TopQuadrant™

ENDORSE 16-19 MARCH 2021

THE EUROPEAN DATA CONFERENCE ON REFERENCE DATA AND SEMANTICS

Automating Reference Data Crosswalks Using Knowledge Graphs

Irene Polikoff

TopQuadrant

irene at topquadrant.com

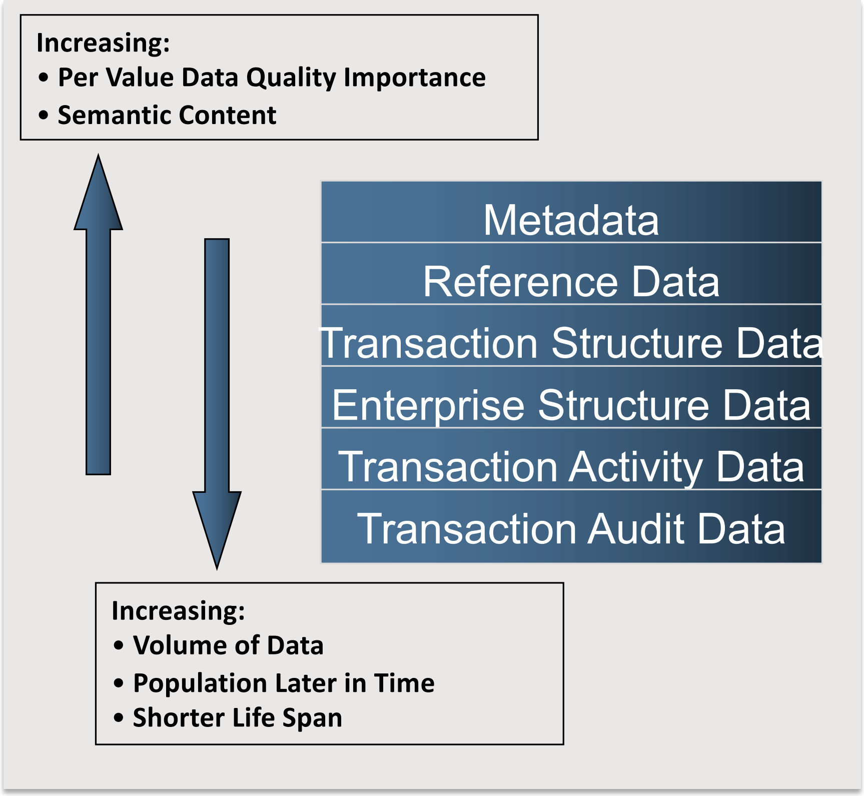
TopBraid EDG™
Enterprise Data Governance

How Reference Data is Used

- Reference data is found in practically every enterprise application
- Mission-critical applications use reference data to classify and categorize master and operational data
- Data warehouses use reference data for aggregation, reporting and analytics

Why Connect Reference Datasets

- Legacy code tables are defined on an application-by-application basis
- How do you then align, aggregate, integrate data?
- Thus, the need to connect or map disparate reference data


Increasing:
• Per Value Data Quality Importance
• Semantic Content
Metadata
Reference Data
Transaction Structure Data
Enterprise Structure Data
Transaction Activity Data
Transaction Audit Data
Increasing:
• Volume of Data
• Population Later in Time
• Shorter Life Span

Reproduced with permission from Malcolm Chisholm

TopQuadrant™

TopBraid EDG Crosswalks

TopBraid EDG™
Enterprise Data Governance

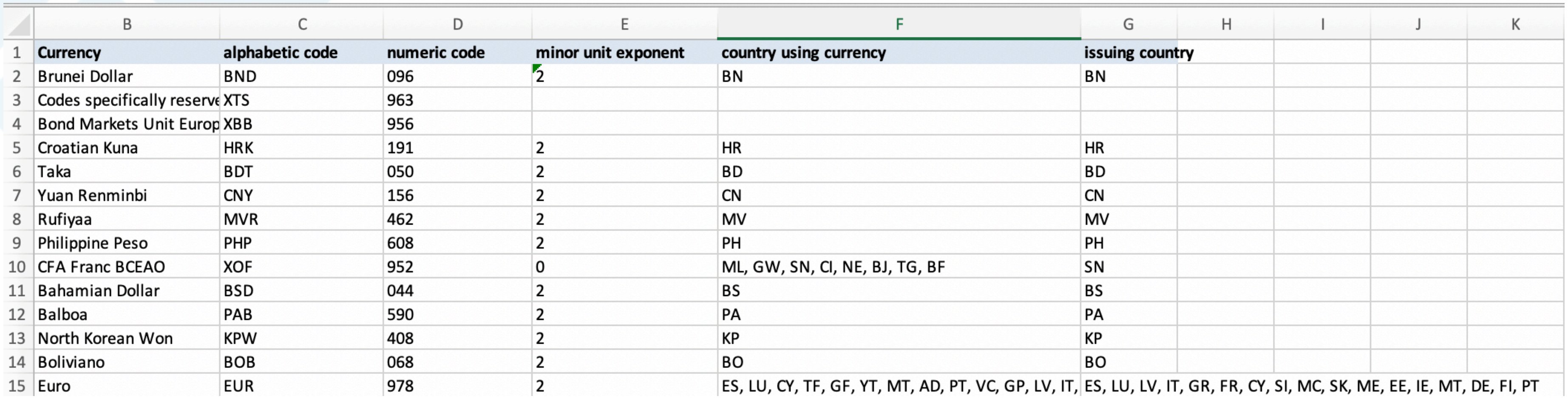
Enterprise Data Governance (**EDG**) addresses data governance needs using heterogeneous data stores, data processing, and applications

Reference Datasets in EDG are graphs

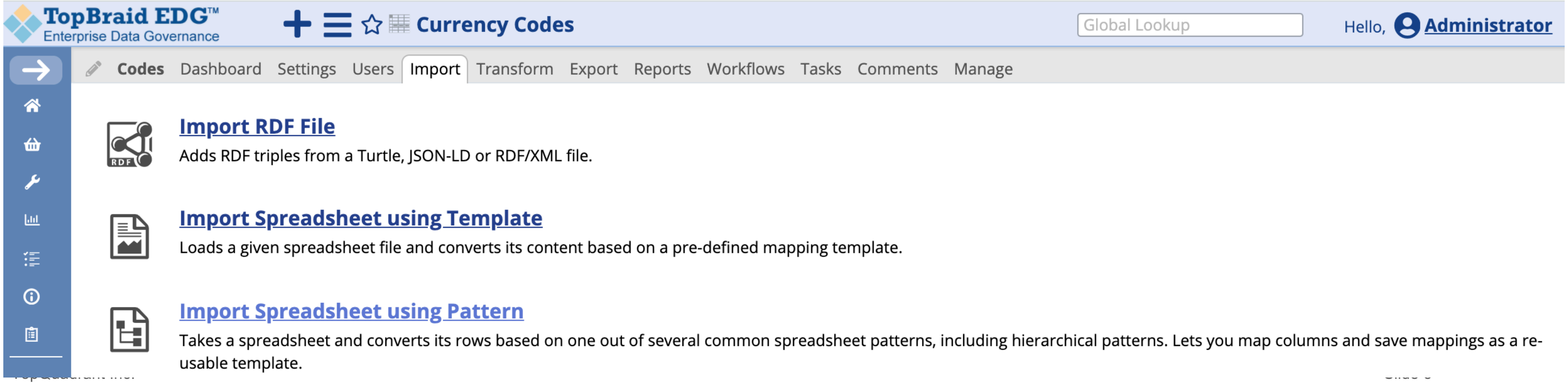
Crosswalks are special graphs that connect terms from two different vocabularies/reference datasets

TopQuadrant™

From Spreadsheet to RDF



	B	C	D	E	F	G	H	I	J	K
1	Currency	alphabetic code	numeric code	minor unit exponent	country using currency	issuing country				
2	Brunei Dollar	BND	096	2	BN	BN				
3	Codes specifically reserve	XTS	963							
4	Bond Markets Unit Europ	XBB	956							
5	Croatian Kuna	HRK	191	2	HR	HR				
6	Taka	BDT	050	2	BD	BD				
7	Yuan Renminbi	CNY	156	2	CN	CN				
8	Rufiyaa	MVR	462	2	MV	MV				
9	Philippine Peso	PHP	608	2	PH	PH				
10	CFA Franc BCEAO	XOF	952	0	ML, GW, SN, CI, NE, BJ, TG, BF	SN				
11	Bahamian Dollar	BSD	044	2	BS	BS				
12	Balboa	PAB	590	2	PA	PA				
13	North Korean Won	KPW	408	2	KP	KP				
14	Boliviano	BOB	068	2	BO	BO				
15	Euro	EUR	978	2	ES, LU, CY, TF, GF, YT, MT, AD, PT, VC, GP, LV, IT,	ES, LU, LV, IT, GR, FR, CY, SI, MC, SK, ME, EE, IE, MT, DE, FI, PT				



Import RDF File

Adds RDF triples from a Turtle, JSON-LD or RDF/XML file.

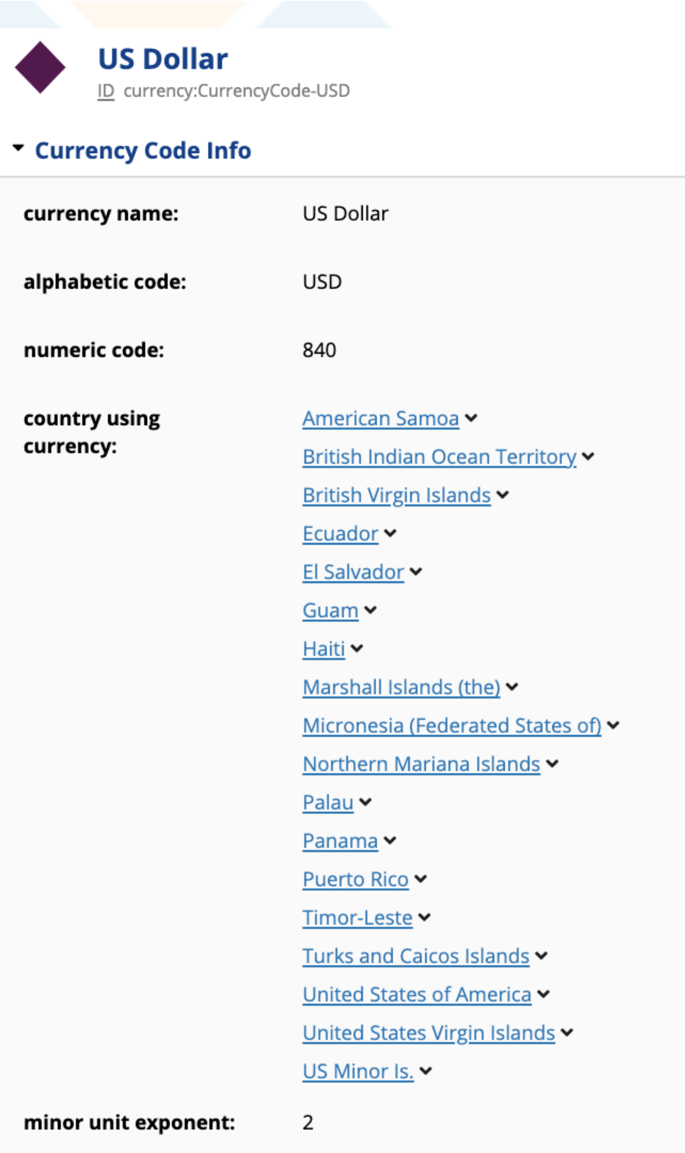
Import Spreadsheet using Template

Loads a given spreadsheet file and converts its content based on a pre-defined mapping template.

Import Spreadsheet using Pattern

Takes a spreadsheet and converts its rows based on one out of several common spreadsheet patterns, including hierarchical patterns. Lets you map columns and save mappings as a re-usable template.

Reference Data as a Graph



US Dollar

ID currency:CurrencyCode-USD

Currency Code Info

currency name: US Dollar

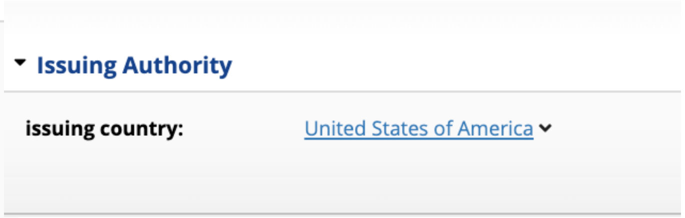
alphabetic code: USD

numeric code: 840

country using currency:

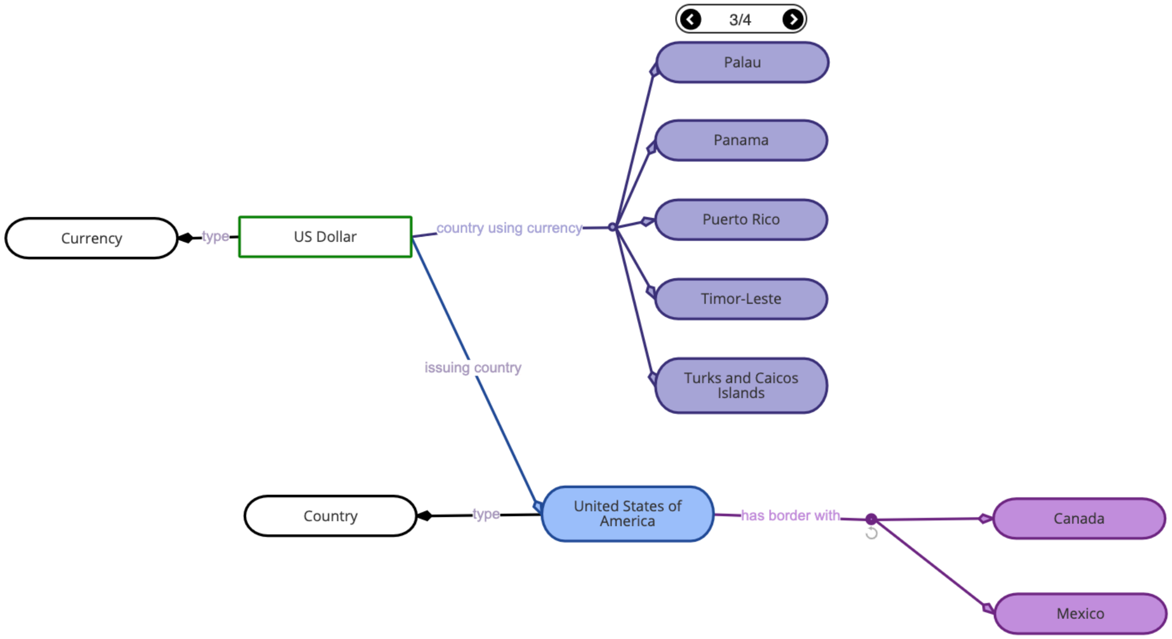
- American Samoa
- British Indian Ocean Territory
- British Virgin Islands
- Ecuador
- El Salvador
- Guam
- Haiti
- Marshall Islands (the)
- Micronesia (Federated States of)
- Northern Mariana Islands
- Palau
- Panama
- Puerto Rico
- Timor-Leste
- Turks and Caicos Islands
- United States of America
- United States Virgin Islands
- US Minor Is.

minor unit exponent: 2

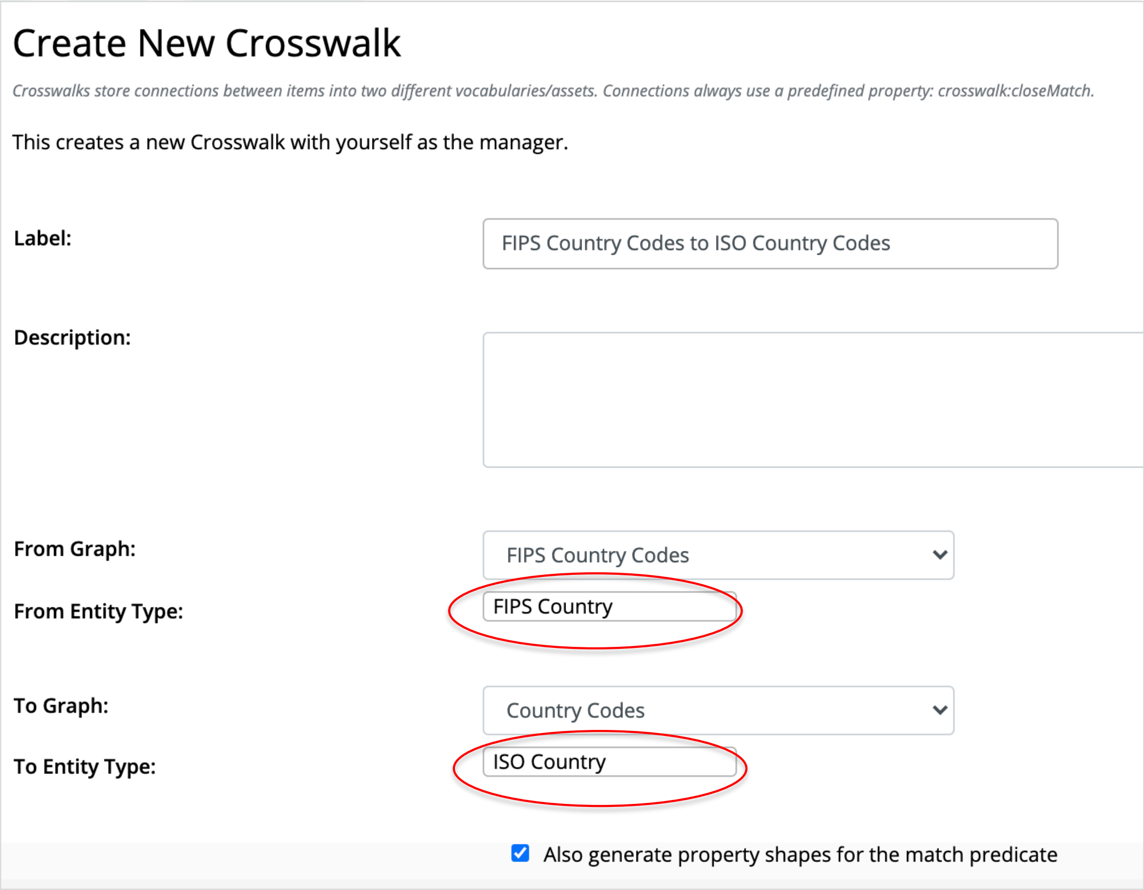


Issuing Authority

issuing country: United States of America



Creating Crosswalks



Create New Crosswalk

Crosswalks store connections between items into two different vocabularies/assets. Connections always use a predefined property: crosswalk:closeMatch.

This creates a new Crosswalk with yourself as the manager.

Label: FIPS Country Codes to ISO Country Codes

Description:

From Graph: FIPS Country Codes

From Entity Type: FIPS Country

To Graph: Country Codes

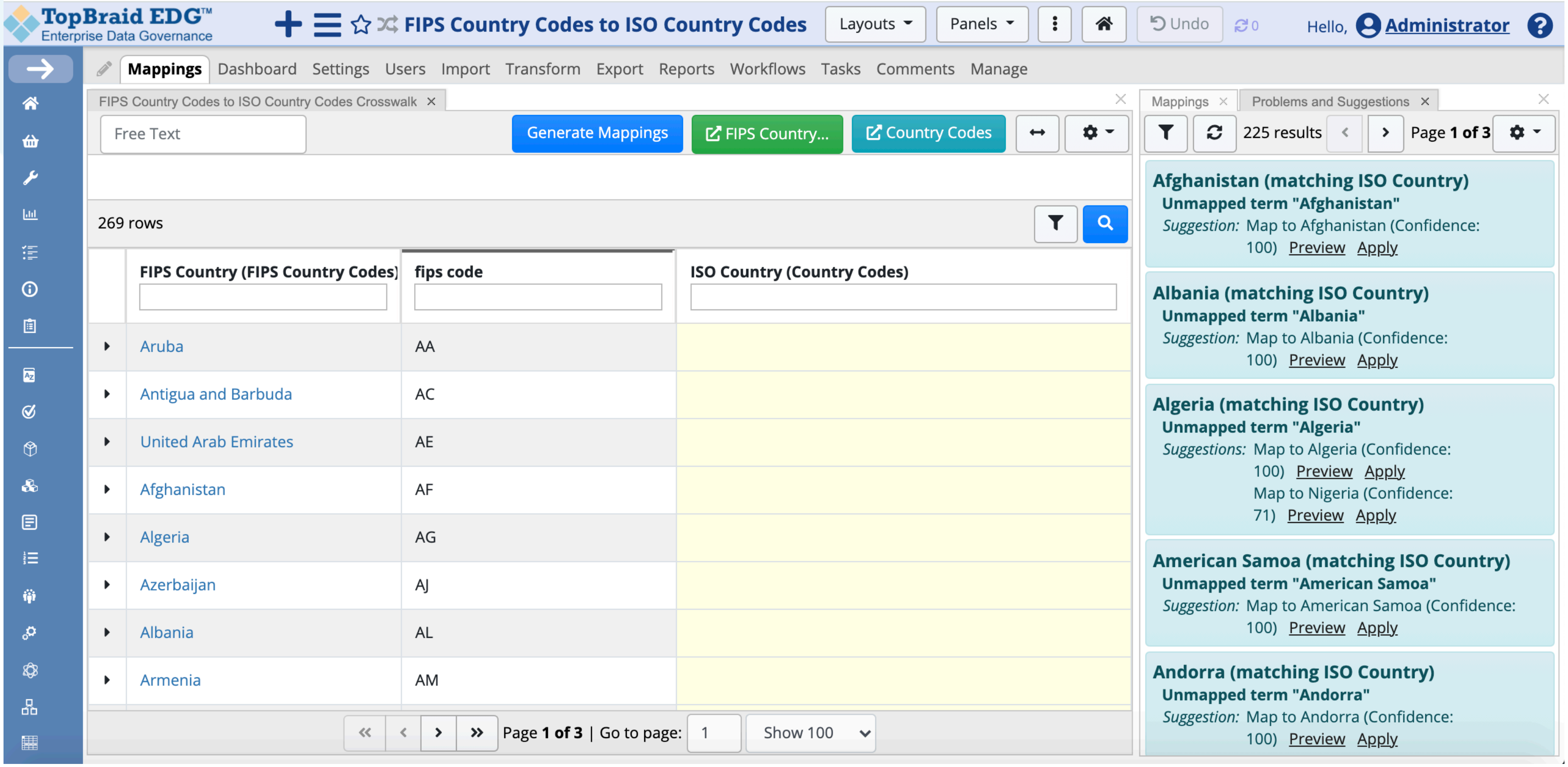
To Entity Type: ISO Country

☑ Also generate property shapes for the match predicate

- Select source and target data
- Specify source and target classes. Only instances of these classes will be candidates for matching

TopQuadrant™

Generating matches



269 rows

	FIPS Country (FIPS Country Codes)	fips code	ISO Country (Country Codes)
▸	Aruba	AA	
▸	Antigua and Barbuda	AC	
▸	United Arab Emirates	AE	
▸	Afghanistan	AF	
▸	Algeria	AG	
▸	Azerbaijan	AJ	
▸	Albania	AL	
▸	Armenia	AM	

Page **1 of 3** | Go to page: 1 Show 100

Afghanistan (matching ISO Country)
Unmapped term "Afghanistan"
Suggestion: Map to Afghanistan (Confidence: 100) Preview Apply

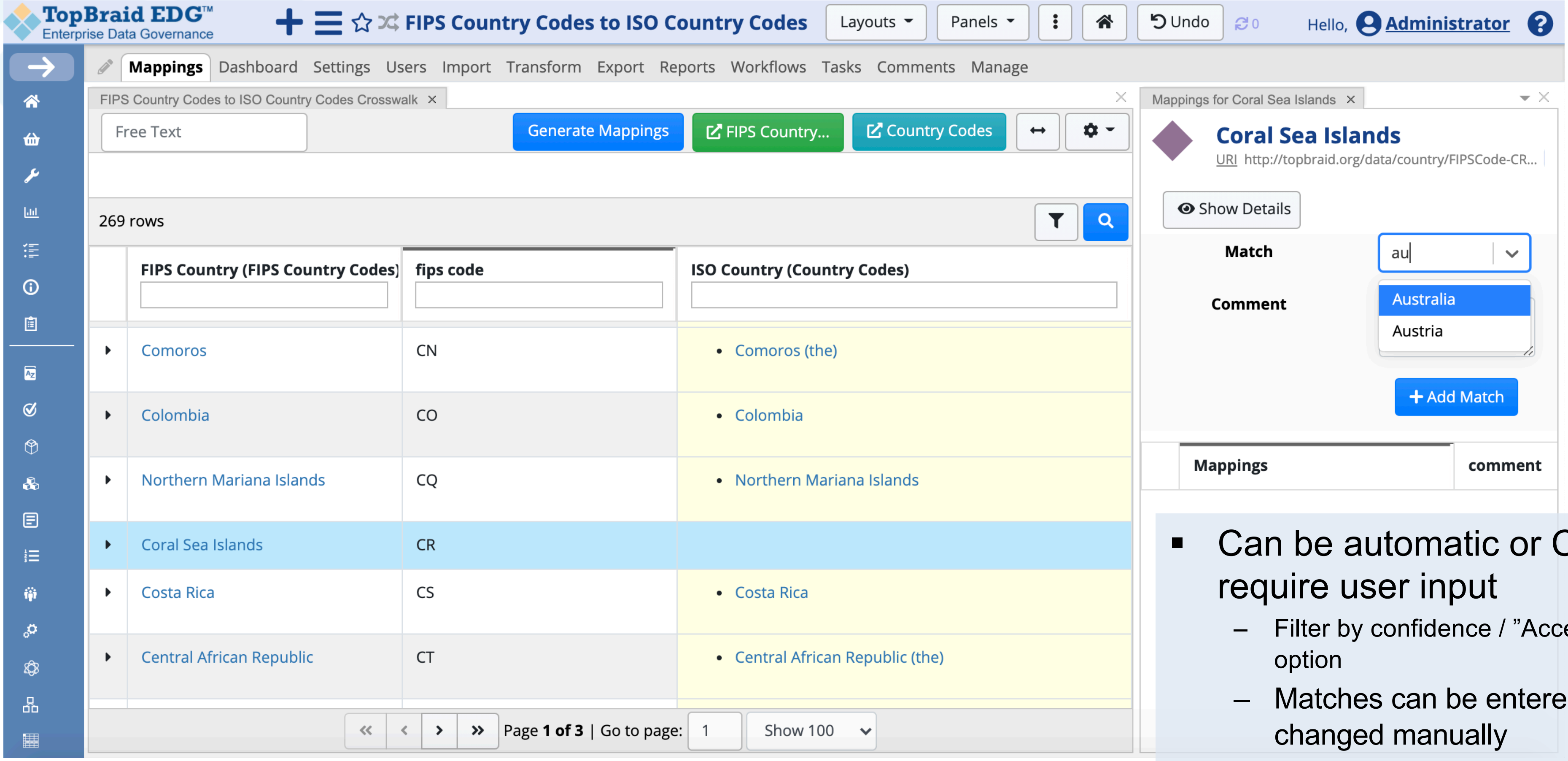
Albania (matching ISO Country)
Unmapped term "Albania"
Suggestion: Map to Albania (Confidence: 100) Preview Apply

Algeria (matching ISO Country)
Unmapped term "Algeria"
Suggestions: Map to Algeria (Confidence: 100) Preview Apply
Map to Nigeria (Confidence: 71) Preview Apply

American Samoa (matching ISO Country)
Unmapped term "American Samoa"
Suggestion: Map to American Samoa (Confidence: 100) Preview Apply

Andorra (matching ISO Country)
Unmapped term "Andorra"
Suggestion: Map to Andorra (Confidence: 100) Preview Apply

Accepting Matches



269 rows

FIPS Country (FIPS Country Codes)	fips code	ISO Country (Country Codes)
Comoros	CN	Comoros (the)
Colombia	CO	Colombia
Northern Mariana Islands	CQ	Northern Mariana Islands
Coral Sea Islands	CR	
Costa Rica	CS	Costa Rica
Central African Republic	CT	Central African Republic (the)

Page **1 of 3** | Go to page: 1 Show 100

Coral Sea Islands

URI http://topbraid.org/data/country/FIPSCode-CR...

Match au

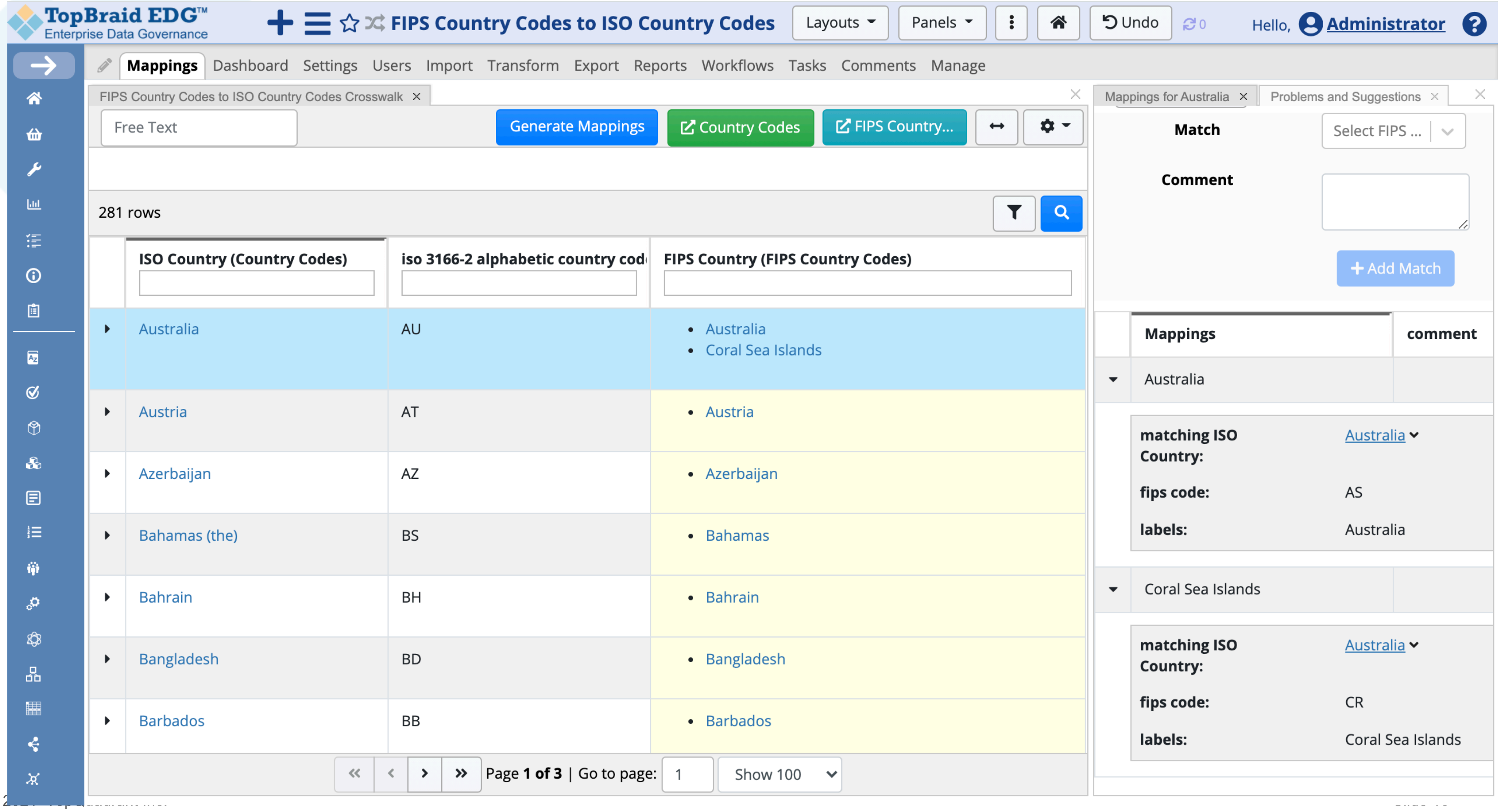
Australia

Austria

Comment

- Can be automatic or Can require user input
  - Filter by confidence / ”Accept all” option
  - Matches can be entered / changed manually

Many to Many Mappings



281 rows

ISO Country (Country Codes)	iso 3166-2 alphabetic country cod	FIPS Country (FIPS Country Codes)
Australia	AU	• Australia • Coral Sea Islands
Austria	AT	• Austria
Azerbaijan	AZ	• Azerbaijan
Bahamas (the)	BS	• Bahamas
Bahrain	BH	• Bahrain
Bangladesh	BD	• Bangladesh
Barbados	BB	• Barbados

Page **1 of 3** | Go to page: 1 Show 100

Match

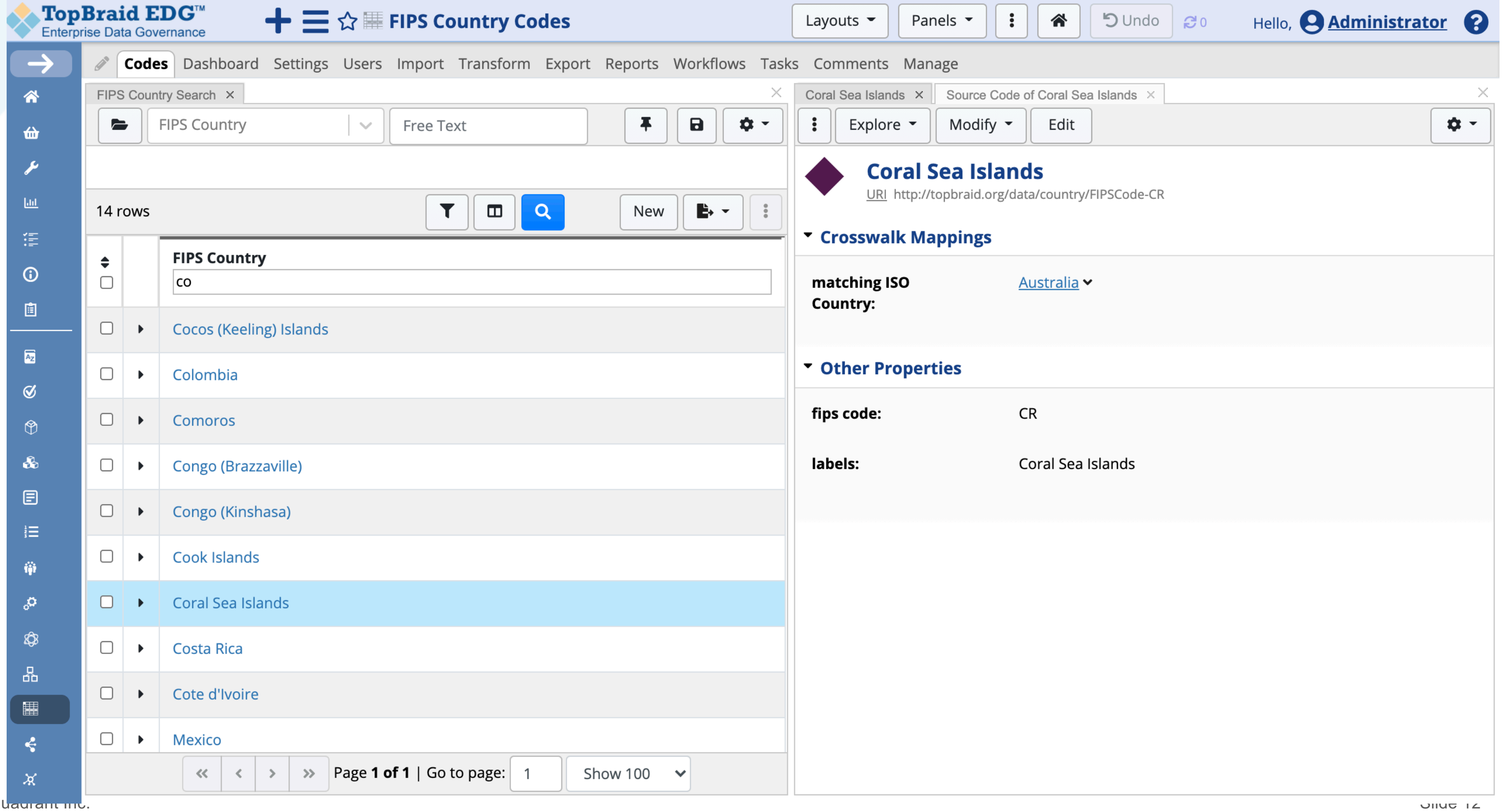
Comment

Mappings	comment
Australia	
matching ISO Country: Australia **fips code:** AS **labels:** Australia	
Coral Sea Islands	
matching ISO Country: Australia **fips code:** CR **labels:** Coral Sea Islands	

Determining Candidate Matches

- Based on the label
  - rdfs:label, skos:prefLabel, skos:altLabel
  - Users can specify custom labels
  - Alternative labels (skos:altLabel) are given lower confidence
- Uses a Lucene index
  - Create an index holding labels of all instances of target class
- Query the index for matches for each instance of the source class
  - By default we do a fuzzy query on the labels
- Customers can customise/extend matching behaviour

Accessing Crosswalk Information in a Reference Dataset



14 rows

FIPS Country
co
Cocos (Keeling) Islands
Colombia
Comoros
Congo (Brazzaville)
Congo (Kinshasa)
Cook Islands
Coral Sea Islands
Costa Rica
Cote d'Ivoire
Mexico

Coral Sea Islands

URI http://topbraid.org/data/country/FIPSCode-CR

Crosswalk Mappings

matching ISO Country:	Australia

Other Properties

fips code:	CR
labels:	Coral Sea Islands

How Well Does Automated Mapping Perform in Real Life Scenarios?

- For a pharma customer, we analyzed performance of automated mapping from a customer specific terminology used for clinical trials to SNOMED and MESH. Calculating:
  - % found mappings from the local terminology to SNOMED and MESH i.e., recall
  - % of correct matches vs Expected results i.e., precision
  - Mappings for 2 columns

Correct match % Calculation

- We considered a correct match one that presents expected mapping, regardless of the confidence.
  - For example, if we returned 3 mappings from the same input value to different concepts (at different confidence levels), as long as one of them matches the expected term, the match is considered *correct*.

How Well Does Automated Mapping Perform in Real Life Scenarios?

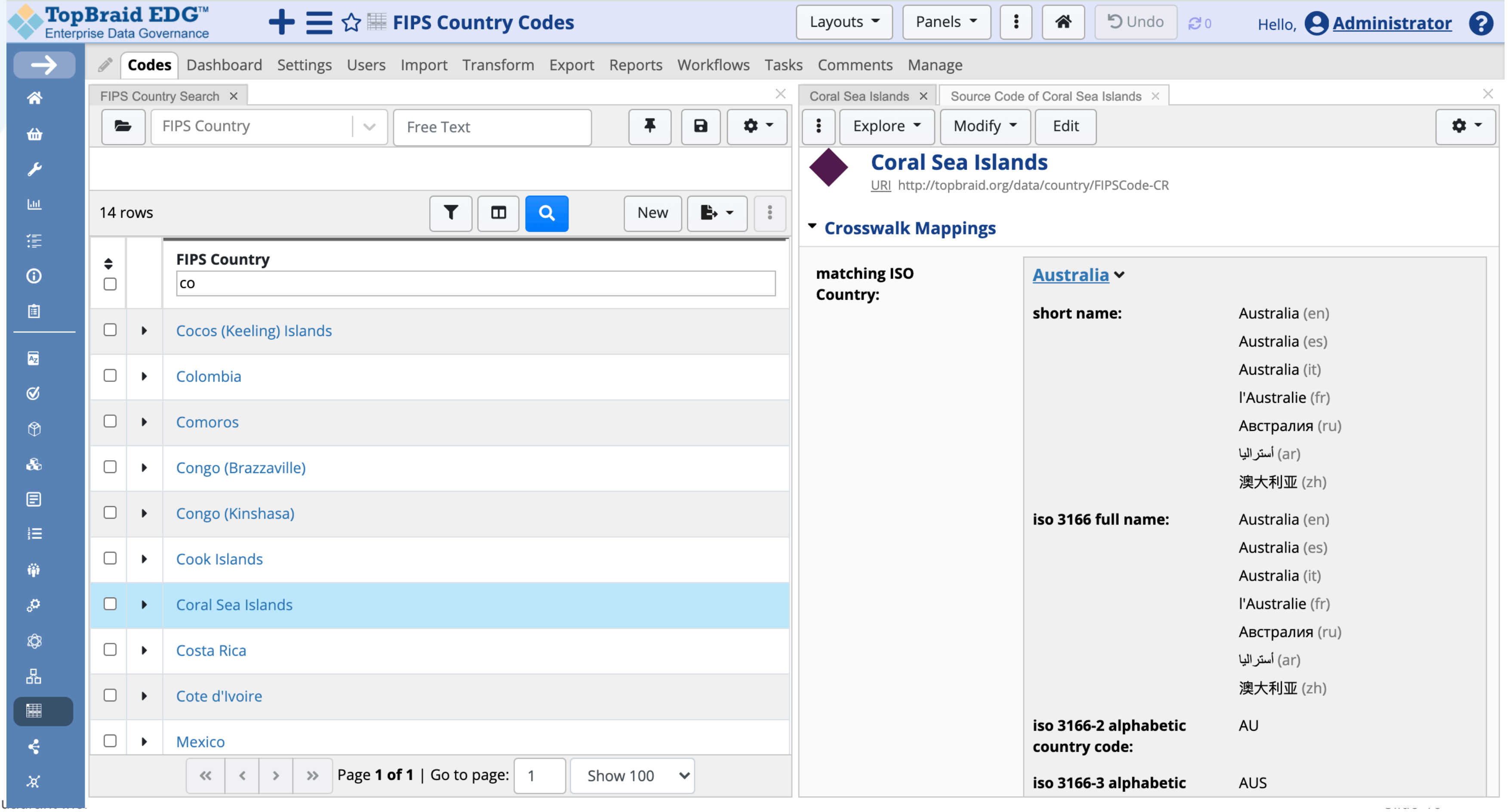
Matches Found

Local Terminology	Standard Vocabulary	Matches found (%)
Col1 - 105 unique cell values	MESH	102 out of 105 (97%)
Col2 - 150 unique cell values	SNOMED	150 out of 150 (100%)
Col2 - 150 unique cell values	MESH	150 out of 150 (100%)

Correct matches (expected results)

Local Terminology	Standard Vocabulary	Correct Matches found (%)
Col1	MESH - 132 expected matches	116 out of 132 (88%)
Col2	SNOMED - 69 expected matches	62 out of 69 (90%)
Col2	MESH - 147 expected matches	116 out of 147 (79%)

Accessing Crosswalk Information in a Reference Dataset



14 rows

		FIPS Country co
□	▸	Cocos (Keeling) Islands
□	▸	Colombia
□	▸	Comoros
□	▸	Congo (Brazzaville)
□	▸	Congo (Kinshasa)
□	▸	Cook Islands
□	▸	Coral Sea Islands
□	▸	Costa Rica
□	▸	Cote d'Ivoire
□	▸	Mexico

Page **1 of 1** | Go to page: 1 Show 100

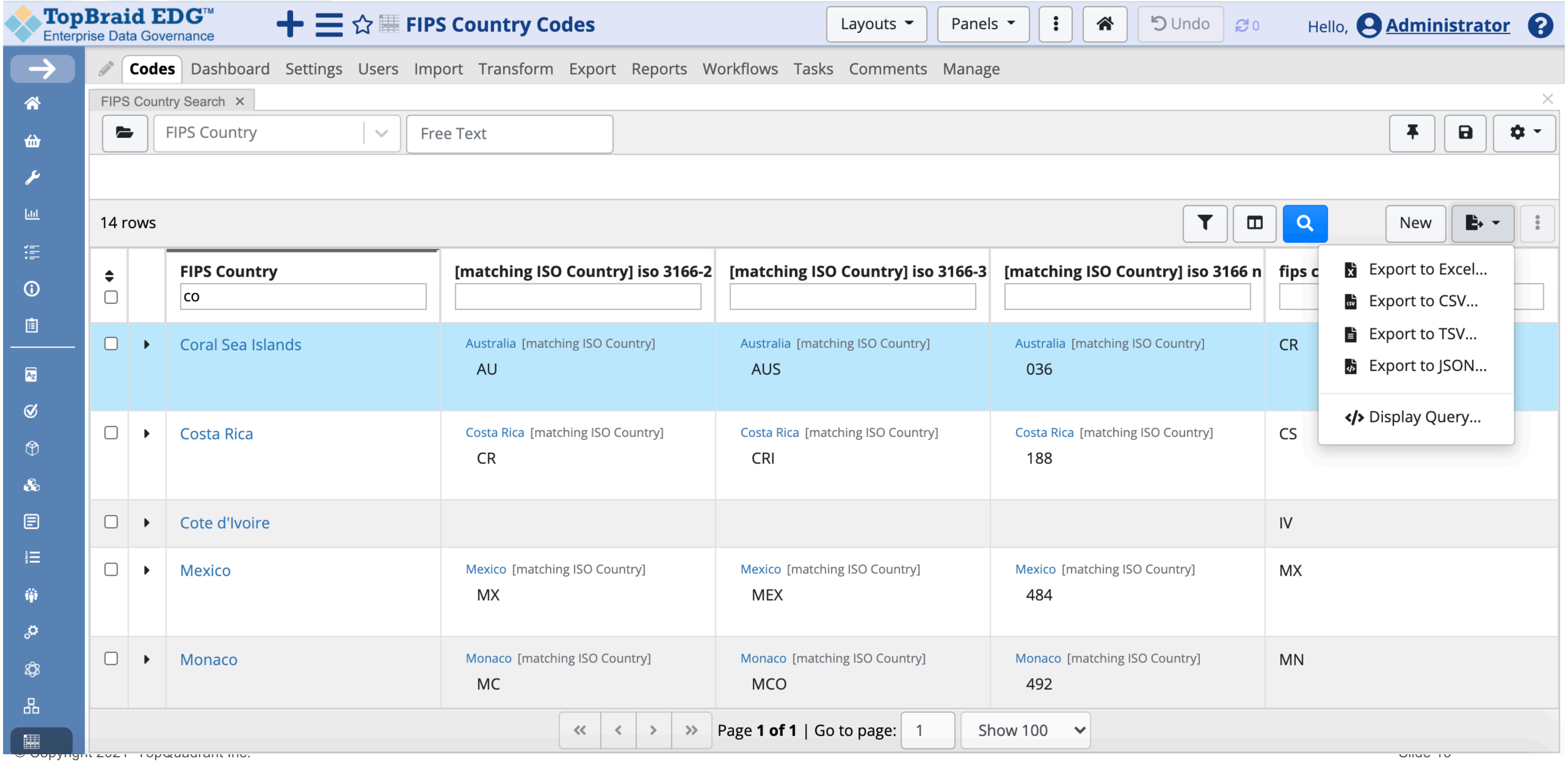
Coral Sea Islands
URI http://topbraid.org/data/country/FIPSCode-CR

▾ **Crosswalk Mappings**

matching ISO Country: Australia

short name:	Australia (en) Australia (es) Australia (it) l'Australie (fr) Австралия (ru) أستراليا (ar) 澳大利亚 (zh)
iso 3166 full name:	Australia (en) Australia (es) Australia (it) l'Australie (fr) Австралия (ru) أستراليا (ar) 澳大利亚 (zh)
iso 3166-2 alphabetic country code:	AU
iso 3166-3 alphabetic	AUS

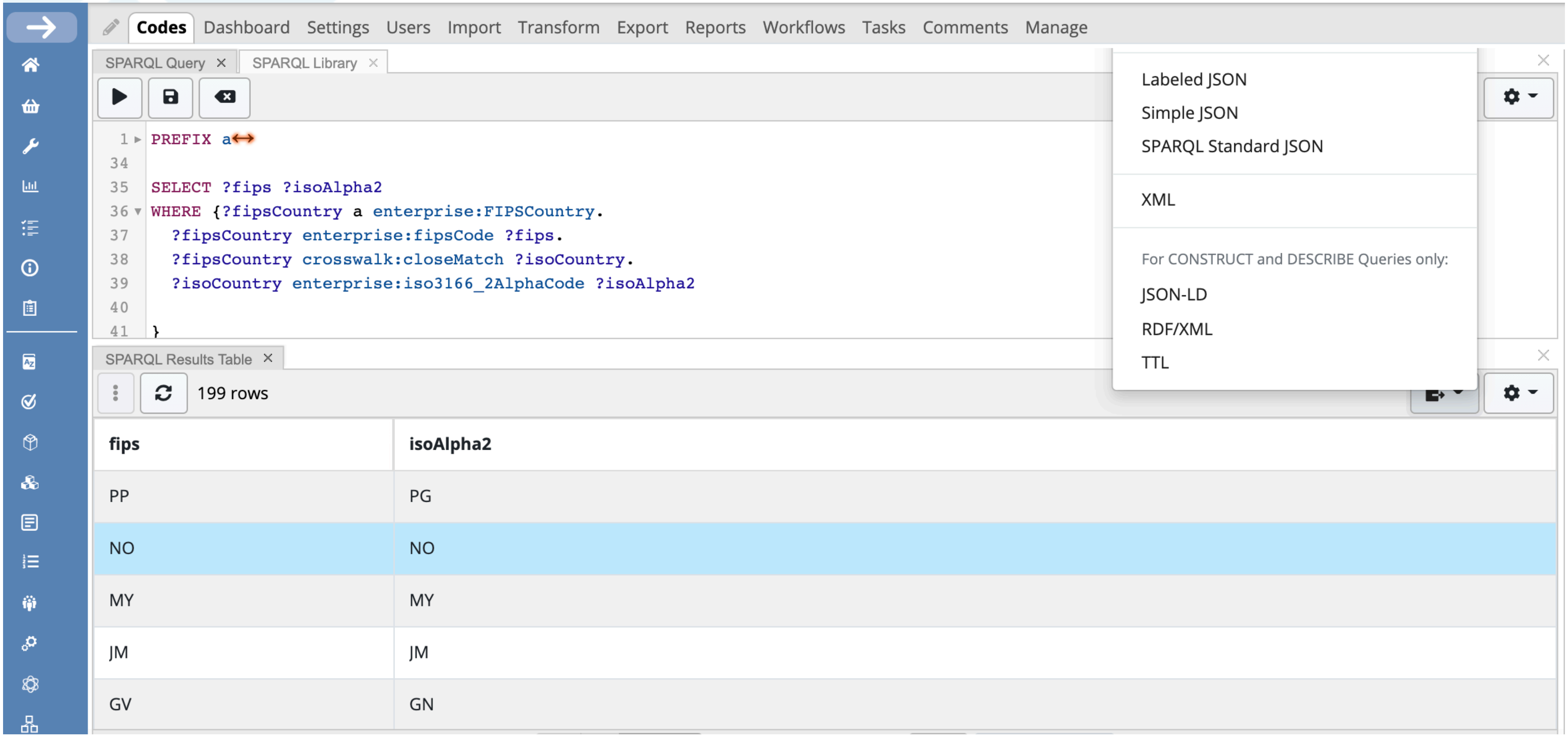
Exporting matches



14 rows

FIPS Country	[matching ISO Country] iso 3166-2	[matching ISO Country] iso 3166-3	[matching ISO Country] iso 3166 n	fips c
co				
Coral Sea Islands	Australia [matching ISO Country] AU	Australia [matching ISO Country] AUS	Australia [matching ISO Country] 036	CR
Costa Rica	Costa Rica [matching ISO Country] CR	Costa Rica [matching ISO Country] CRI	Costa Rica [matching ISO Country] 188	CS
Cote d'Ivoire				IV
Mexico	Mexico [matching ISO Country] MX	Mexico [matching ISO Country] MEX	Mexico [matching ISO Country] 484	MX
Monaco	Monaco [matching ISO Country] MC	Monaco [matching ISO Country] MCO	Monaco [matching ISO Country] 492	MN

Querying and Exporting via SPARQL

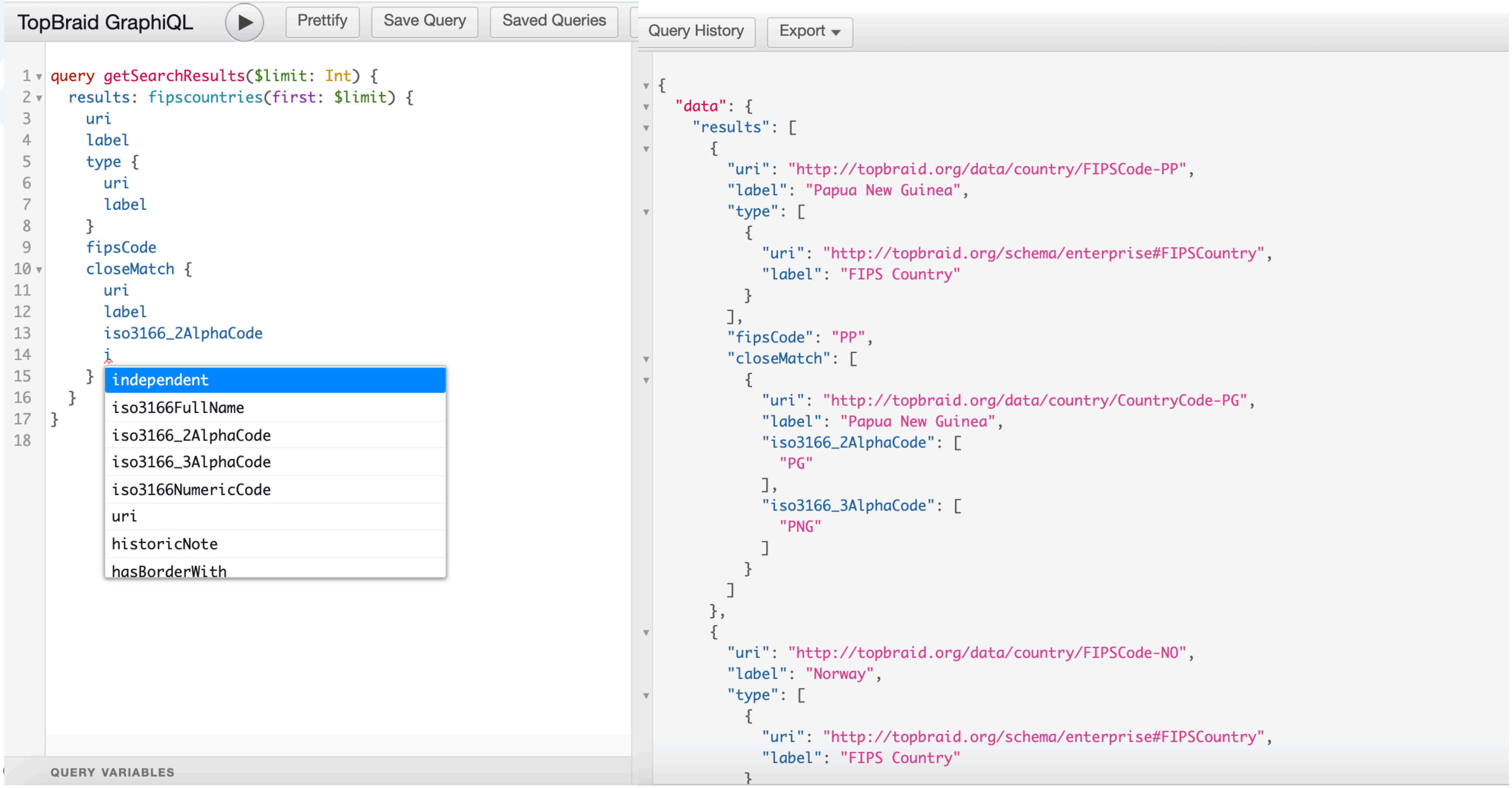


```
PREFIX a↔

SELECT ?fips ?isoAlpha2
WHERE {?fipsCountry a enterprise:FIPSCountry.
  ?fipsCountry enterprise:fipsCode ?fips.
  ?fipsCountry crosswalk:closeMatch ?isoCountry.
  ?isoCountry enterprise:iso3166_2AlphaCode ?isoAlpha2

}
```

199 rows

fips	isoAlpha2
PP	PG
NO	NO
MY	MY
JM	JM
GV	GN

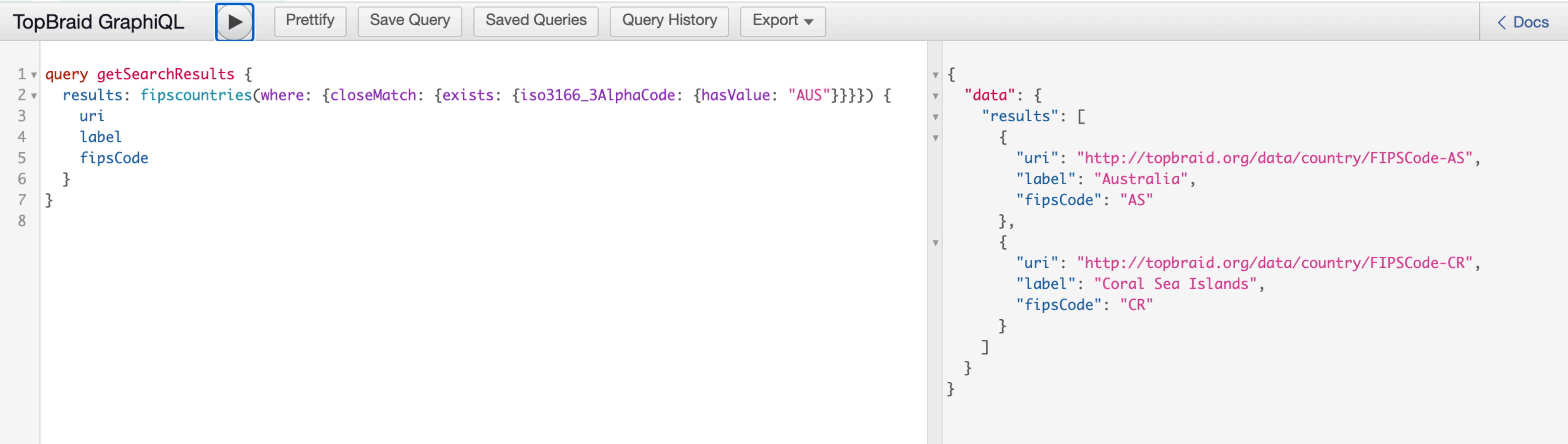
Querying and Exporting via GraphQL



```
query getSearchResults($limit: Int) {
  results: fipscountries(first: $limit) {
    uri
    label
    type {
      uri
      label
    }
    fipsCode
    closeMatch {
      uri
      label
      iso3166_2AlphaCode
      i
    }
  }
}
```

```
{
  "data": {
    "results": [
      {
        "uri": "http://topbraid.org/data/country/FIPSCode-PP",
        "label": "Papua New Guinea",
        "type": [
          {
            "uri": "http://topbraid.org/schema/enterprise#FIPSCountry",
            "label": "FIPS Country"
          }
        ],
        "fipsCode": "PP",
        "closeMatch": [
          {
            "uri": "http://topbraid.org/data/country/CountryCode-PG",
            "label": "Papua New Guinea",
            "iso3166_2AlphaCode": [
              "PG"
            ],
            "iso3166_3AlphaCode": [
              "PNG"
            ]
          }
        ]
      },
      {
        "uri": "http://topbraid.org/data/country/FIPSCode-NO",
        "label": "Norway",
        "type": [
          {
            "uri": "http://topbraid.org/schema/enterprise#FIPSCountry",
            "label": "FIPS Country"
          }
```

Crosswalk Services



```
query getSearchResults {
  results: fipscountries(where: {closeMatch: {exists: {iso3166_3AlphaCode: {hasValue: "AUS"}}}}) {
    uri
    label
    fipsCode
  }
}
```

```
{
  "data": {
    "results": [
      {
        "uri": "http://topbraid.org/data/country/FIPSCode-AS",
        "label": "Australia",
        "fipsCode": "AS"
      },
      {
        "uri": "http://topbraid.org/data/country/FIPSCode-CR",
        "label": "Coral Sea Islands",
        "fipsCode": "CR"
      }
    ]
  }
}
```

What is mapped to a reference data item with the ISO 3-character code = AUS?

TopQuadrant™

Thank You!

ENDORSE 16-19 MARCH 2021
THE EUROPEAN DATA CONFERENCE ON REFERENCE DATA AND SEMANTICS

- Irene Polikoff

  irene at topquadrant.com

- TopQuadrant

  https://www.topquadrant.com/

  Request demo: sales at topquadrant.com

  Request evaluation account: https://www.topquadrant.com/products/topbraid-enterprise-data-governance/request-edg-evaluation-account/

- Enterprise Data Governance (EDG)

  https://www.topquadrant.com/products/topbraid-enterprise-data-governance/