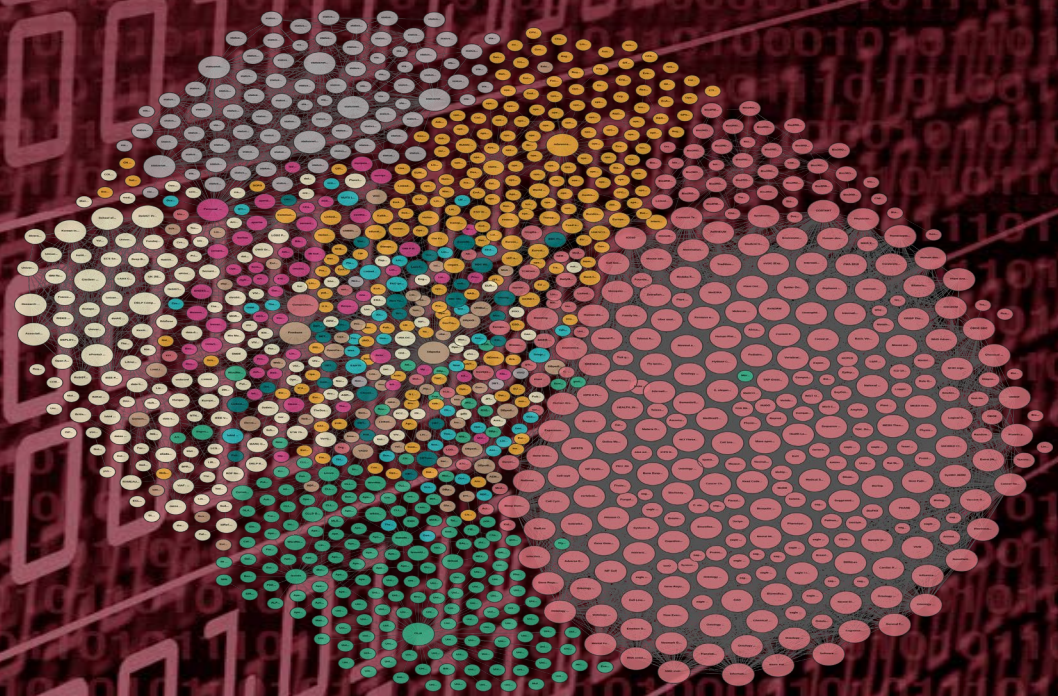
Linguistic Linked Data: open challenges and a roadmap



ENDORSE follow-up event
Publications Office of the European Union
14/05/2024

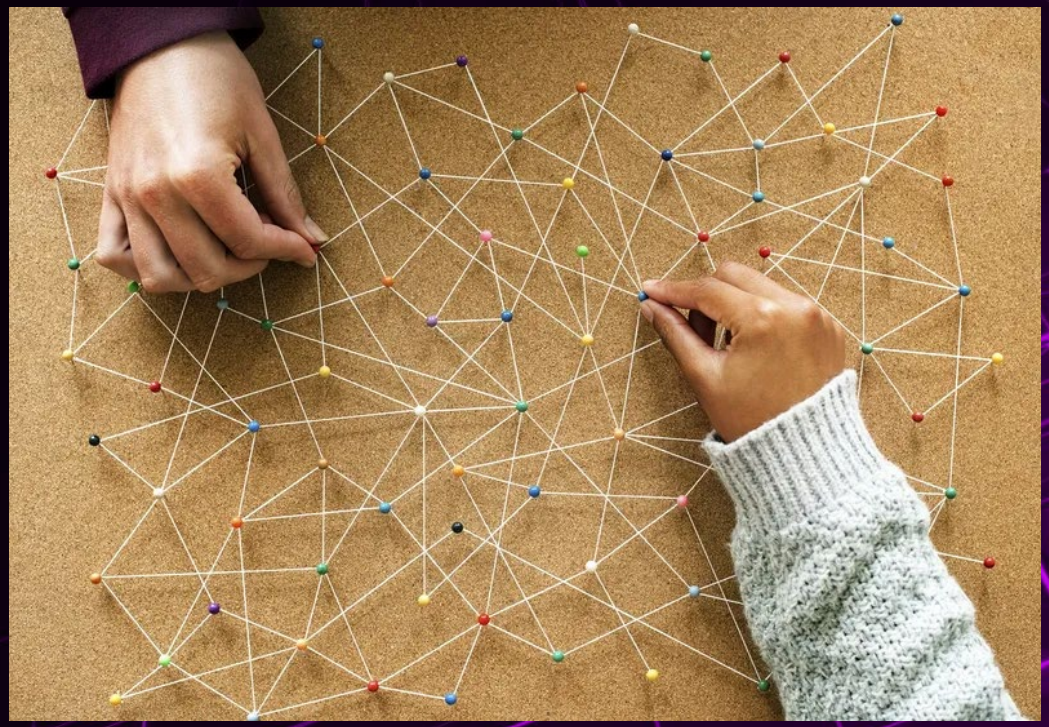
Jorge Gracia

Aragon Institute of Engineering Research
University of Zaragoza
jogracia@unizar.es
http://jogracia.url.ph/web/

Acknolwedgements:

Nexus Linguarum

LLD Motivation



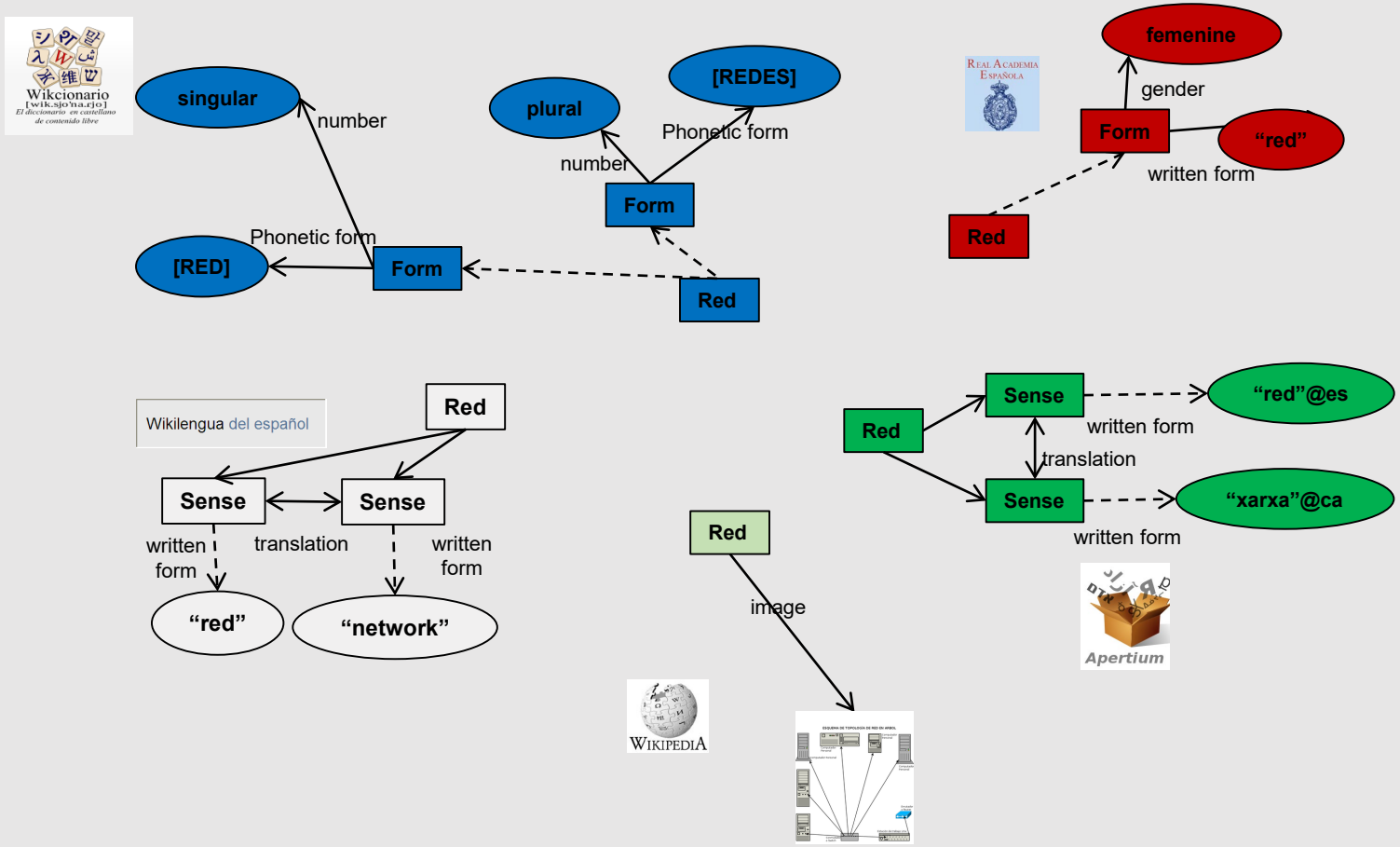
[Image from rawpixel]

LLOD - Motivation

Language Resources as DATA SILOS

[image by Doc Searls in flickr]

LLD - Motivation



lemon

[taken from previous presentations at OEG, UPM]

LLD - Motivation

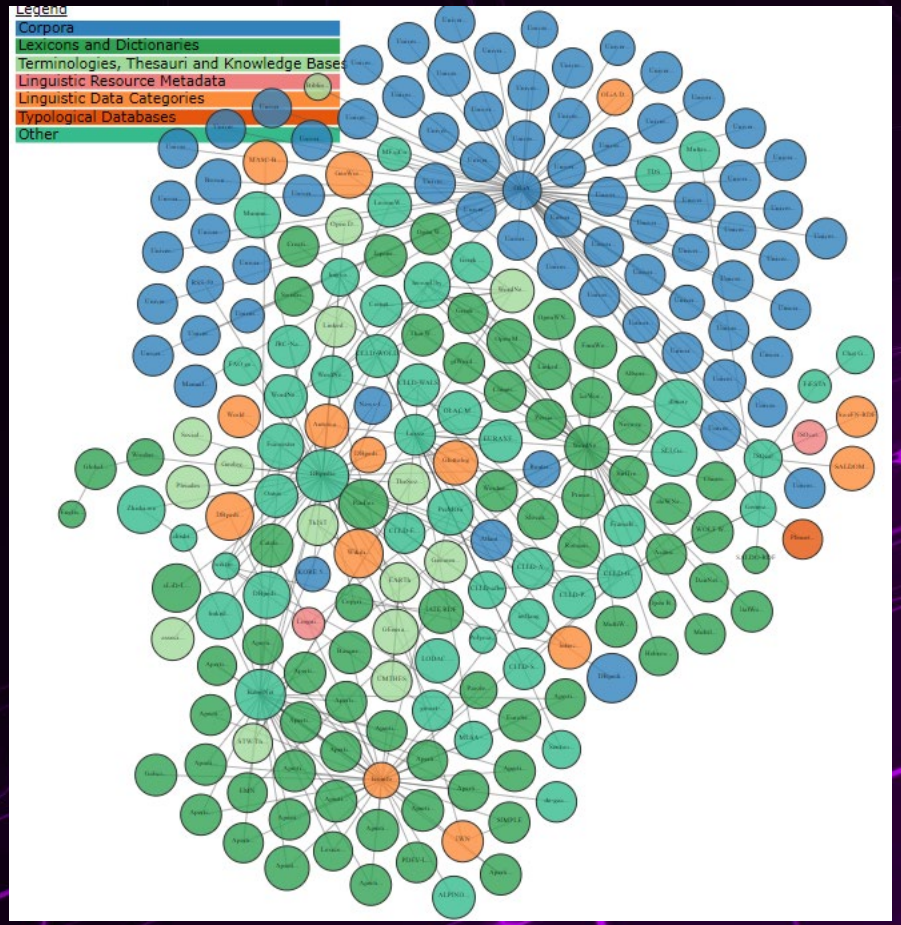
Some BENEFITS of LRs as Linked Data

- Aggregation and integration of linguistic resources
- Resources are explicitly linked
- Data exposed in a standardized way
- Improved discovery of dataset and services
- Use of common vocabularies for representing language content and metadata

lemon

NIF
NLP Interchange Format

lexInfo
['lɛksɪnfəʊ]



LLD – Origins and evolution

Timeline

2001 "foundational" paper of the Semantic Web

2005 development of OLiA (Ontologies of Linguistic Annotation) starts

2010 Open Linguistics Working Group founded at OKF,
Lexical model for ontologies (lemon) published

2011 W3C Ontology-lexica (Ontolex) community group founded

2012 First LDL workshop, first implementation of the LLOD cloud

Timeline

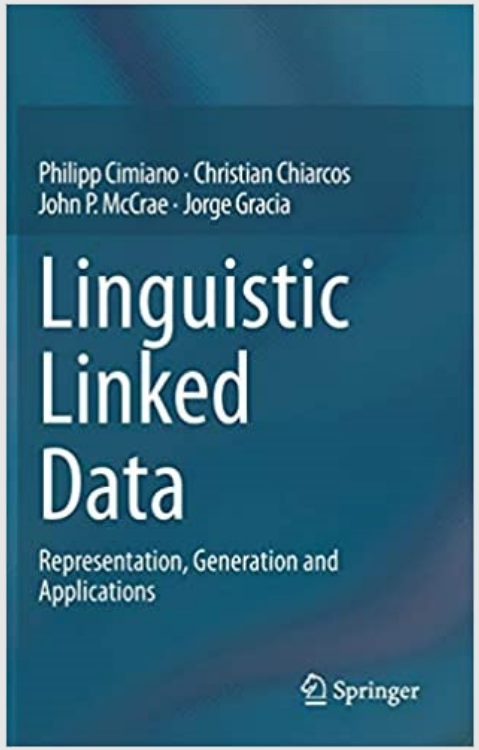
2013 W3C Best Practices for Multilingual Linked Open Data (BPMLOD) group

2014 W3C Linked Data for Language Technology (LD4LT) group

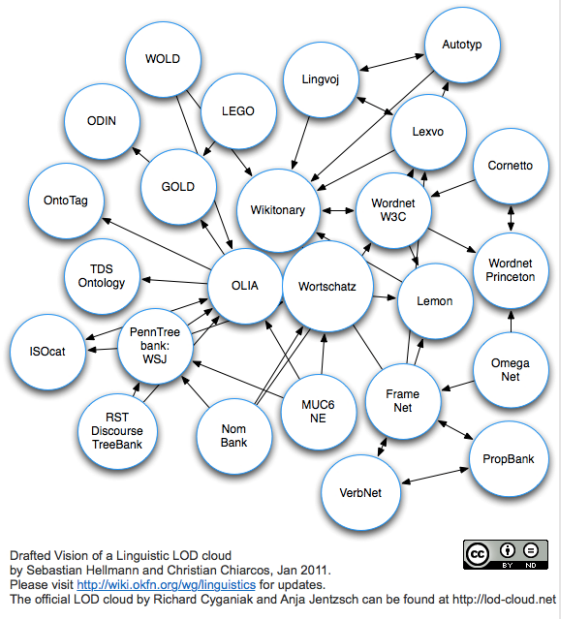
2016 Ontolex lemon specification released

2019 Ontolex lexicography module (lexicog) released

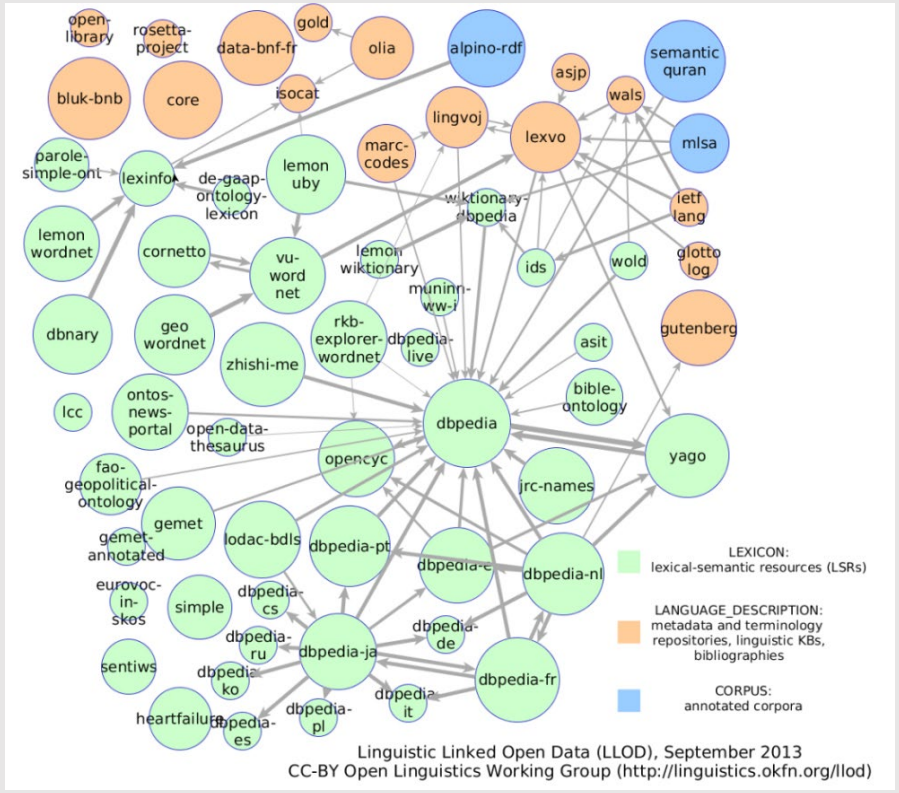
2020 Linguistic Linked Data book



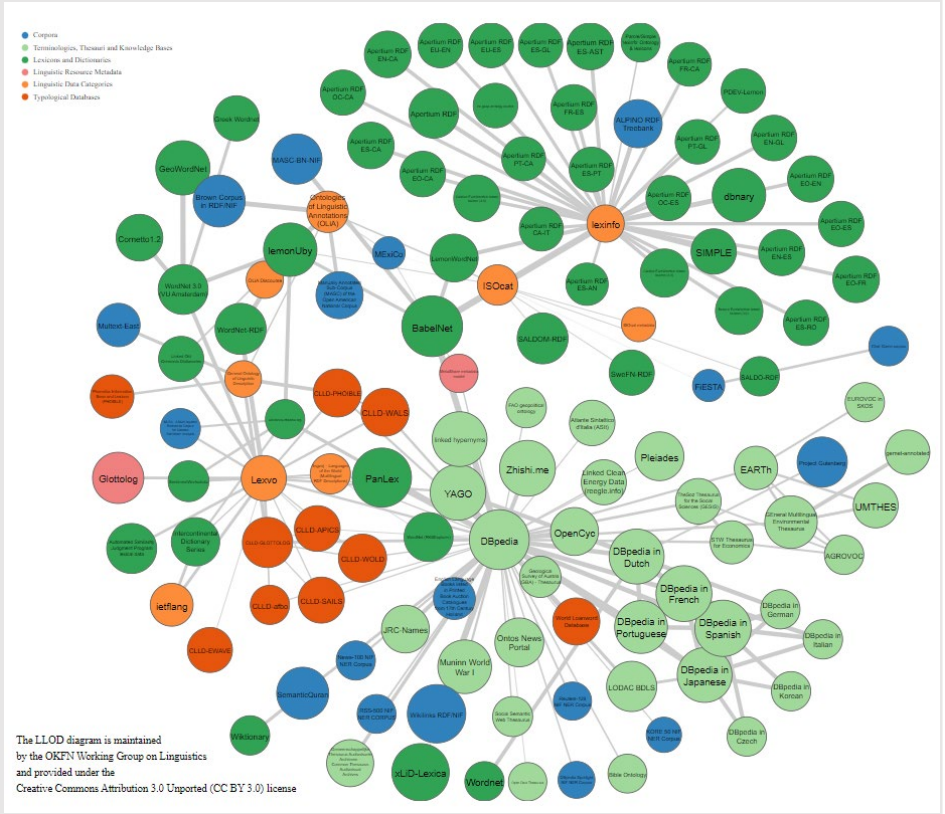
LLOD cloud evolution



2011

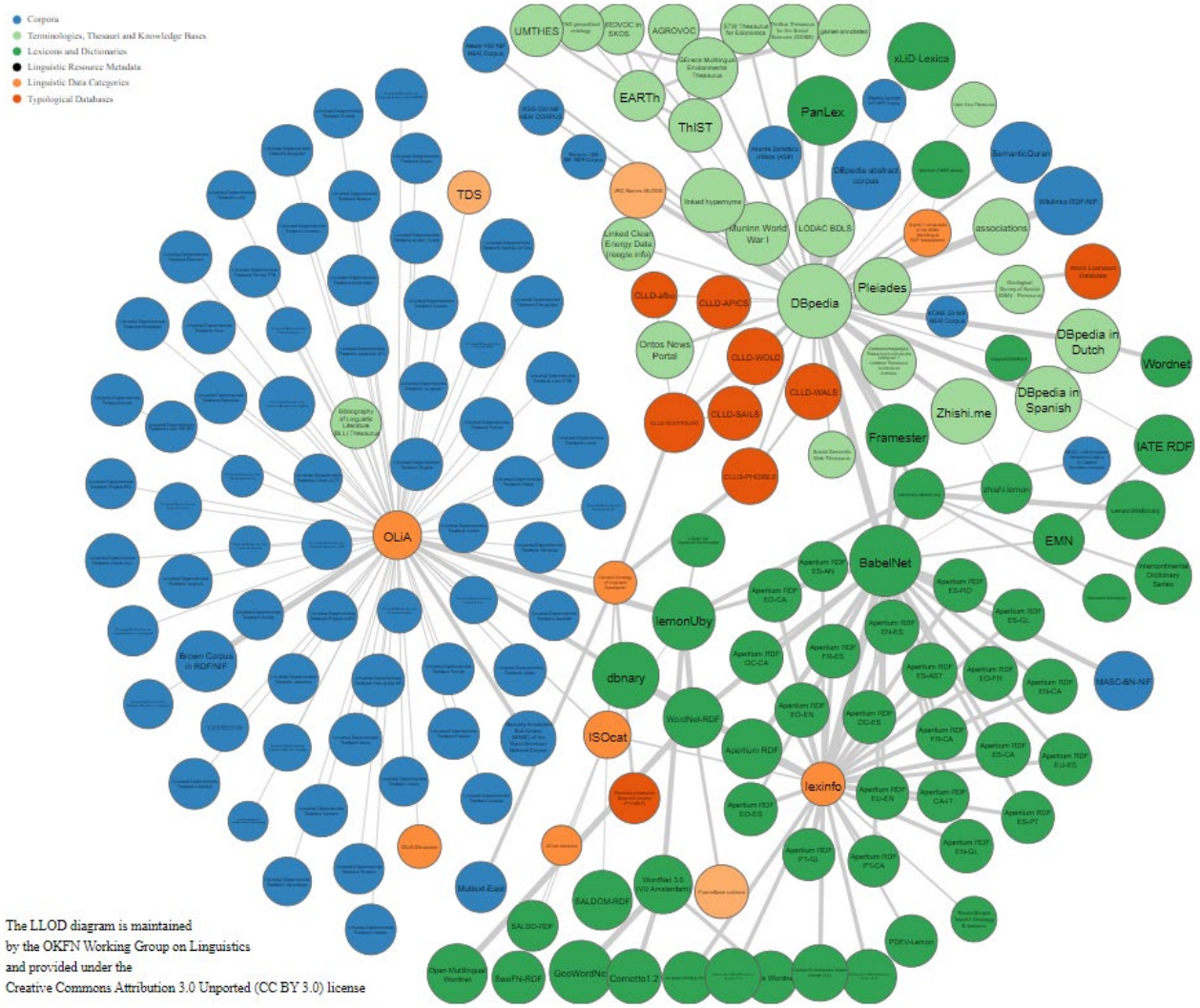


2013

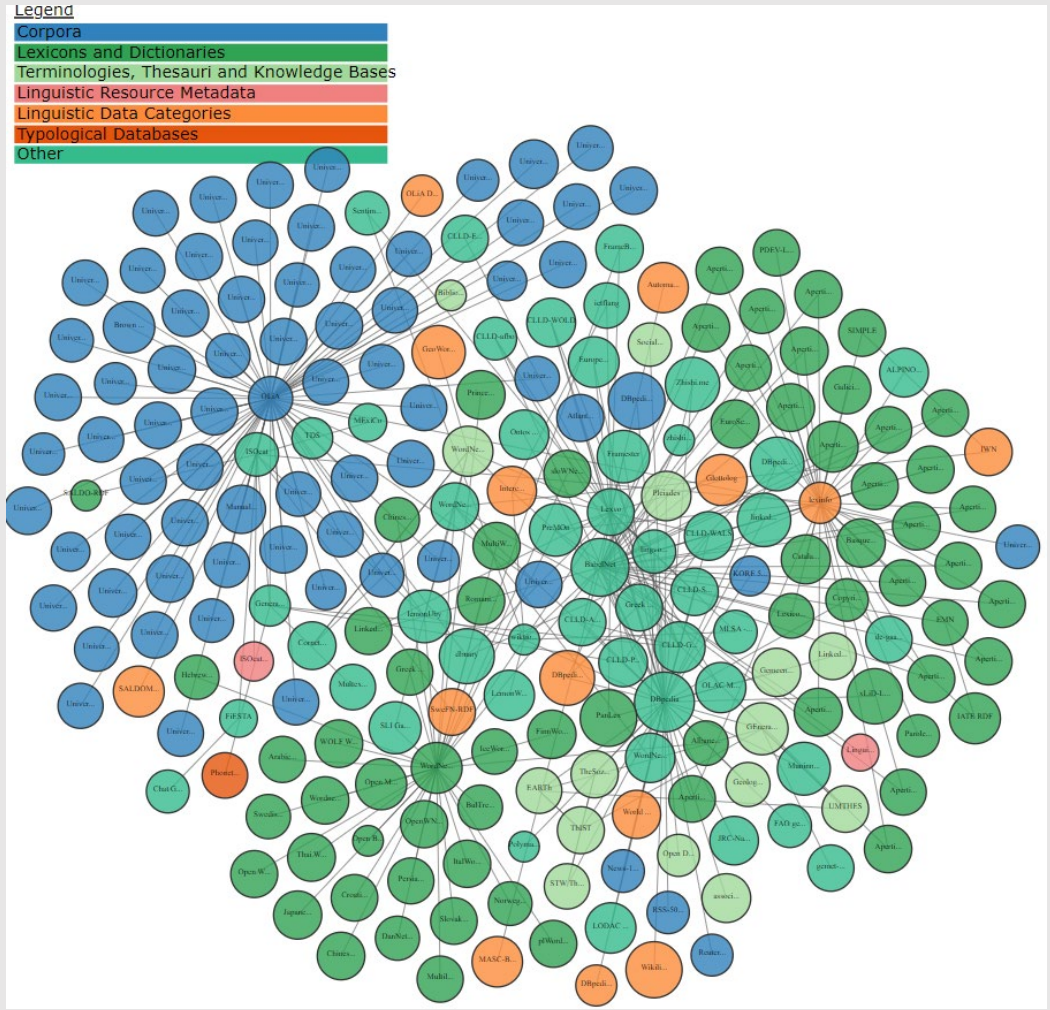


2015

LLOD cloud evolution

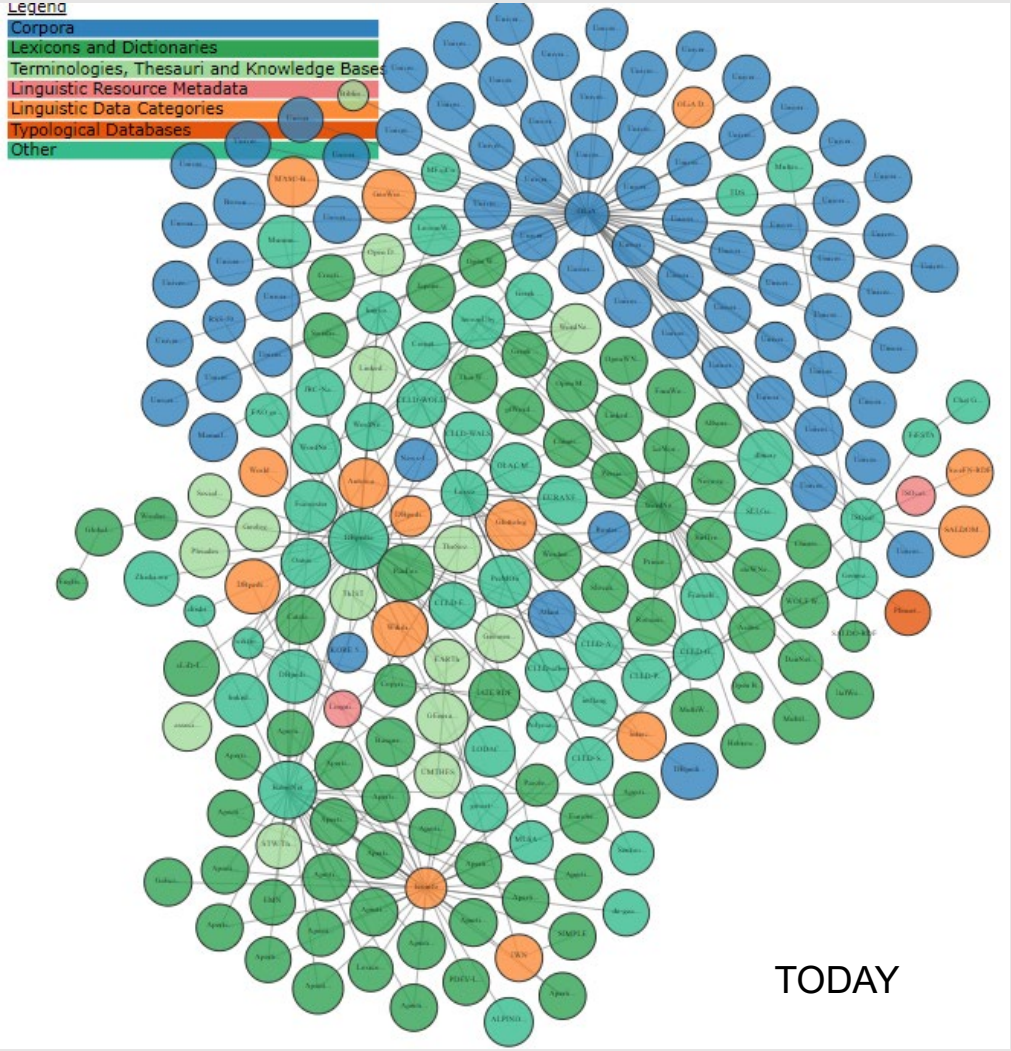


2017



2019

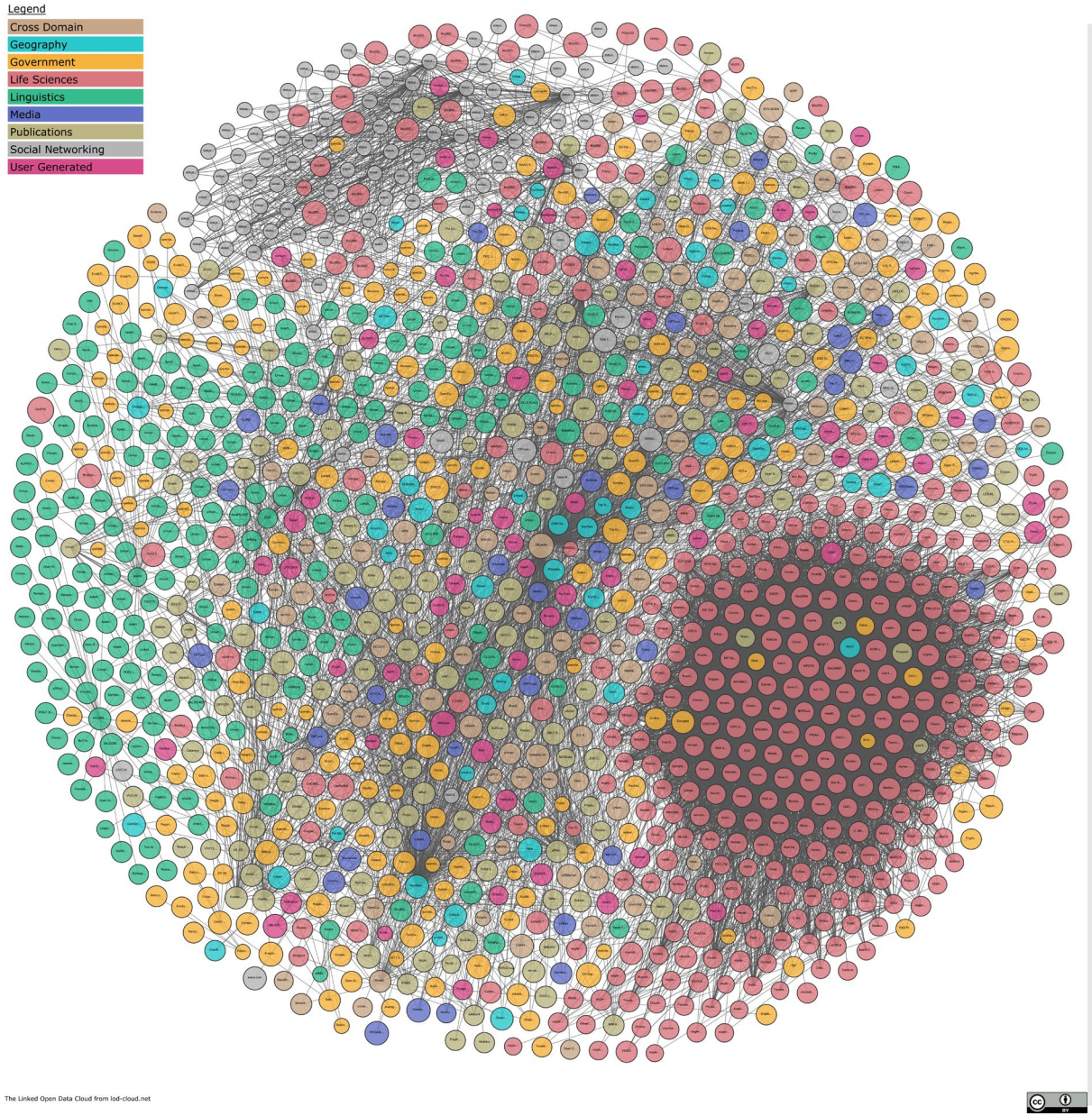
LLOD cloud today



http://linguistic-lod.org/

LOD cloud today

“Linguistics” in green



Well-known Vocabularies for Linguistic Linked Data

- Lexicons and Dictionaries:
  - OntoLex-Lemon, lexicog
- Metadata vocabularies:
  - Basic Metadata: Dublin Core , FOAF, DCTERMS, Prov-O
  - Linguistic Metadata: lime, METASHARE
- Terminology and Thesauri:
  - SKOS (-XL)
- Corpora and Annotation:
  - NIF, Web Annotation
- Data Categories:
  - Lexinfo, Lexvo, OLiA

Large adoption of LLD methods and models

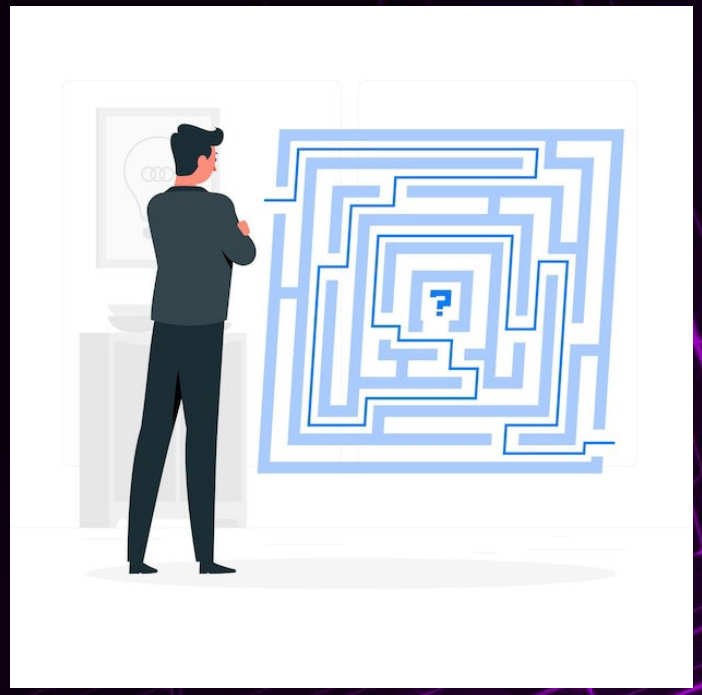
- WordNet
- Global WordNet
- BabelNet
- DBnary
- Apertium RDF
- Lemon-UBY
- Lexvo.org
- Glottolog

- Panlex
- Parole/simple
- KDictionaries
- LiLa
- Wikidata

…

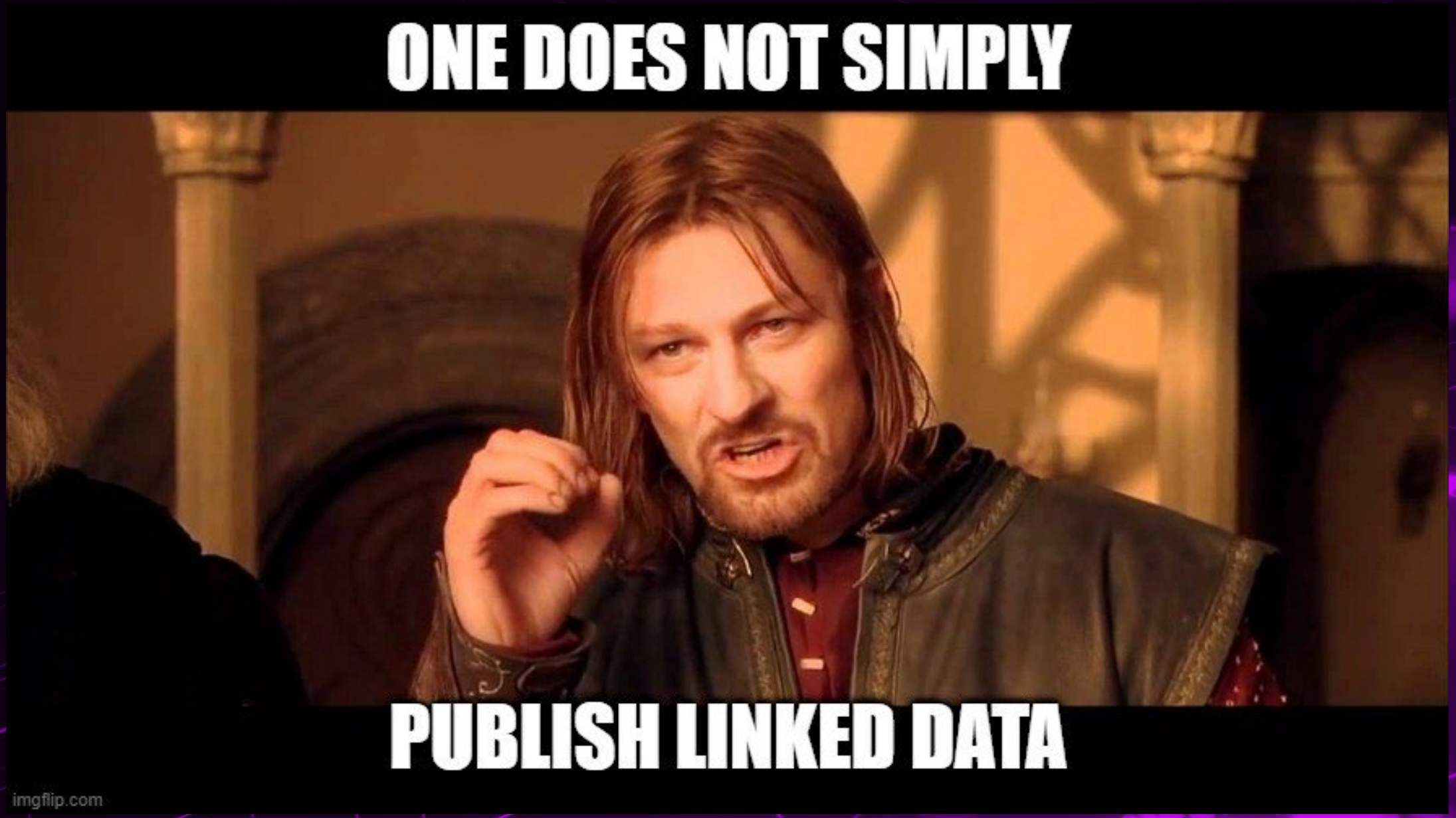
[and many more!]

challenges of LLD



[source of images: freepik.com]

Dagmar Gromann et al. “Multilinguality and LLOD: A Survey Across Linguistic Description Levels”. Semantic Web Journal, 2024

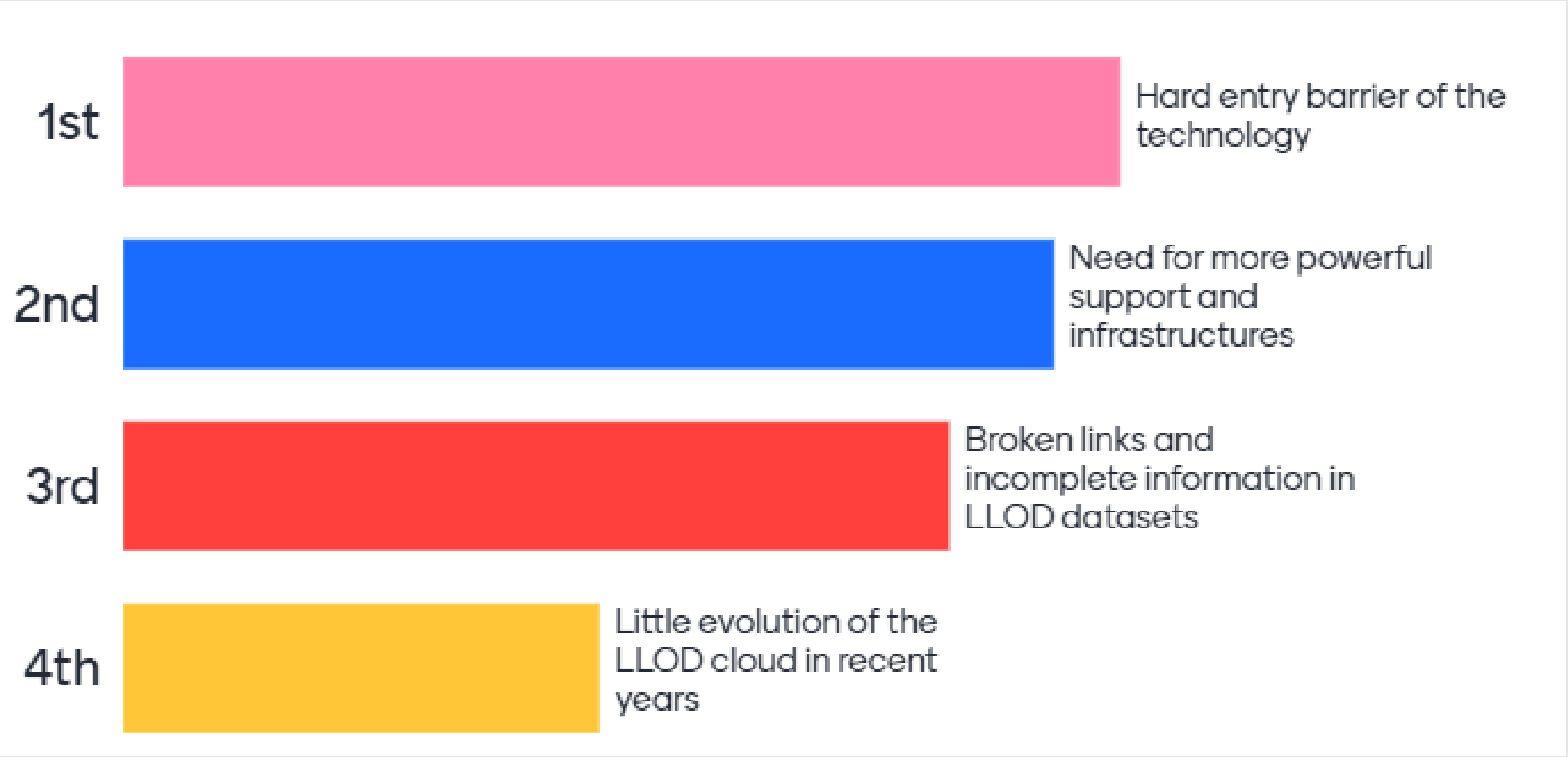

ONE DOES NOT SIMPLY
PUBLISH LINKED DATA
imgflip.com

Some challenges of LLD

- Entry barriers to the technology
- Sustainability
- Coverage of representation models
- Cross-lingual linking

NexusLinguarum questionnaire

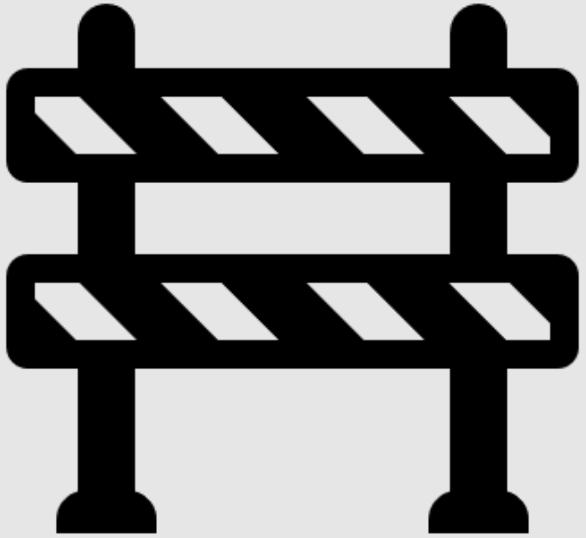
‘‘RANK the following **issues** of LLOD according to importance’’



Some challenges of LLD

Entry barriers to the technology

- steep learning curve
- need of technical support
- accessing isues to some datasets

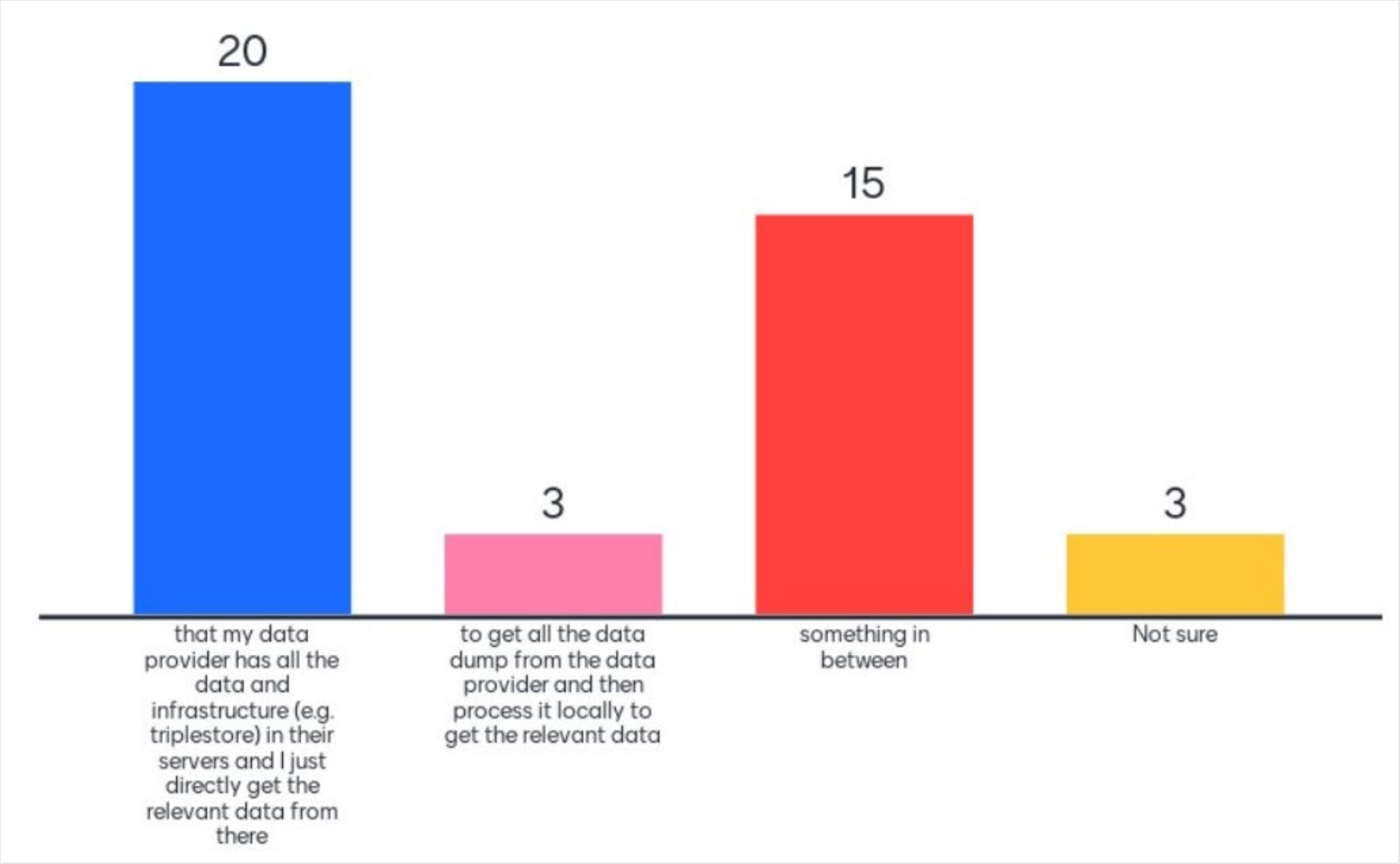


$\Rightarrow$ Investment in education

$\Rightarrow$ more visual and user-centered tools

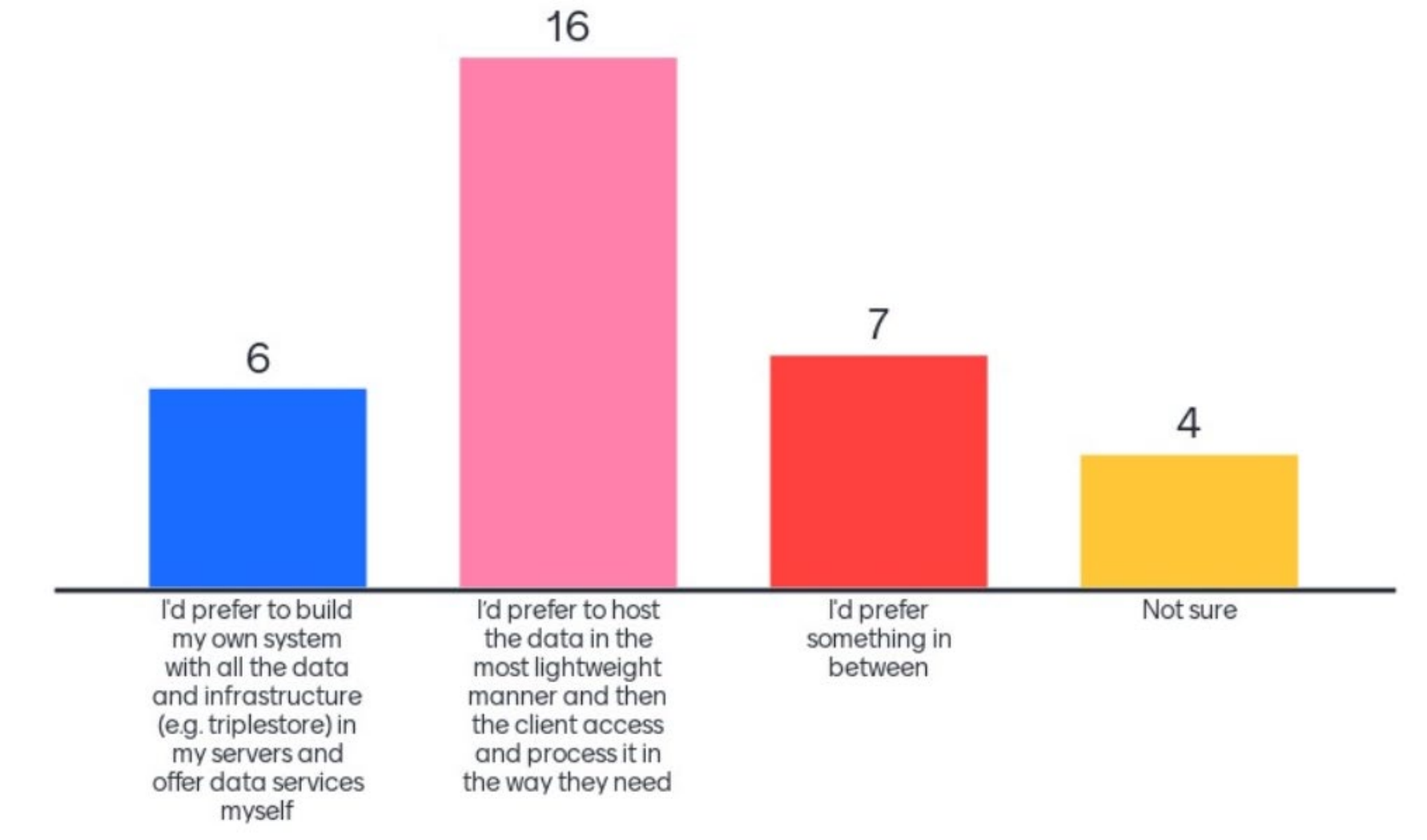
NexusLinguarum questionnaire

"If you were a data **consumer** of the LLOD cloud, you would prefer…"



NexusLinguarum questionnaire

"If you were a data **provider** of the LLOD cloud…"



Some challenges of LLD

Sustainability

- need of sustainable hosting solutions
- burden is either on data providers or data consumers

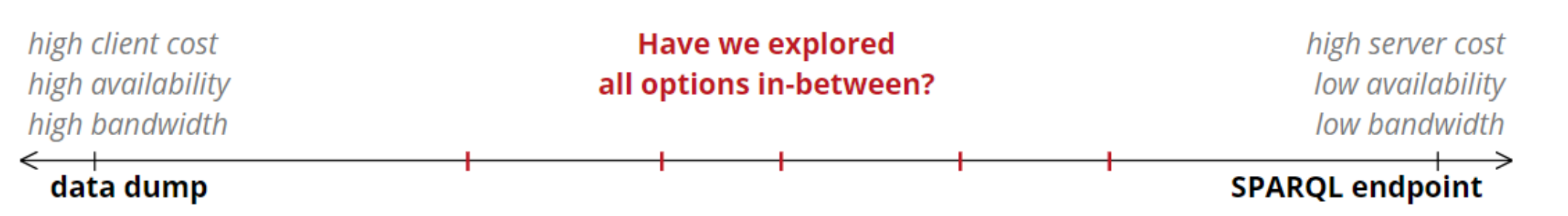
[Source of the image:http://smartcalling.com]

⇒ large infrastructures like CLARIN, ELG, or the European Language Data Space could play a role here

Some challenges of LLD

Sustainability

How to balance efforts between data provider, consumer and host?



[Source of the image: https://linkeddatafragments.org/]

Intermediate solutions have been proposed, like:

- Linked Data Fragments (https://linkeddatafragments.org/)
- SPARQLer (http://www.sparql.org/)
- RDF-HDT (https://www.rdfhdt.org/)
- Hosting of uncompressed RDF dumps with RDF mediatypes

Some challenges of LLD

Sustainability

How to lower the (technical) entry barrier for language resource providers (and consumers)?

Need of more powerful **support and infrastructures**. Something analogous to www.wordpress.org for web sites, but for small linked data providers

Some steps in this direction:

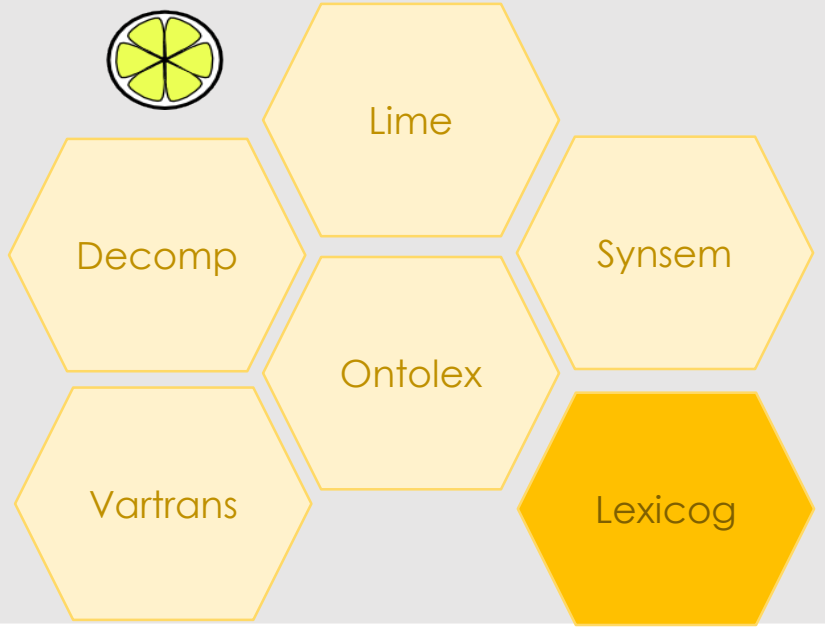
- Databus (https://databus.dbpedia.org/)
- Wikibase (https://wikiba.se/)
- Semantic media wiki (https://www.semantic-mediawiki.org/)
- TriplyDB (https://triply.cc/)

Some challenges of LLD

Coverage of representation models

- some linguistic levels well covered, particularly lexica and semantics
- other levels not so well covered, e.g., phonetics or pragmatics

⇒ work on new models and modules



Some challenges of LLD

Cross-lingual linking

- need of dealing with conceptual mismatches and cultural specificities

⇒ New algorithms / techniques needed

⇒ More benchmarks needed

[source of image: @cristinn at Adobe Stock - 158423491]

a roadmap for LLD

NexusLinguarum deliverable D5.1 on ”Roadmap and common agenda for future research on linguistic data science”, 2024

Roadmap – STEP 1

1.1 More robust and sustainable open infrastructures

- Technology is already in place -> adoption through new projects with a clear focus on infrastructure development

1.2 More educational efforts needed

- To make the advantages of LLOD visible to a new generation of researchers and practitioners.

Roadmap – STEP 1

Nexus Linguarum

Erasmus Mundus

Joint Master Degree in “linguistic data science” (submitted)

MOOC on LLD (coming soon)

Roadmap – STEP 2

2.1 New models developed

- To cover linguistic description levels currently under-represented in the LLOD cloud

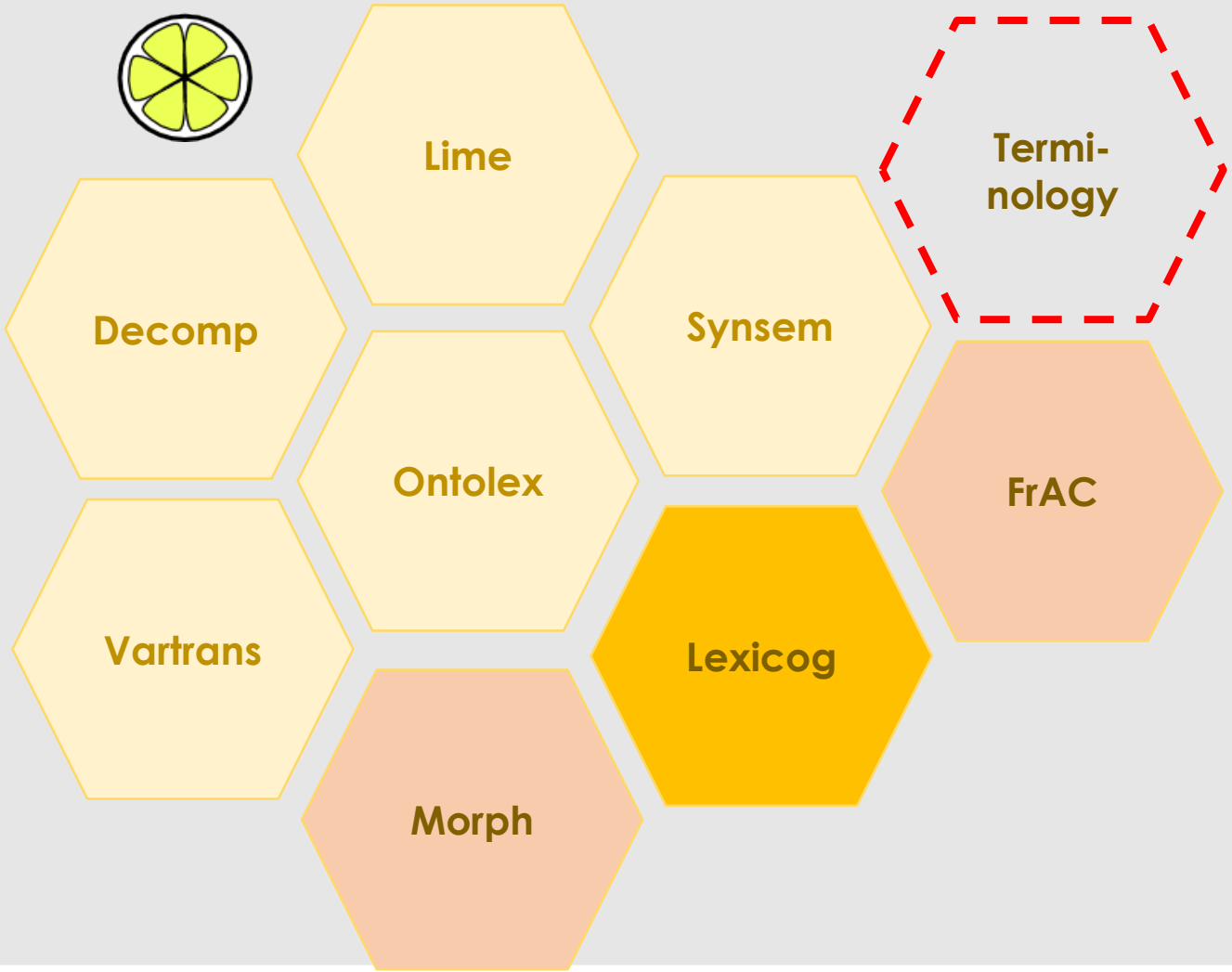
2.2 New guidelines/best practises

- To consolidate existing models and techniques and suport new ones

2.3 New systems for RDF generation and linking

Roadmap – STEP 2

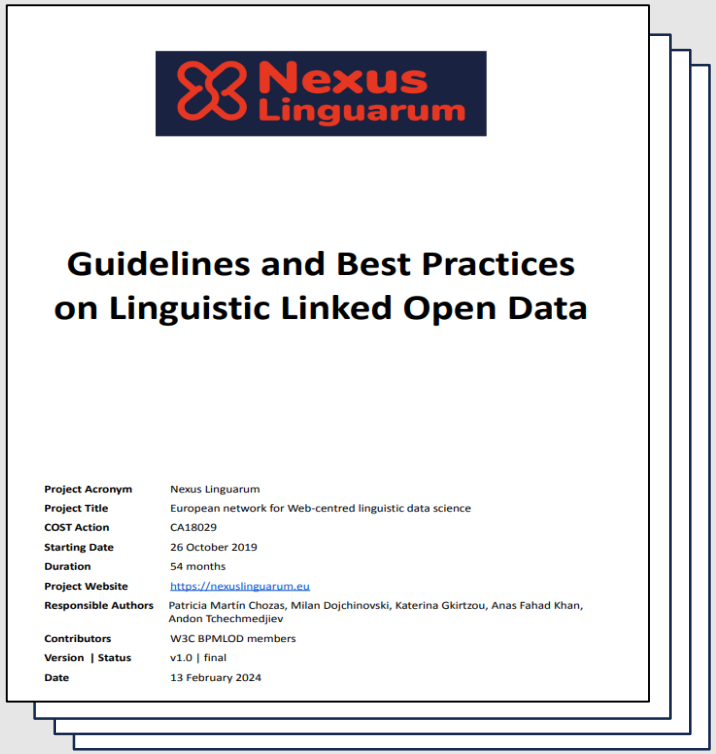
Ontolex new modules



Roadmap – STEP 2

Guidelines and best practises

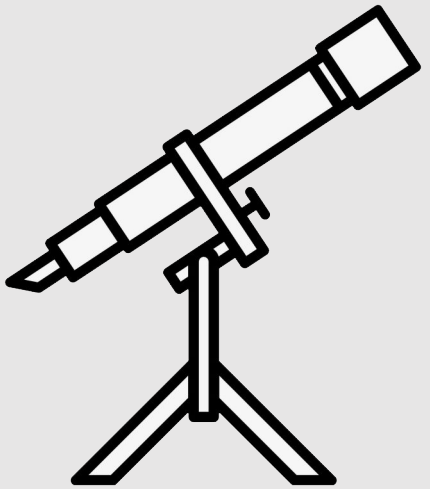
Best Practices for Multilingual Linked Open Data Community Group



Topic	main proposer	contributors	status	link to working document	link to report	comments / relevant links
Terminology generation (TBX)	Patricia	Andon, Maria Pia, Fahad, (Christian), Dagmar	tbc	tbc	tbc	https://www.w3.org/2015/09/bpmlod-reports/multilingual-terminologies/
Bilingual dictionaries	Jorge	Ilan, Fahad, Gilles, (Christian)	draft	github	preview draft report	Update of existing one at http://www.w3.org/2015/09/bpmlod-reports
Crosslingual linking	Mike	Jorge, Katerina, Ilan, Gilles, Thierry	started	See here	tbc	related to Nexus T1.3 activities and position paper
Corpora annotation (NIF, Web annotation)	Milan	Ranka, Christian	tbc	tbc	tbc	see reference card and guidelines document
Guidelines for Developing NIF-based NLP Services	Milan	Ranka, Andon, Katerina, Christian	tbc	tbc	tbc	Update of existing guidelines at http://bpmlod.github.io/report/NIF-based-NLP-WebServices/index.html
Wordnets	Fahad	Mike, Thierry, (Christian)	tbc	tbc		
Guidelines for LLOD aware services	Andon	Katerina, Patricia, (Christian)	tbc	See here	tbc	https://bpmlod.github.io/report/LLOD-aware-services/index.html
LLOD for Under-resourced languages		Giedre, Jorge, Ranka, Fahad, (Christian)	tbc	tbc	tbc	see policy brief and LREC paper
Neuro-symbolic LLOD		Andon, Hugo, Dagmar, Katerina, Cosimo, Christian	tbc	See here	tbc	
Licensing and metadata	Penny	(Ilan), Fahad, Dagmar	tbc	See here	tbc	See Lider related reference card
~~Multimodal data~~		Andon, Fahad, Thierry	dropped	tbc	tbc	focus is mostly on sign languages and there is already an ongoing effort on that direction
Sign languages	Ineke, Andon	Fahad	tbc	tbc	tbc	Task T3.4 of NexusLinguarum
Lexicographic LLOD		Jorge, Marco, Francesco, Ilan, Fahad	tbc	tbc	tbc	

Roadmap – STEP 3

3.1 Development of an “observatory” to measure the quality and evolution of linguistic data on the Web

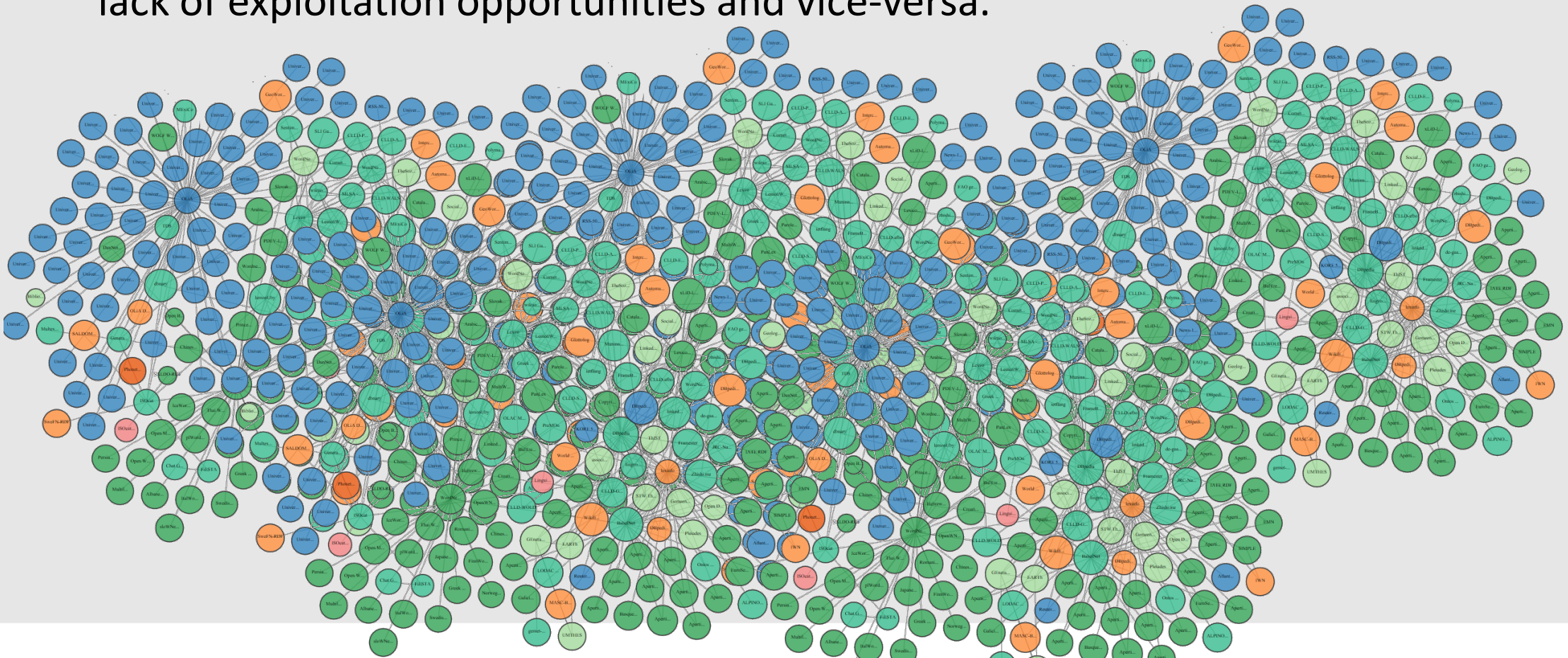


3.2 Stable metadata models and repositories

- not only for discovery of relevant language resources, but really accessing their data and enabling their direct re-use and inter-operation

Roadmap – STEP 4

4. Massive population of the LLOD cloud
   - To cut the vicious circle resulting in lack of data caused by lack of exploitation opportunities and vice-versa.



Roadmap – STEP 5

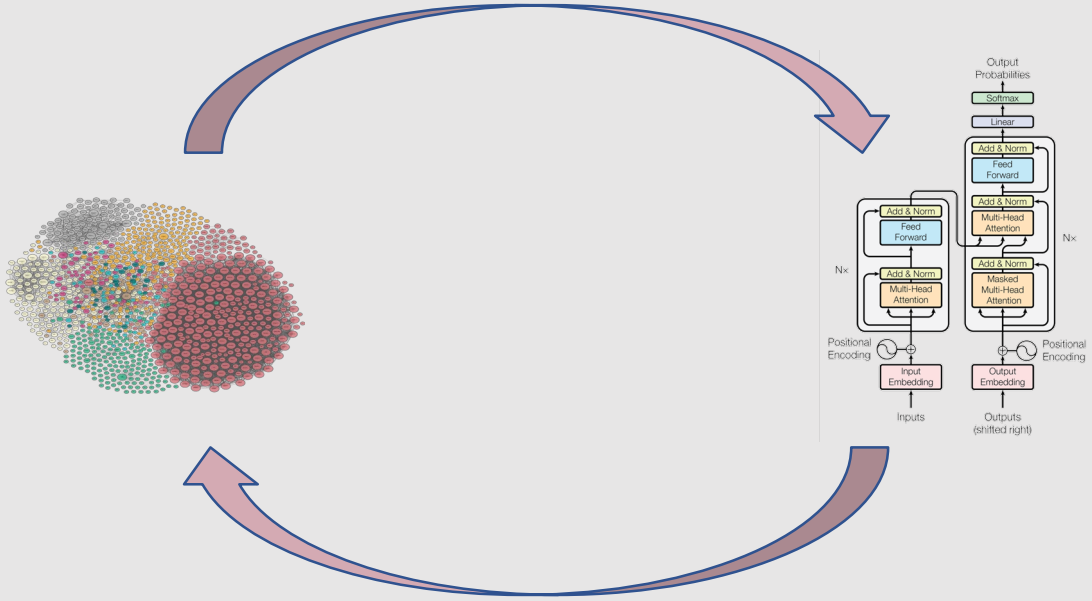
5. Development of a fully fledged family of services
   - For upload, integration, access, querying, browsing, editing of multilingual linguistic data

[Source: DALL-E 3]

Roadmap - BONUS STEP

Research on Hybrid symbolic (linguistic) + non-symbolic approaches

Linguistic knowledge injection into LLMs
(potentially better and more explainable results)



Linguistic graphs enrichment, completion, and querying

Conclusions

Conclusions

Linguistic Linked data:

- Rich story
- Increasing adoption

But

- Many challenges still to address
- New scenarios to cope with (“LLMs”)

A lot of potential for

- Data driven artificial intelligence
- Linguistic Data Science

Some LLD seminal papers

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. "Towards a Linguistic Linked Open Data cloud: The Open Linguistics Working Group." *TAL* (*Traitement Automatique des Langues)*, *52*(3), 245-275. 2011

John McCrae, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, Tobias Wunner. "Interchanging lexical resources on the semantic web" Language Resources and Evaluation 46, 701-719. 2012

Jorge Gracia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar, and John McCrae. "Challenges for the multilingual Web of Data." Journal ofWeb Semantics, vol. 11, pp. 63–71. Elsevier B.V., 2012.

Christian Chiarcos, Steven Moran, Pablo N. Mendes, Sebastian Nordhoff, and Richard Littauer. "Building a Linked Open Data cloud of linguistic resources: Motivations and developments." In Iryna Gurevych and Jungi Kim (eds.), The People's Web Meets NLP. Collaboratively Constructed Language Resources. Springer, Heidelberg, 2013.

Some papers with a recent overview on LLD

Gromann, D., Apostol, E.-S., Chiarcos, C., Cremaschi, M., Gracia, J., Gkirtzou, K., Liebeskind, C., Mockiene, L., Rosner, M., Schuurman, I., Sérasset, G., Silvano, P., Spahiu, B., Utka, A., Truica, C.-O., & Oleškevičienė, G. V. (2024). "Multilinguality and LLOD: A Survey Across Linguistic Description Levels." Semantic Web Journal. 2024

Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambrini, John P McCrae. "When linguistics meets web technologies. Recent advances in modelling linguistic linked open data". Semantic Web Journal. 2022

Thierry Declerck, John Philip McCrae, Matthias Hartung, Jorge Gracia, Christian Chiarcos, Elena Montiel-Ponsoda, Philipp Cimiano, Artem Revenko, Roser Sauri, Deirdre Lee, Stefania Racioppa, Jamal Abdul Nasir, Matthias Orlikowsk, Marta Lanau-Coronas, Christian Fäth, Mariano Rico, Mohammad Fazleh Elahi, Maria Khvalchik, Meritxell Gonzalez, Katharine Cooney. "Recent developments for the linguistic linked open data infrastructure". Proc. of the 12th Language Resources and Evaluation Conference (LREC'22). 2022

Pareja-Lora, A., Lust, B., Blume, M., & Chiarcos, C. (2020). "Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences". The MIT Press.

Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda, Asunción Gómez-Pérez. "Models to represent linguistic linked data" Natural Language Engineering 24 (6), 811-859, 2018

ENDORSE

Follow-up events

[picture by Ashashyou from Wikimedia commons]

Jorge Gracia del Río

Aragon Institute of Engineering Research
University of Zaragoza
jogracia@unizar.es
http://jogracia.url.ph/web/

Nexus Linguarum