name of target Class instance (Target) and an 'Open in new window' icon [5], which opens new web browser window displaying Class instance content. By clicking on this link, a Class instance editor opens in the new browser window.

By clicking on the selective element [3] the association role is assigned and Association instance is effectively set. By clicking on the selective element again, Association instance is removed.

With this method, more associations for the default Class can be set/removed and more Association instances can be created in case of multiple associations. In this case, selective element is displayed as a checkbox and more Class instances can be selected.

For 1-1 type association, only one selection from the Class instances list is permitted. Selective element is displayed as a switch button (radio) and only one Association instance can be selected. For removing association instance, right mouse click is performed on the selective element.

By clicking on one of the multiple selection icons [4], all displayed instances can be selected and all instances, which are displayed according to used filters, can be unselected.

Figure 83: Study database - Editing of associations

1. *Link for edited Class instance*
2. *List of Class instances, that can be linked by a relation (association)*
3. *Selective element for association assignment.*
4. *Multiple selection settings of displayed instances*
5. *Link for Class instance content view in a new window.*
6. *Full text search settings.*
7. *Class selection for filtering the list of connected Class instances.*
8. *Type of use selection for filtering the list of connected Class instances.*
9. *Icon for creating new instance of a relevant Class.*

Full text search (filtering) can be performed in the list [6]. Settings are described in the chapter about full text search settings. The list of instances [2] can be further limited by Mask (described in chapter about chart controls), Class selection [7] or by criterion for use, or non-use in some associations [8]. All can be displayed, or only those already used for an association, or those which are free, unused.

By clicking on plus icon [9], a page for establishing new relevant Class instance appears, which is used for creating new Class instance and displaying Class instance editor page for editing the instance. For returning to Association editing, use browser return button (left arrow) along with page refresh (for example F5 key). Association role instance is automatically created for created Class instances.

2.5.8. Hierarchy

Hierarchy is an attribute determining class instance position in the predefined hierarchical structure. Hierarchy is always edited on the superior instance's side (parent) by attaching subordinate instances (children). Access to editing is situated on Class instance editor page.

Hierarchy link is similar to the association within Classes, into which the Hierarchy Attribute is included. Only the subordination role of other instances to currently edited instance is set. Hierarchy instances editor page is displayed by clicking on edit icon in Class instance editor page, and this Class is set as a Parent role [1] in the hierarchy, with selected records classified as Children roles [2]. Editor control elements layout controls itself is similar to Association editing. A chart with records selection contains extra Sort[n] column [3], which allows assigning a serial number to each subordinate instance to order them. Instances with empty Sort[n] field are ordered to the end. In case of matching serial number value (or if empty), the records are ordered alphabetically according to Target column.

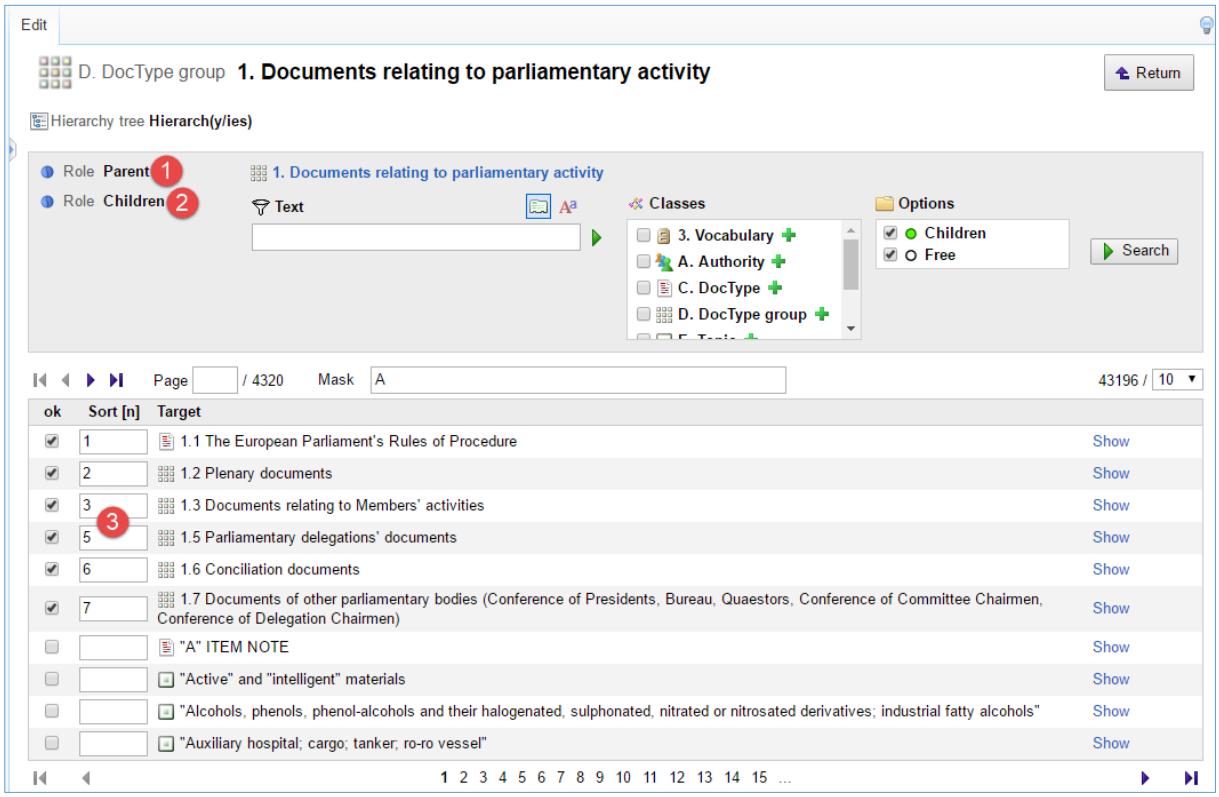


Figure 84: Study database - Hierarchy editing

1. *Superior role in Hierarchy*
2. *Subordinate role in Hierarchy*
3. *Field for entering order when displayed*

Annex 2: Use of NAL entries

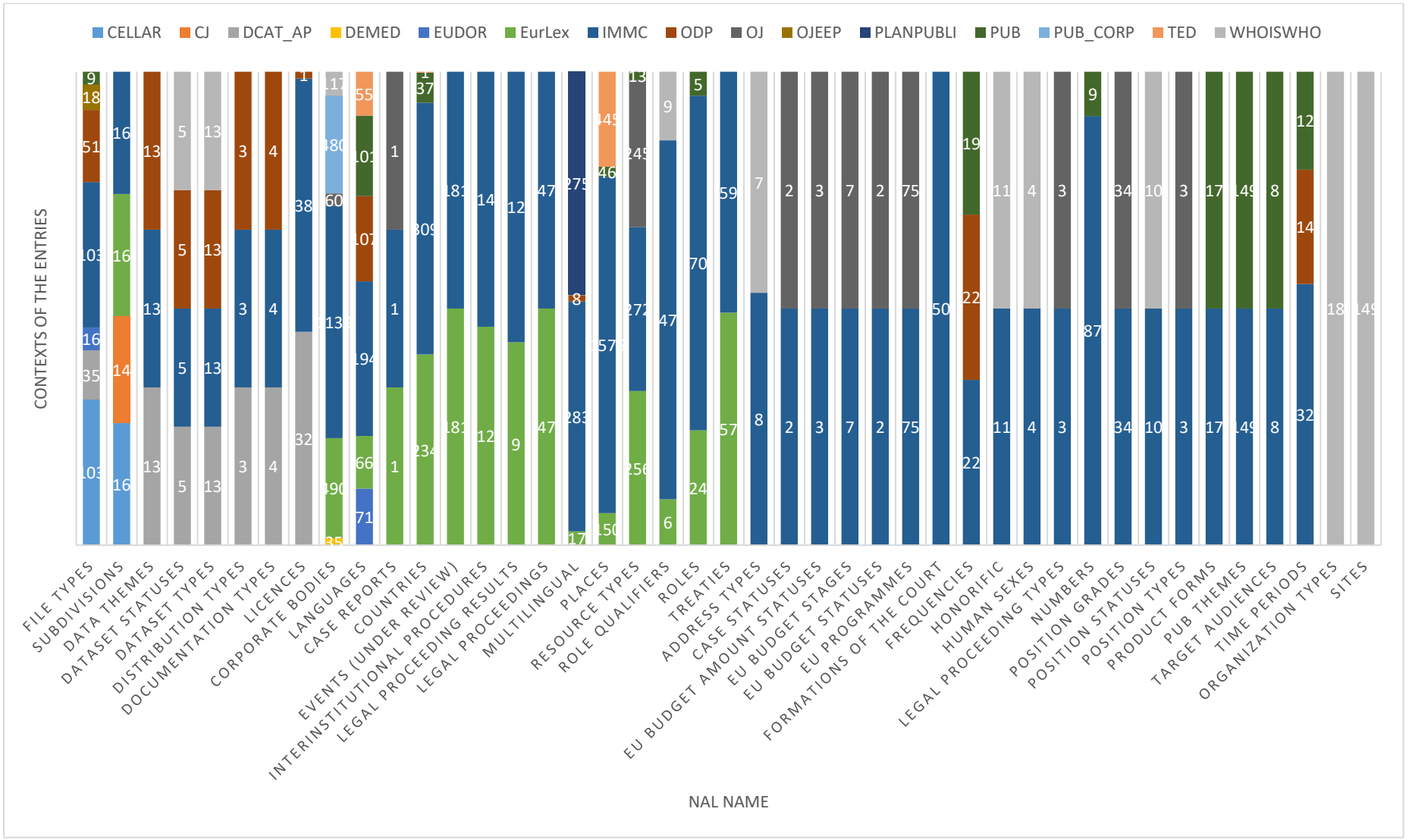


Figure 85: Use of NAL entries

List of Tables

List of Figures

List of the abbreviations used

A	
API	Application Programming Interface
ATTO	Atelier for Translation Tables in the Office
B	
BEREC	Body of European Regulators for Electronic Communications
C	
CASE	Central Archives Search Engine of the Council of the European Union
CCL	Common Command Language
CDM	Common Data Model
CEU	Council of the European Union
CEU-DAC	Council database of agreements and conventions
CHARTER	Charter of Fundamental Rights of the European Union
CJEU	Court of Justice of the European Union
CMS	Content Management System
CoR	The Committee of the Regions
CoR-DM	Committee of the Regions - Documents Manager
CoR-DS	Committee of the Regions - Documents Search
CPV	Common Procurement Vocabulary
CSV	Comma-separated values (file format)
D	
DDoS	Distributed Denial of Service
DG	Directorate-General
DOC	Microsoft Word file format
DOCX	Microsoft Word open file format
E	
EACEA	Education, Audiovisual and Culture Executive Agency
EASME	Executive agency for small and medium-sized enterprises
EC	European Commission
ECA	European Court of Auditors
ECAS	European Commission Authentication Service
ECB	European Central Bank
ECLI	European Case Law Identifier
EESC	The European Economic and Social Committee
EMM	Europe Media Monitor
EO	European Ombudsman
EO-RC	European Ombudsman - Register of Cases
EO-RR	European Ombudsman - Register of Resources
EP	European Parliament
EPUB	E-Book file format
ERC	European Research Council
ERCEA	European Research Council Executive Agency
EU	European Union
EUIPO	European Union Intellectual Property Office

EUIPO-CLR	European Union Intellectual Property Office - Case Law Register
	F
FORMEX	Formalised Exchange of Electronic Publications
FRBR	Functional Requirements for Bibliographic Records
FTP	File Transfer Protocol
	H
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
	I
IMMC	Interinstitutional Metadata Maintenance Committee
INEA	Innovations and Networks Executive Agency
IPEX	The InterParliamentary EU information eXchange
ISBN	International Standard Book Number
ISSN	International Standard Serial Number
	J
JEL	Journal of Economic Literature
JEX	JRC EuroVoc Indexer
JSON	JavaScript Object Notation
	M
MDR	Metadata Registry
MEPs	Members of the European Parliament
	N
NAL	Named Authority List
NER	Named Entity Recognition
NLP	Natural Language Processing
NRAs	National Regulatory Authorities
NUTS	Nomenclature of Territorial Units of statistics
	O
ODP	European Union Open Data Portal
OIB	Office for Infrastructure and Logistics in Brussels
OJ	Official Journal of the European Union
OP	Publications Office of the European Union
OWL	Ontology Web Language
	P
Study	PublicAccess.eu Study (this document)
Study database	PublicAccess.eu Study database (database accompanying the study)
PDF	Portable Document Format
	R
RD-BEREC	Body of European Regulators for Electronic Communications - Document Register
RD-CEU	Public register of Council documents
RD-EC	Register of Commission documents
RD-EESC	The European Economic and Social Committee - Register of documents
RD-EP	The European Parliament's Register of Documents

RED	Committee of the Regions & European Economic and Social Committee - Public Documents
RSS	Rich Site Summary, Really Simple Syndication
	S
SEO	Search Engine Optimization
SPARQL	Simple Protocol and RDF (Resource Description Framework) Query Language
SQL	Structured Query Language
	T
TED	Tenders Electronic Daily
TFEU	Treaty on Functioning of the European Union
TOAD	Committee of the Regions - Transfer of Administrative Documents
TSV	Tab-separated values (text file format)
	U
URL	Uniform Resource Locator
	W
WEMI	Work, Expression, Manifestation, Item
	X
XLS	Microsoft Excel file format
XML	Extensible Markup Language
XSD	XML Schema Definition

Glossary of the terms used

The following table lists and explains the terms used in the study.

A	
Adobe ColdFusion	A commercial rapid web application development platform owned by Adobe Systems.
Advanced search	The *search* engine user interface allowing the use of *search queries* that consist of several search criteria.
Apache Solr Search engine	An open source enterprise *search* platform, written in Java, from the Apache Lucene project.
Application programming interface (API)	Set of routines, protocols, and tools for building software applications, retrieving or storing the data.
Attribute (of a document)	A meta-information directly describing a property of the *document*. It is typically displayed as the value of a variable in the user interface. Usually, it is a character string, number or code. Attributes are also used as a description of a certain filtering condition in *advanced search* forms which is directly typed (that means not selected from a list).
Authentication	Unique identifying information from each system user, generally in the form of a username and password.
Authority	The institution, body, person, etc. with the right to exercise power over the document or responsible for the creation of the document.
Authorization	The process of adding or denying individual user access to a computer network and its resources. Users may be given different authorization levels that limit their access to the associated resources.
B	
Back-end	An application, program or infrastructure serving indirectly in support of the *front-end* services, usually by being closer to the required resource or having the capability to communicate with the required resource.
C	
Cardinality	The number of elements in a given mathematical set, it may refer to the uniqueness of data values. In the view of the study, cardinality refers to the relationships between *document* and entry in the *vocabulary*. It can be one-to-one, many-to-one or many-to-many.
Celex	The unique identifier of each *document* in EUR-Lex, regardless of language.
Comma-separated values (CSV) file format	A file format which stores tabular data (numbers and text) in plain-text. Each line of the file is a data record. Each data record consists of one or more fields, separated by commas.
Content (of the document)	Information or communication expressed in the digital *document*/publication in a specific form (e. g. text, picture, audio, video) or through a combination of various forms, providing something that is to be presented by an *originator* of a *document* and perceived by the user.
Content management system (CMS)	A computer application that supports the creation and modification of the *documents* (both the *content* and the *context*) using a common user interface and thus usually supporting multiple users working in a collaborative environment.

Context (of the document)	Set of *metadata* or meta-information structures describing the properties of the *document* designed by the document *originator* following a certain intention (sometimes also represented by the *ontology* of the *document source*). Context, as used in the Study, is more general term than the *metadata*. While *metadata* describes the single property of the document, context represents all the *metadata* (all the properties). Up to three types of *metadata* can create the *context* of the *document*: 1. *Vocabulary* 2. *Attribute* 3. Relationship to other document
Contextual suggest	A system providing the user with direct contextual help at the time of them making their *search query.*
D	
Data pump	A feature providing automated data load, typically communicating with the *API*.
Database	A collection of information organised in such a way that a computer program can quickly select desired pieces of data.
DOC file format	A file format for word processing documents, most commonly used in connection with the proprietary Microsoft Word Binary File Format.
Document	A digital record of information which may be preserved or represented in order to serve as evidence for a specific purpose. A document, as used in the study, should meet the following three conditions: 1. It captures the *content* of communication (typically text but also picture/audio/video information) in an appropriate digital format. 2. It exists in the *context* of the other documents described by *metadata* (‘*context*’). 3. It does not change over time. All changes made must lead to the creation of a new and appropriately labelled version. This can be ensured by technical means (e.g. electronically signed *PDF*), but it can also remain the responsibility of the document’s *originator* (i.e. the website). A typical example of document in compliance with this definition is a *PDF* file published in the document registry or a website of an institution, agency or other body. On the contrary, a search result is not a document, because it is dynamically generated based on the input *query* and may vary in time or as well as other dynamic extracts from the databases.
Document register	A repository of the *documents* that an institution controls and maintains. Every document register is also a *document source* but not every *document source* is a document register.
Document source	A web application operated by a specific subject – either an EU institution or an agency which provides access to *documents* to its users in a certain way.
DOCX file format	An open XML-based file format used for documents in the Microsoft Word application.
E	

EPUB file format	An e-book file format that can be downloaded and read on devices like smartphones, tablets, computers, or e-readers. It is a free and open standard published by the International Digital Publishing Forum (IDPF).
e-TrustEx	A platform offered by European Commission to Public Administrations at European, national and regional level to set up the secure exchange of natively digital *documents* or scanned *documents* from system to system via standardized interfaces.
EuroVoc	A multilingual, multidisciplinary *thesaurus* covering the activities of the EU. It contains terms in 23 EU languages, plus in three languages of countries which are a candidate for EU accession; managed by the Publications Office, which moved forward to *ontology*-based *thesaurus* management and *semantic web* technologies conformant to W3C recommendations as well as the latest trends in *thesaurus* standards.
External relations	In view of the study associations between *documents* from one *document source* with another one or from one website to another.
F	
Facet search/Facet filter	A technique for accessing a collection of information represented using a *faceted* classification, allowing users to explore by filtering available information. A *faceted* classification system allows the assignment of multiple classifications to an object, enabling the classifications to be ordered in multiple ways, rather than in a single, pre-determined, *taxonomic* order. Each *facet* typically corresponds to the possible values of a property common to a set of digital objects.
Formalised exchange of electronic publications (FORMEX)	Format for the exchange of data between the Publications Office and its contractors. In particular (but not exclusively), it defines the logical markup for documents which are published in the different series of the Official Journal of the European Union. It is based on the international standard XML.
Front-end	Application or program that the user interacts with directly.
File Transfer Protocol (FTP)	A standard network protocol used to transfer computer files between a client and server on a computer network.
Full text search	Techniques for retrieving a *document* or a collection of documents from the document storage, where the search engine examines all of the words of all *documents* previously stored in the database called index containing the relationships between the words and the documents.
G	
GET Method	Type of HTTP request requesting data from a specified resource.
Google search	A web search engine owned by Google Inc.
H	
HyperText Markup Language (HTML)	The standard markup language used for formal description of the web pages displayable by web browsers.
HyperText Markup Language5 (HTML5)	The fifth and the most up-to-date version of the *HTML* language.
HyperText Transfer Protocol (HTTP)	An application protocol for distributed, collaborative, hypermedia information systems.
I	

Indexing	1. Describing a *document* by means of *metadata* to make them easier to retrieve. 2. In the sense of *search* engine, it means collecting, parsing, and storing data to facilitate fast and accurate information retrieval.
Internal relations	Associations between *documents* within one *document source*.
J	
Jahia	A web content management system with a user interface built using Google Web Toolkit.
JavaScript Object Notation (JSON) format	A lightweight data-interchange format. It is easier for humans to read and write than the XML. It is also easier for machines to parse and generate than the XML.
L	
Logical operators	Operators that denote a logical operation - an instruction in which the quantity being operated on and the results of the operation can each have two values. Logical operations include AND, OR, NAND, XOR, and NOR.
M	
Machine readability	The level how well the data encoded on an appropriate medium in a suitable form can be processed by a computer.
Metadata Registry (MDR)	It registers and maintains definition data (*metadata* elements, named authority lists, schemas, etc.) used by the different European Institutions involved in the legal decision-making process gathered in the Interinstitutional Metadata Maintenance Committee (IMMC) and by the Publications Office in its production and dissemination process.
Metadata	As used in the study, the metadata provides information about one certain aspect or property of the *document*; there are three types of metadata: 1. *Vocabulary* 2. *Attribute* 3. Relationship to other document All the metadata of the *document* create the *context* of the *document*.
Microsoft SharePoint	A web application platform in the Microsoft Office server suite that combines various functions that are traditionally separate applications (intranet, extranet, *content* management, *document* management, personal cloud, social networking, workflow management, etc.).
O	
Ontology	A formal representation of a set of concepts within a domain and the relationships between those concepts used to reason about the properties of that domain and to define the domain. As used in the study, ontology describes the *context* of the *documents* at the formal level.
Originator	The author of the *document*.
Ontology Web Language (OWL)	A World Wide Web Consortium (W3C) standard; stands for a family of knowledge representation languages for authoring *ontologies*.
P	

Parsing	The process of analysing the text of a *document* by a parser, which is a software component that takes input data (frequently text) and results in a kind of a data structure – a kind of *metadata* or structure recognized in the *document*.
Portable Document Format (PDF)	A file format used to present *documents* in a manner independent of application software, hardware, and operating systems. *PDF* was a proprietary format controlled by Adobe Systems, until it was officially released as an open standard and published by the International Organization for Standardization as ISO standard.
PULL data transfer	An act of transmitting files over a computer network where the receiver initiates a data transmission.
PUSH data transfer	An act of transmitting files over a computer network where the sender initiates a data transmission.
R	
Resource Description Framework (RDF)	The family of World Wide Web Consortium (W3C) specifications of data model for conceptual description or modelling, of information that is implemented in web resources.
Responsive design	An approach to web design aimed at crafting sites to provide an optimal viewing and interaction experience - easy reading and navigation with a minimum of resizing, panning, and scrolling - across a wide range of devices (from desktop computer monitors to mobile phones).
Rich Site Summary or Really Simple Syndication (RSS)	A family of standard web feed formats to publish frequently updated information (blog entries, news headlines, audio, video).
S	
Sample document	As used in the study, general overview on the *document* and its *context* from particular *document source*.
Search	*Document* or information retrieval.
Search query	Question made through *search* engine interface.
Semantic web	An evolving extension of the World Wide Web in which the semantics of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the web *content*.
Simple search form	Search engine interface with one query line (or very limited number of query lines).
SPARQL	A semantic query language for databases able to retrieve and manipulate data stored in the Resource Description Framework (RDF) format.
Structured information (data)	Information (data) contained in fields. Its nature and function are identified by *metadata*.
Study database	Special database created as part of the study, gathering several *vocabularies* and other descriptive information regarding analysed *document sources*.
SWOT analysis	An acronym for Strengths, Weaknesses, Opportunities, and Threats – a structured planning method that evaluates those four elements of a project or business venture.
T	

Tab-separated values (TSV) file format	A simple text format for storing data in a tabular structure (e. g. database or spreadsheet data). Each record in the table is one line of the text file. Each field value of a record is separated from the next by a tab stop character.
Taxonomy	A controlled and formal *vocabulary* with parent-child hierarchical relations between concepts.
Thesaurus	A controlled *vocabulary* based on *taxonomy* but adding relations between concepts in various branches and various descriptive fields like synonyms, preferable terms or definitions. Current specifications of the thesaurus are defined by the ISO norm 25964.
Topic	Representation of any concept; subject of the *document*.
Topic Maps	An international industry standard (ISO 13250) for information management and interchange. The Topic Maps Data Model is the heart of the Topic Maps standards and is supported by several file formats, query languages and modelling languages. A topic map in a software system is usually managed using a Topic Maps engine. (http://www.topicmaps.org/)
U	
Uniform Resource Locator (URL)	A reference to a web resource that specifies its location on a computer network and a mechanism for retrieving it.
V	
Verity Autonomy Engine	The search engine that performs searches against collections, not against the actual *documents* owned by a company called Autonomy.
Vocabulary	A file/list of *metadata* entries created for a specific purpose, organised and named, and existing independently on the *document*. Its role is to capture the *metadata*, i.e. specific *document* properties in the form of a relationship between a *document* and entries/entries in it. It should be possible to describe individual vocabulary entries. It may take the form of a flat list, *taxonomy* (tree) or *thesaurus*.
X	
XLS file format	A Microsoft proprietary binary file format used for the Excel application.
XML (Extensible Markup Language)	A markup language that defines a set of rules for encoding *documents* in a format that is both human-readable and machine-readable.
XSD (XML Schema Definition)	A recommendation of the World Wide Web Consortium (W3C) specifying how to formally describe the elements in an Extensible Markup Language (XML) *document*.
Z	
ZIP file format	An archive file format that supports lossless data compression. A ZIP file may contain one or more files or directories that may have been compressed.