Linked SDMX Data

Sarven Capadisli
Universität Leipzig, Institut für Informatik, AKSW
Postfach 100920, D-04009 Leipzig, Germany
info@csarven.ca

Sören Auer
Universität Leipzig, Institut für Informatik, AKSW
Postfach 100920, D-04009 Leipzig, Germany
auer@informatik.uni-leipzig.de

Axel-Cyrille Ngonga Ngomo
Universität Leipzig, Institut für Informatik, AKSW
Postfach 100920, D-04009 Leipzig, Germany
ngonga@informatik.uni-leipzig.de

Purpose

Path to high fidelity Statistical Linked Data

ABSTRACT

As statistical data is inherently highly structured and comes with rich metadata (in form of code lists, data cubes etc.), it would be a missed opportunity to not tap into it from the Linked Data angle. At the time of this writing, there exists no simple way to transform statistical data into Linked Data since the raw data comes in different shapes and forms. Given that SDMX (Statistical Data and Metadata eXchange) is arguably the most widely used standard for statistical data exchange, a great amount of statistical data about our societies is yet to be discoverable and identifiable in a uniform way. In this article, we present the design and implementation of SDMX-ML to RDF/XML XSL transformations, as well as the publication of OECD, BFS, FAO, ECB, IMF, UIS, and FRB dataspaces with that tooling.

Categories And Subject Descriptors

H.4 [**Information Systems Applications**]: Linked Data; D.2 [**Software Engineering**]: Semantic Web

Keywords

Linked Data, SDMX, Statistics, Data modeling, Data transformation, Dataspaces, Knowledge management

INTRODUCTION

While access to statistical data in the public sector has increased in recent years, a range of technical challenges makes it difficult for data consumers to tap into this data at ease. These are particularly related to the following two areas:

- Automation of data transformation of data from high profile statistical organizations.
- Minimization of third-party interpretation of the source data and metadata and lossless transformations.

Development teams often face low-level repetitive data management tasks to deal with someone else's data. Within the context of Linked Data, one aspect is to transform this raw statistical data (e.g., SDMX-ML) into an RDF representation in order to be able to start tapping into what's out there in a uniform way.

The contributions of this article are two-fold. We present an approach for transforming SDMX-ML based on XSLT 2.0 templates and showcase our implementation which transforms SDMX-ML data to RDF/XML. Following this, SDMX-ML data from OECD (Organisation for Economic Co-operation and Development), BFS (Bundesamt für Statistik@de, Swiss Federal Statistical Office@en), FAO (Food and Agriculture Organization of the United Nations), ECB (European Central Bank), and IMF (International Monetary Fund), UIS (UNESCO Institute for Statistics), FRB (Federal Reserve Board) are retrieved, transformed and published as Linked Data.

1. BACKGROUND

As pointed out in *Statistical Linked Dataspaces* (Capadisli, S., 2012), what linked statistics provide, and in fact enable, are queries across datasets: Given that the dimension concepts are interlinked, one can learn from a certain observation's dimension value, and enable the automation of cross-dataset queries.

Moreover, a number of approaches have been undertaken in the past to go from raw statistical data from the publisher to linked statistical data, as discussed in great detail in *Official statistics and the Practice of Data Fidelity* (Cyganiak, R., 2011). These approaches go from retrieval of the data by majority; in tabular formats: Microsoft Excel or CSV, tree formats: XML with a custom schema, SDMX-ML, PC-Axis, to transformation into different RDF serialization formats. As far as graph formats go, majority of datasets in those formats not published by the owners. However, there are number of statistical linked dataspaces in the LOD Cloud already.

A number of transformation efforts are performed by the Linked Data community based on various formats. For example, the World Bank Linked Dataspace is based on custom XML that the World Bank provides through their APIs with the application of XSL Templates. The Transparency International Linked Dataspace's data is based on CSV files with the transformation step through Google Refine and the RDF Extension. That is, data sources provide different data formats for the public, with or without accompanying metadata e.g., vocabularies, provenance. Hence, this repetitive work is no exception to Linked Data teams as they have to constantly be involved either by way of hand-held transformation efforts, or in best-case scenarios, it is done semi-automatically. Currently, there is no automation of the transformation step to the best of our knowledge. This is generally due to the difficulty of the task when dealing with the quality and consistency of the statistical data that is published on the Web, as well as the data formats that are typically focused on consumption. Although SDMX-ML is the primary format of the high profile statistical data organizations, it is yet to be taken advantage of.

2. SDMX-ML TO LINKED DATA

Recently, SDMX is approved by ISO as International Standard. It is a standard which provides the possibility to consistently carry out data flows between publishers and consumers. SDMX-ML (using XML syntax) is considered to be the gold standard for expressing statistical data. It has a highly structured mechanism to represent statistical observations, classifications, and data structures. Organizations behind SDMX are BIS (Bank for International Settlements), OECD, UN (United Nations), ECB, World Bank, IMF, FAO, and Eurostat.

We argue that high-fidelity statistical data representation in Linked Data should take advantage of SDMX-ML as it is widely adopted by data producers with rich data about our societies, making the need for transforming SDMX-ML to RDF and publishing accompanying Linked Dataspaces of paramount importance

2.1 Data Sources

As a demonstration of the SDMX-ML to RDF transformations, the selection of datasets here are from the following organizations. Instead of regurgitating publisher's profile about themselves, and to keep this part brief, I'll quote the first sentences from their about pages:

OECD

> “The mission of the Organisation for Economic Co-operation and Development (OECD) is to promote policies that will improve the economic and social well-being of people around the world” from OECD Our mission.

BFS

> “Swiss Statistics, the Federal Statistical Office’s web portal. Offering a modern and appealing interface, our website proposes a wide range of statistical information on the most important areas of life: population, health, economy, employment, education and much more” from BFS Welcome.

FAO

> “Achieving food security for all is at the heart of FAO's efforts - to make sure people have regular access to enough high-quality

food to lead active, healthy lives" from FAO's mandate.

ECB

> "Whose main task is to maintain the euro's purchasing power and thus price stability in the euro area" from ECB home.

IMF

> "Working to foster global monetary cooperation, secure financial stability, facilitate international trade, promote high employment and sustainable economic growth, and reduce poverty around the world" from IMF home.

UIS

> "The primary source for cross-nationally comparable statistics on education, science and technology, culture, and communication for more than 200 countries and territories" from UIS home.

FRB

> "The Federal Reserve, the central bank of the United States, provides the nation with a safe, flexible, and stable monetary and financial system" from FRB home.

The OECD, FAO, ECB, IMF, UIS datasets consisted of observational and structural data. The OECD, ECB, and UIS provided complete coverage (to the best of our knowledge), whereas FAO had partial fishery related data, and IMF partial data over their REST service. BFS had all of their classifications available, with no observational data in SDMX-ML.

2.2 Data Retrieval

As SDMX-ML publishers have their own publishing processes, availability and accessibility of the data varied. After obtaining the dataset codes, names, and URLs with common Linux command-line work, a Bash script was created to retrieve the data.

2.2.1 OECD

On how to retrieve all of the datasets from the OECD website was not entirely clear. In order to automatically get a hold of list of datasets, I copied the innerHTML of the DOM tree that contained all the dataset codes from OECD.StatExtracts to a temporary file. This was done due to the fact that a simple scrape of an HTTP GET wasn't possible as the data on the page was populated via JavaScript on `document` ready. After constructing a list of datasets and structures to get, two REST API endpoints were called.

2.2.2 BFS

BFS offered a Microsoft Excel document which contained a catalog of their classifications and URLs for retrieval.

2.2.3 FAO

After searching for keywords along the lines of `SDMX site:fao.org` at a search-engine nearby, FAO Fisheries and data.fao.org SDMX Registry and Repository, and its children pages were marked for SDMX-ML retrieval.

2.2.4 ECB

ECB had a similar REST API to OECD. Additionally, SDMX Dataflows was retrieved to get a primary list of datasets to retrieve. Some of the large datasets was retrieved by making multiple smaller calls to the API using a call per refereance area.

2.2.5 IMF

Same procedure as ECB and OECD.

2.2.6 UIS

Same procedure as ECB, OECD, and IMF.

2.3 Provenance

2.3.1 Provenance At Retrieval

At the time of data retrieval, information pertaining to provenance was captured using the PROV Ontology in order to further enrich the data. This RDF/XML document contains `prov:Activity` information which indicates the location of the XML document on the local filesystem. It contains other provenance data like when it was retrieved, with what tools, etc. This provenance data from retrieval may be provided to the XSL Transformer during the transformation phase and VoID enrichment.

2.3.2 Provenance At Transformation

Resources of type `qb:DataStructureDefinition`, `qb:DataSet`, `skos:ConceptScheme` are also typed with the `prov:Entity` class. Also properties `prov:wasAttributedTo` were added to these resources with the `creator` value which is of type `prov:Agent` obtained from XSLT configuration. There is a unique `prov:Activity` for each transformation, and it has a `dcterms:title`, and contains values for `prov:startedAtTime`, `prov:wasAssociatedWith` (the creator), `prov:used` (i.e., source XML, XSL to transform) to what was `prov:generated` (and source data URI that it `prov:wasDerivedFrom`). It also declares `dcterms:license` where value taken from XSLT configuration. The provenance document from the retrieval phase may be provided to the transformer. In this case, it establishes a link between the current provenance activity (i.e., the transformation), with the earlier provenance activity (i.e., the retrieval) using the `prov:wasInformedBy` property.

2.3.3 Provenance At Post-processing

The post-processing step for provenance is intended to retain provenance data for future use. As datasets get updated, it is important to preserve information about past activities by way of exporting all instances of the `prov:Activity` class from the RDF store. Activities are unique artifacts, on a conceptual level as well as with regard to referencing them. Since one of the main concerns of provenance is to keep track of activities, this post-processing step also allows us to retain a historical account of all activities during the data lifecycle, and to preserve all previously published URIs (cf. Cool URIs don't change).

2.4 Data Preprocessing

By in large, there was no need to pre-process the data as the transformation dealt with the data as it was. However, some non-vital SDMX components were omitted from the output. For instance, one type of attribute in OECD and ECB observations contained free-text as opposed to its corresponding code from a codelist. Since the RDF Data Cube required codes as opposed to free-text for dimension values, some attributes were excluded. The decision here was to trade-off some precision in favour of retaining the dataset.

2.5 Data Modeling

This section goves over several areas which are at the heart of representing statistical data in SDMX-ML as Linked Data. The approach taken was to provide a level of consistency for data consumers and tool builders for all statistical Linked Data with its origins from data in SDMX-ML.

2.5.1 Vocabularies

Besides the common vocabularies: RDF RDFS, XSD, OWL, XSD, the RDF Data Cube vocabulary is used to describe multi-dimensional statistical data, and SDMX-RDF for the statistical information model. PROV-O is used for provenance coverage. SKOS and XKOS to cover concepts, concept schemes and their relationships to one another.

2.5.2 Versioning

SDMX data publishers version their classifications and the generated cubes refer to particular versions of those classifications. Consequently, versions need to be explicitly part of classification URIs in order to uniquely identify them. Although including version information in the URI is disputed by some authors, it is a good exception for identifying different concepts and data structures. Jeni Tennison et al discussed Versioning URIs, and concluded that there was no one-size-fits all solution. An alternative approach using named graphs for a series of changes was proposed in Linking UK Government Data.

2.5.3 URI Patterns

An outline for the URI patterns is given in table below: `authority` is replaced with the domain (see also: Agency identifiers and URIs) followed with `class`, `code`, `concept`, `dataset`, `property`, `provenance`, or `slice` as example. These tokens as well as `/` which is used to separate the dimension concepts in URIs can be configured in the toolkit.

In order to construct the URIs for the above patterns, some of the data values are normalized to make them URI safe but not altered in other ways (e.g., lower-casing). The rationale for this was to keep the consistency of terms in SDMX and RDF.

URI Patterns

Entity type	URI Pattern
qb:DataStructureDefinition	http://{authority}/structure /{KeyFamilyID}
qb:ComponentSpecification	http://{authority}/component /{KeyFamilyID} /{dimension|measure|attribute} /{version}/{conceptSchemeID} /{conceptID}
qb:DataSet	http://{authority}/dataset/{datasetID}
qb:Observation	http://{authority}/dataset/{datasetID} /{dimension-1}/../{dimension-n}
qb:Slice	http://{authority}/slice/{KeyFamilyID} /{dimension-1}/../{dimension-n-1}/{slice}
skos:Collection	http://{authority}/code/{version} /{hierarchicalCodeListID}, http://{authority}/code/{version} /{hierarchyID}
sdmx:CodeList	http://{authority}/code/{version} /{codeListID}
skos:ConceptScheme	http://{authority}/concept/{version} /{conceptSchemeID}
skos:Concept, sdmx:Concept	http://{authority}/code/{version} /{codeListID}/{codeID} http://{authority}/concept/{version} /{conceptSchemeID}/{conceptID}
owl:Class, rdfs:Class	http://{authority}/class/{version} /{codeListID}
rdf:Property, qb:ComponentProperty	http://{authority} /{property|dimension|measure|attribute} /{version}/{conceptSchemeID} /{conceptID}
qb:DimensionProperty	http://{authority}/dimension/{version} /{conceptSchemeID}/{conceptID}
qb:MeasureProperty	http://{authority}/measure/{version} /{conceptSchemeID}/{conceptID}
qb:AttributeProperty	http://{authority}/attribute/{version} /{conceptSchemeID}/{conceptID}

2.5.4 Datatypes

XSD datatypes are assigned to literals are based on the value of the measure component (e.g., decimal, year). In the absence of this datatype, observation values are checked whether they can be casted to `xsd:decimal`. Otherwise, they are left as plain literals.

3. LINKED SDMX DATA TRANSFORMATION

The Linked SDMX XSLT 2.0 templates and scripts are developed to transform SDMX-ML data and metadata to RDF/XML. Its goals are:

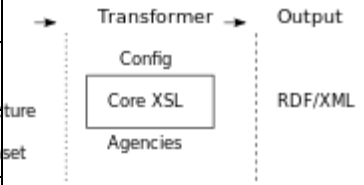
- To improve access and discovery of cross-domain statistical data.
- To perform the transformation in a lossless and semantics preserving way.
- To support and encourage statistical agencies to publish their data using RDF and integrating the transformation into their workflow.

The key advantage of this transformation approach is that additional interpretations are not required by the data modeler especially in comparison to alternative transformation (e.g., CSV or XML to RDF serialization). Since the SDMX-RDF vocabulary is based on SDMX-ML standard, and the RDF Data Cube vocabulary is closely aligned with the SDMX information model, the transformation is to a large extent a matter of mapping the source SDMX-ML data to its counter parts in RDF.

3.1 Features Of The Transformation

- Transforms SDMX KeyFamilies, ConceptSchemes and Concepts, CodeLists and Codes, Hierarchical CodeLists, DataSets.
- Configurability for SDMX publisher's needs.
- Detection and referencing CodeLists and Codes of external agencies.
- Support of interlinking publisher-specific annotation types.
- Support for omission of components.
- Inclusion of provenance data.

Figure : Transformation Process



3.2 What Is Inside?

It comes with scripts and sample data:

- XSLT 2.0 templates to transform Generic SDMX-ML data and metadata. It includes the main XSL template for generic SDMX-ML, an XSL for common templates and functions, and an RDF/XML configuration file to set preferences like base URIs, delimiters in URIs, how to map annotation types.
- Bash script that transforms sample data using saxonb-xslt.
- Sample SDMX Message and Structure retrieved from those organizations that are initially involved in the SDMX standard, as well as from BFS.

3.3 Requirements

The requirements for the Linked SDMX toolkit are an XSLT 2.0 processor to transform, and optionally to configure some of the settings in the transformation with provided `config.ttl` (in RDF Turtle) and transforming that to an abbreviated version of RDF/XML. In sequel some of they key features are described in more detail.

The transformation follows some common Linked Data practices as well as other ones out of thin air.

3.4 Configuration

The config file is used to pre-configure some of the decisions that are made in XSL templates. Here is an outline for some of the noteworthy things.

3.4.1 Agency Identifiers And URIs

`agencies.ttl` is used to track some of the mappings for maintenance agencies. It includes the maintenance agency's i.e., the SDMX publisher's, identifier that's in the SDMX Registry, as well as the base URI for that agency. This file allows references to external agency identifiers to be looked up for their base URI and used in the transformations. Currently this agency recognition is treated as either "SDMX" or some agency that's publishing the actual statistics.

In the case of SDMX, when there is a reference to SDMX CodeLists and Codes, it is typically indicated by the component agency being set to `SDMX` e.g., `codelistAgency="SDMX"` of a `structure:Component` and/or `agencyID="SDMX"` of a CodeList with `id="CL_FREQ"`. When this is detected, corresponding URIs from the SDMX-RDF vocabulary is used e.g., for metadata; `http://purl.org/linked-data/sdmx/2009/code#freq`, and data; `http://purl.org/linked-data/sdmx/2009/code#freq-A`.

Similarly, an agency might use some other agency's codes. By following the same URI pattern conventions, the agency file is used to find the corresponding base URI in order to make a reference. For example, here is a coded property that's used by European Central Bank (`ECB`) to associate a code list that's defined by Eurostat (`eurostat`):

```
<http://ecb.270a.info/property/OBS_STATUS>
 <http://purl.org/linked-data/cube#codeList>
  <http://eurostat.270a.info/code/1.0/CL_OBS_STATUS>
```

Naturally, the transformation does not re-define metadata that's from an external agency as the owners of the data would define

them under their authority.

3.4.2 URI Configurations

Base URIs can be set for classes, codelists, concept schemes, datasets, slices, properties, provenance, as well as for the source SDMX data.

The value for `uriThingSeparator` e.g., `/`, lets one set the delimiter to separate the "thing" from the rest of the URI. This is typically either a `/` or `#`. For example, if slash is used, an URI would end up like `http://{authority}/code/{version}/CL_GEO` (note the last slash before CL_GEO). If hash is used, an URI would end up like `http://{authority}/code/{version}#CL_GEO`.

Similarly, `uriDimensionSeparator` can be set to separate dimension values that's used in RDF Data Cube observation URIs. As observation should have its own unique URI, the method to construct URIs is done by taking dimension values as safe terms to be used in URIs separated by the value in `uriDimensionSeparator`. For example, here is a crazy looking observation URI where `uriDimensionSeparator` is set to `/`: `http://{authority}/dataset/HEALTH_STAT/EVIEFE00/EVIDUREV/AUS/1960`. But with `uriThingSeparator` set to `#` and `uriDimensionSeparator` set to `-`, it could end up like `http://{authority}/dataset/HEALTH_STAT#EVIEFE00-EVIDUREV-AUS-1960`. `HEALTH_STAT` is the dataset id.

Creator's URI can also be set which is also used for provenance data.

3.4.3 Default Language

From the configuration, it is possible to force a default `xml:lang` on `skos:prefLabel` and `skos:definition` when lang is not originally in the data. If `config.rdf`contains a non-empty lang value it will use it. Default language may also be applied in the case of Annotations. See Interlinking SDMX Annotations for example.

3.4.4 Interlinking SDMX Annotations

SDMX Annotations contain important information that can be put to use by the publisher. Data in AnnotationTypes are typically used as publisher's internal conventions. Hence, there is no standardization on how they are used across all SDMX publishers. In order not to leave this information behind in the final transformation, the configuration allows publishers to define the way they should be transformed. This is done by setting `interlinkAnnotationTypes`: the AnnotationType to detect (in `rdfs:type`), the predicate (as an XML QName) to use (in `rdf:predicate`), whether to apply instances of Concepts or Codes to apply to, or as Literals (in `rdf:range`), and whether to target AnnotationText or AnnotationTitle (in `rdfs:label`). Currently this feature is only applied to Annotations in Concepts and Codes. Only the AnnotationTypes with a corresponding configuration will be applied, and unspecific ones will be skipped.

3.4.5 Omitting Components

There are cases in which certain data parts contain errors. To get around this until the data is fixed at source, and without giving up on rest of the data at hand, as well as without making any significant assumptions or changes to the remaining data, `omitComponents` is a configuration option to explicitly skip over those parts. For example, if the Attribute values in a DataSet don't correspond to coded values - where they may contain whitespace - they can be skipped without damaging the rest of the data. This obviously gives up on precision in favour of still making use of the data.

4. LINKED DATASETS

This section provides information on the publication of OECD, BFS, FAO, and ECB datasets.

The original SDMX-ML files were transformed to RDF/XML using XSLT 2.0. Saxon's command-line XSLT and XQuery Processor tool was used for the transformations, and employed as part of Bash scripts to iterate through all the files in the datasets.

4.1 RDF Datasets

Here are some statistics for the transformations.

The command-line tool `saxonb-xslt` was used to conduct the XSL transformations. 12000M of memory was allocated on a machine with Intel(R) Xeon(R) CPU E5620 @ 2.40GHz. Linux kernel 3.2.0-33-generic was used. Table [Transformation time] provides information on datasets; input SDMX-ML size, output RDF/XML size, their size difference in ratio, and the total amount transformation time.

Transformation time

Dataset	Input size	Output size	Ratio	Time
OECD	3400 MB	24000 MB	1:7.1	131m25.795s
BFS	111 MB	154 MB	1:1.4	2m38.225s
FAO	902 MB	4400 MB	1:4.9	31m48.207s
ECB	11000 MB	35000 MB	1:3.2	316m46.329s
IMF	392 MB	3400 MB	1:8.7	28m11.826s
UIS	115 MB	896 MB	1:7.8	2m33.214s
FRB	783 MB	11000 MB	1:14.1	96m20.676s
Input size (rounded) refers to the original data in XML, and the output (rounded) is the RDF/XML size. Time is the real process time.				

Table [Transformed data] provides data on the transformed data; number of triples it contains, as well as the number of `qb:Observation`s, and the ratio.

Transformed data

Data	Number of triples	Number of observations	Ratio
OECD Dataset	305 million	30 million	10.2:1
OECD Metadata	1.15 million	N/A	N/A
BFS Metadata	1.5 million	N/A	N/A
FAO Dataset	53 million	7.2 million	7.4:1
FAO Metadata	0.37 million	N/A	N/A
ECB Dataset	469 million	26 million	18:1
ECB Metadata	0.47 million	N/A	N/A
IMF Dataset	36 million	3.3 million	10.9:1
IMF Metadata	0.05 million	N/A	N/A
UIS Dataset	10.4 million	1.4 million	7.4:1
UIS Metadata	0.09 million	N/A	N/A
FRB Dataset	135 million	9.8 million	13.8:1
FRB Metadata	0.04 million	N/A	N/A
Metadata (from `graph/meta`) includes dataset structures and classifications. Ratio refers to rounded ratio of total number of triples (rounded) to number of observations (rounded) in the dataset.			

Table [Resource counts] provides further statistics on prominent resources. It gives a contrast between the classifications and the dataset.

Resource counts

Resource	OECD	BFS	FAO	ECB	IMF	UIS	FRB
skos:ConceptScheme	1209	216	32	147	25	35	76
skos:Concept	80918	120202	28115	55609	3397	4515	2935
rdf:Property	129	0	12	231	42	12	64
qb:Observation	30183484	0	7186764	25791005	3603719	1437651	976829
Count of resources in datasets							

4.2 Interlinking

Initial interlinking is done among the classifications themselves in the datasets. The OECD classifications in particular contained highly similar codes (in some cases the same) throughout its codelists. Hence, majority of the codes within the codelist `CL_*_LOCATION` was interlinked with one another using `skos:exactMatch`.

The consequent interlinking was done with DBpedia, World Bank Linked Data, Transparency International Linked Data, and EUNIS using LInk discovery framework for MEtric Spaces (LIMES): *A Time-Efficient Hybrid Approach to Link Discovery* (Ngonga Ngomo, A.-C., 2011)

Interlinks between datasets

Source dataset	Target dataset	Entity type	Link relation	Link count
OECD	World Bank	skos:Concept	skos:exactMatch	3487
OECD	Transparency International	skos:Concept	skos:exactMatch	3335
OECD	DBpedia	skos:Concept	skos:exactMatch	3613
OECD	BFS	skos:Concept	skos:exactMatch	3383
OECD	FAO	skos:Concept	skos:exactMatch	3360
OECD	ECB	skos:Concept	skos:exactMatch	3495
BFS	World Bank	skos:Concept	skos:exactMatch	185
BFS	DBpedia	skos:Concept	skos:exactMatch	261
FAO	World Bank	skos:Concept	skos:exactMatch	178
FAO	Transparency International	skos:Concept	skos:exactMatch	167
FAO	DBpedia	skos:Concept	skos:exactMatch	875
FAO	EUNIS	skos:Concept	skos:exactMatch	359
FAO	ECB	skos:Concept	skos:exactMatch	210
ECB	World Bank	skos:Concept	skos:exactMatch	188
ECB	Transparency International	skos:Concept	skos:exactMatch	167
ECB	DBpedia	skos:Concept	skos:exactMatch	239
ECB	BFS	skos:Concept	skos:exactMatch	221
ECB	FAO	skos:Concept	skos:exactMatch	210
IMF	World Bank	skos:Concept	skos:exactMatch	26
IMF	Transparency International	skos:Concept	skos:exactMatch	23
IMF	DBpedia	skos:Concept	skos:exactMatch	25
IMF	BFS	skos:Concept	skos:exactMatch	24
IMF	FAO	skos:Concept	skos:exactMatch	23
IMF	ECB	skos:Concept	skos:exactMatch	26
UIS	World Bank	skos:Concept	skos:exactMatch	964
UIS	Transparency International	skos:Concept	skos:exactMatch	854
UIS	DBpedia	skos:Concept	skos:exactMatch	964
UIS	BFS	skos:Concept	skos:exactMatch	849
UIS	FAO	skos:Concept	skos:exactMatch	825
UIS	ECB	skos:Concept	skos:exactMatch	855
UIS	IMF	skos:Concept	skos:exactMatch	119
UIS	OECD	skos:Concept	skos:exactMatch	17337
UIS	FRB	skos:Concept	skos:exactMatch	800
UIS	Geonames	skos:Concept	skos:exactMatch	959
UIS	IATI	skos:Concept	skos:exactMatch	964
UIS	Humanitarian Response	skos:Concept	skos:exactMatch	835
UIS	Eurostat	skos:Concept	skos:exactMatch	964
FRB	DBpedia	skos:Concept	skos:exactMatch	280
FRB	World Bank	skos:Concept	skos:exactMatch	280

Figure [SDMX concept links] gives an overview of the complete connectivity of a concept that's linked internally, externally, and with sdmx-codes where applicable, as well as the interlinking that was done to an external concept.

Figure : SDMX Concept links

4.3 RDF Data Storage

Apache Jena's TDB storage system is used to load the RDFized data using TDB's incremental `tdbloader` utility. `tdbstats`, the tool for TDB Optimizer is executed after a complete load to internally update the count of resources in order for TDB to make the best decision to come up with future query results.

Individual datasets from each organization were transformed to N-Triples format before loading into the store. Each RDF Data Cube dataset was imported into its own `NAMED GRAPH` in the store. Given the significant load speed on an empty database, N-Triples were ordered from largest to smallest, and then loaded.

5. PUBLICATION

The publication steps are described in this section.

5.1 Dataset Discovery And Statistics

As VoID file is generally intended to give an overview of the dataset metadata i.e., what it contains, ways to access it or query it, each dataspace contains Vocabulary of Interlinked Datasets (*VoID*) files accessible through their `.well-known/void`. Each OECD, BFS, FAO, ECB, IMF, UIS, FRB VoID contains locations to RDF datadumps, named graphs that are used in the SPARQL endpoint, used vocabularies, size of the datasets, interlinks to external datasets, as well as the provenance data which was gathered through the retrieval and transformation process. The VoID files were generated automatically by first importing the LODStats information into respective `graph/void` into the store, and then a SPARQL `CONSTRUCT` query to include all triples as well as additional ones which could be actively created based on the available information in all graphs.

Dataset statistics are generated and are included in the VoID file using LODStats, *LODStats – An Extensible Framework for High-performance Dataset Analytics* (Demter, J., 2012).

5.2 User-interface

The HTML pages are generated by the Linked Data Pages framework, where Moriarty, Paget, and ARC2 does the heavy lifting for it. Given the lessons learned over the years about Linked Data publishing, there is a consideration to either take Linked Data Pages further (originally written in 2010), or to adapt one of the existing frameworks after careful analysis.

5.3 SPARQL Endpoint

Apache Jena Fuseki is used to run the SPARQL server for the three datasets. SPARQL Endpoints are publicly accessible and read only at their respective `/sparql` and `/query` locations for OECD, BFS, FAO, ECB, IMF, UIS, FRB. Currently, 12000MB of memory is allocated for the single Fuseki server running all datasets.

5.4 Data Dumps

The data dumps for the datasets are available from their respective `/data/` directories: OECD, BFS, FAO, ECB, IMF, and UIS. Additionally, they are mentioned in the VoID files. The the Data Hub entries (see below) also contains links to the dumps.

5.5 Source Code

The code for transformations is at csarven/linked-sdmx, and for retrieval and data loading to RDF store for OECD is at csarven/oecd-linked-data, for BFS is at csarven/bfs-linked-data, for FAO is at csarven/fao-linked-data, for ECB is at csarven/ecb-linked-data, for IMF is at csarven/imf-linked-data, for UIS is at csarven/uis-linked-data, for FRB is at csarven/frb-linked-data. All using the Apache License 2.0.

5.6 Data License

All published Linked Data adheres to original data publisher's data license and terms of use. Additionally attributions are given on the websites. The *Linked Data* version of the data is licensed under CC0 1.0 Universal (CC0 1.0) Public Domain Dedication.

5.7 Announcing The Datasets

For other ways for these datasets to be discovered, they are announced at mailing lists, status update services, and at the Data Hub: OECD is at data/oecd-linked-data, BFS is at dataset/bfs-linked-data, FAO is at dataset/fao-linked-data, ECB is at dataset/ecb-linked-data, IMF is at dataset/imf-linked-data, UIS is at dataset/uis-linked-data, FRB is at dataset/frb-linked-data.

6. CONCLUSIONS

With this work we provided an automated approach for transforming statistical SDMX-ML data to Linked Data in a single step. As a result, this effort helps to publish and consume large amounts of quality statistical Linked Data. Its goal is to shift focus from mundane development efforts to automating the generation of quality statistical data. Moreover, it facilitates to provide RDF serializations alongside the existing formats used by high profile statistical data owners. Our approach to employ XSLT transformations does not require changes to well established workflows at the statistical agencies.

One aspect of future work is to improve the SDMX-ML to RDF transformation quality and quantity. Regarding quality, we aim to test our transformation with further datasets to identify shortcomings and special cases being currently not yet covered by the implementation. Also, we plan the development of a coherent approach for (semi-)automatically interlinking different statistical dataspaces, which establishes links on all possible levels (e.g. classifications, observations). With regard to quantity, we plan to publish statistical dataspaces for Bank for International Settlements (BIS), World Bank and Eurostat based on SDMX-ML data.

The current transformation is mostly based on the generic SDMX format. Since some of the publishers make their data available in compact SDMX format, the transformation toolkit has to be extended. Alternatively, the compact format can be transformed to the generic format first (for which tools exist) and then Linked SDMX transformations can be applied. Ultimately, we hope that Linked Data publishing will become a direct part of the original data owners workflows and data publishing efforts. Therefore, further collaboration on this will expedite the provision of uniform access to statistical Linked Data.

7. ACKNOWLEDGEMENTS

We thank Richard Cyganiak for his ongoing support, as well as graciously offering to host the dataspaces on a server at Digital Enterprise Research Institute. We also acknowledge the support of Bern University of Applied Sciences for partially funding the transformation effort for the pilot Swiss Statistics Linked Data project and thank Swiss Federal Statistical Office for the excellent collaboration from the very beginning.