Linked SDMX Data

Path to high fidelity Statistical Linked Data

Editor(s): Oscar Corcho, Universidad Politécnica de Madrid, Spain; Jens Lehmann, University of Leipzig, Germany
Solicited review(s): Francois Scharffe, University of Montpellier 2, France; 2 anonymous reviewers

Sarven Capadisli [a,*], Sören Auer [b] , Axel-Cyrille Ngonga Ngomo [b]
[a] *Universität Leipzig, Institut für Informatik, AKSW, Postfach 100920, D-04009 Leipzig, Germany*
E-mail: info@csarven.ca
[b] *Universität Leipzig, Institut für Informatik, AKSW, Postfach 100920, D-04009 Leipzig, Germany*
E-mail: {lastname}@informatik.uni-leipzig.de

Document Identifier http://csarven.ca/linked-sdmx-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, and IMF datasets with that tooling.

Keywords: Linked Data, Statistics, SDMX, Data transformation, Dataspaces

1. 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 Organisation for Economic Co-operation and Development (OECD)[1], Bundesamt für Statistik (BFS, Swiss Federal Statistical Office)[2], Food and Agriculture Organization of the United Nations (FAO)[3], European Central Bank (ECB)[4], International Monetary Fund (IMF)[5] are retrieved, transformed and published as Linked Data.

*Corresponding author. E-mail: info@csarven.ca.

[1] `http://www.oecd.org/`
[2] `http://www.bfs.admin.ch/`
[3] `http://www.fao.org/`
[4] `http://www.ecb.int/`
[5] `http://www.imf.org/`

1570-0844/13/$27.50

2. Background

As pointed out in Statistical Linked Dataspaces [2], 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 [3]. 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 [5]. As far as graph formats go, majority of datasets in those formats are not published by the owners. However, there are number of statistical linked dataspaces in the LOD Cloud already[6].

A number of transformation efforts are performed by the Linked Data community based on various formats. For example, the World Bank Linked Dataspace[7] is based on custom XML that the World Bank[8] provides through their APIs with the application of XSL Templates. The Transparency International Linked Dataspace[9] is based on CSV files with the transformation step through Google Refine[10] and the RDF Extension[11]. 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.

[6]http://lod-cloud.net/
[7]http://worldbank.270a.info/
[8]http://worldbank.org/
[9]http://transparency.270a.info/
[10]http://code.google.com/p/google-refine/
[11]http://refine.deri.ie/

3. SDMX-ML to Linked Data

Recently, SDMX was approved by ISO as an international standard: ISO 17369:2013[12]. 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 industry standard for expressing statistical data. It has a highly structured mechanism to represent statistical observations, classifications, and data structures. Organizations supporting SDMX are Bank for International Settlements (BIS)[13], Organisation for Economic Co-operation and Development (OECD), United Nations (UN)[14], European Central Bank (ECB), World Bank (WB), International Monetary Fund (IMF)[15], Food and Agriculture Organization of the United Nations (FAO) and Eurostat[16].

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.

Data Sources As a demonstration of the SDMX-ML to RDF transformations, we selected datasets from the following organizations:

- OECD, whose mission is to promote policies that will improve the economic and social well-being of people around the world.
- BFS Swiss Statistics, due to the Federal Statistical Office's web portal offering a wide range of statistical information including population, health, economy, employment and education.
- FAO, which works on achieving food security for all to make sure people have regular access to enough high-quality food.
- ECB, whose main task is to maintain the euro's purchasing power and thus price stability in the euro area.
- 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.

[12]http://www.iso.org/iso/catalogue_detail.htm?csnumber=52500
[13]http://www.bis.org/
[14]http://www.un.org/
[15]http://imf.org/
[16]http://epp.eurostat.ec.europa.eu/

The OECD, FAO, ECB, and IMF datasets consisted of observational and structural data. The OECD and ECB data 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.

The architectural workflow of the dataspaces consists of data retrieval, transformations, enrichment, storage and publication. Along the way, information about provenance is incorporated in some of the phases.

Data Retrieval As SDMX-ML publishers have their own publishing processes, availability and accessibility of the data varied. We performed a combination of HTML scraping, site search for SDMX files and data catalog retrieval to obtain the dataset codes, names, and URLs, which we then fed into a Bash script to retrieve the actual data. Details can be found in [17].

By in large, there was no need to pre-process the data as the transformation dealt with the data as it was.

4. Provenance

Provenance at Retrieval At the time of data retrieval, information pertaining to provenance was captured using the *PROV Ontology*[18] 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 such as the tools that were used to process the data. This provenance data from retrieval may be provided to the XSL Transformer during the transformation phase and VoID enrichment.

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.

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[19]).

5. Data Modeling

In this section we go over several areas which are at the heart of representing SDMX-ML data as Linked Data. The approach taken was to provide a level of consistency for data consumers and developers.

Vocabularies In addition to RDF, RDFS, XSD, OWL, the RDF Data Cube vocabulary [4] isused to describe multi-dimensional statistical data, SDMX-RDF is used for the statistical information model. PROV-O is used for capturing provenance data. SKOS and XKOS to cover concepts, concept schemes and their relationships.

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, we deem it is as a good practice for identifying different concepts and data structures. Jeni Tennison et al discussed Versioning URIs[20], and concluded that

[17] `http://csarven.ca/linked-sdmx-data`
[18] `http://www.w3.org/TR/prov-o/`
[19] `http://www.w3.org/Provider/Style/URI.html`
[20] `http://www.jenitennison.com/blog/node/112`

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 [7].

URI Patterns An outline for the URI patterns is given in Table 1. `authority` is replaced with the domain (see also: Agency identifiers and URIs) followed by `class`, `code`, `concept`, `dataset`, `property`, `provenance`, or `slice` for each prominent area in SDMX structures. These tokens as well as `/` which is used to separate the dimension concepts in URIs can be configured in our 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.

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.

6. Linked SDMX Data Transformation

The Linked SDMX XSLT 2.0 templates and scripts[21] 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 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.

[21] `https://github.com/csarven/linked-sdmx`

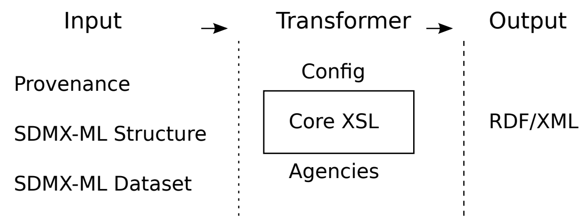

Fig. 1. Transformation process.

Features of the transformation

– Transformation of SDMX KeyFamilies, ConceptSchemes and Concepts, CodeLists and Codes, Hierarchical CodeLists, and 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.

Configuration 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. In sequel some of they key features are described in more detail.

Agency identifiers and URIs An RDF file is used to lookup information on maintenance agencies (i.e., the data owner and publisher). It includes maintenance agencies' identifiers in the SDMX Registry, as well as their base URI. It allows to look up base URIs using the agency identifier. For example, Listing 1 shows a coded property that is used by European Central Bank to associate a code list defined by Eurostat as an external agency:

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

Listing 1: Referencing external agencies.

We decided to avoid re-defining metadata from external agencies, since the owners of the data would define them under their authority. If the agency identifier is `SDMX` the corresponding URIs from the SDMX-RDF vocabulary are used instead.

URI configurations Separate base URIs can be set for classes, codelists, concept schemes, datasets, slices, properties, provenance, as well as for the location of the source data for transformations. The value

Table 1
URI patterns

Entity type	URI Pattern
qb:DataStructureDefinition	http://{authority}/structure/{KeyFamilyID}
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-no-FREQ}
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	http://{authority}/property/{conceptID}
qb:DimensionProperty	http://{authority}/property/{conceptID}
qb:MeasureProperty	http://{authority}/property/{conceptID}
qb:AttributeProperty	http://{authority}/property/{conceptID}

for `uriThingSeparator` (e.g., `/`), sets the delimiter for separating the "thing" from the rest of the URI. This is typically either a `/` or `#`. Similarly, `uriDimensionSeparator` can be set to separate dimension values used in RDF Data Cube observation URIs. Each observation requires a unique URI construction. One simple and user-friendly approach is to construct URIs by using URI-safe dimension values as tokens separated by the `uriDimensionSeparator`. Listing 2 shows an example observation URI with `/` as `uriDimensionSeparator`.

```
http://{authority}/dataset/HEALTH_STAT/EVIEFE00/
    EVIDUREV/AUS/1960
```

Listing 2: Example observation URI

Default language From the configuration, it is possible to assign a default language on `skos:prefLabel` and `skos:definition` property values, when language is not originally set for a data item. Default language may also be applied in the case of SDMX Annotations.

Interlinking SDMX Annotations The conventions in annotations typically differ from one SDMX publisher to another as there is no standardization. In order to retain this valuable information, the configuration file allows publishers to define the way annotations should be transformed. This is accomplished by defining the annotation types that should be interlinked or described with, by providing the range i.e., either an URL or a literal. The predicate to connect both resources are also defined here.

Omitting components There are cases in which certain data parts contain errors. The configuration option `omitComponents` allows to omit this erroneous data without effecting other parts, as well as to abstain from making any significant assumptions or changes to the data.

7. Linked Datasets

This section describes the transformation result and the publication of the OECD, BFS, FAO, and ECB datasets.

RDF Datasets The original SDMX-ML files were transformed to RDF/XML using XSLT 2.0. Saxon's command-line XSLT tool `saxonb-xslt` was used and employed as part of shell scripts to iterate through all the files in the datasets. Although only 4 GB of memory was necessary, 12 GB were allocated on a machine with Linux kernel 3.2.0-33-generic running on an Intel(R) Xeon(R) CPU E5620 @ 2.40GHz. Table 2 provides information on datasets; input SDMX-ML size, output RDF/XML size, their size difference in ratio, and the total amount transformation time. Table 3 summarizes the transformed data; number of triples it contains, as well as the number of `qb:Observation`, and the ratio. Table 4 provides further statistics on prominent resources. It gives a contrast between the classifications and the dataset.

Interlinking SDMX concept schemes and code lists are two valuable artifacts that are used by data owners to precisely denote the meaning of observational data. The concepts and codes within are also reused by

Table 2
Transformation time

Dataset	Input size	Output size	Ratio	Time
OECD	3,430 MB	23,000 MB	1:6.7	7885s
BFS	87 MB	139 MB	1:1.6	158s
FAO	902 MB	5,000 MB	1:5.5	1908s
ECB	5,670 MB	24,000 MB	1:4.2	10863s
IMF	330 MB	3,600 MB	1:10.9	28m11.826s

Table 3
Transformed data

Dataset	Triples	`qb:Observations`	Ratio
OECD Dataset	225M	24M	9.4:1
OECD Metadata	0.77M	N/A	N/A
BFS Metadata	1M	N/A	N/A
FAO Dataset	53M	7.2M	7.4:1
FAO Metadata	0.36M	N/A	N/A
ECB Dataset	241M	12.5M	19.3:1
ECB Metadata	0.45M	N/A	N/A
IMF Dataset	36M	3.3M	10.9:1
IMF Metadata	0.03M	N/A	N/A

Table 4
Resource counts

Dataset	skos:CS*	skos:Concept	rdf:Property	qb:Observation
`OECD`	1,212	43,368	126	24,381,106
`BFS`	185	106,233	0	0
`FAO`	32	28,115	12	7,186,764
`ECB`	149	54,389	209	12,513,494
`IMF`	25	1,565	14	3,227,978

*: `skos:ConceptScheme`

external agencies by way of referring to their unique identifiers. Thus, the interlinking phase that was undertaken for the datasets is complimentary to referencing external code lists as discussed in Agency identifiers and URIs. 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 code lists. Hence, the majority of the codes were interlinked with one another using the property `skos:exactMatch`. Further interlinking was performed among the datasets themselves as well as with other datasets using the LIMES link discovery framework [6], including: DBpedia[22], World Bank[23] (WB), Transparency International[24] (TI), and EUNIS[25]. Table 5 describes the interlinking between the datasets. Figure 2 provides an overview on the complete connectivity of a concept. This comprises linking internally, externally, and with sdmx-codes where applicable, as well as the interlinking with external concepts.

RDF Data Storage Apache Jena's TDB storage system[26] is used to load the RDF data using the TDB incremental `tdbloader` utility. `tdbstats`, the tool for TDB Optimizer is executed after a complete load to internally update the resource counts for query optimization. Each dataset was imported into its own `NAMED GRAPH` in the store. Given the significant load speed on an empty database, N-Triple files were ordered from largest to smallest, and then loaded.

8. Publication

Dataset Discovery and Statistics The Vocabulary of Interlinked Datasets (VoID)[27] file gives an overview of the dataset, for example, what it contains, ways to access or query the dataset. Each dataspace contains files accessible through their `.well-known/void` locations. Each OECD, BFS, FAO, ECB, and IMF VoID[28] 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 [1] information into a `graph/void` named graph, and then executing a SPARQL `CONSTRUCT` query to include all triples as well as relevant additional information from other graphs.

User Interface The HTML pages are generated by the Linked Data Pages[29] framework, which employs Moriarty[30], Paget[31], and ARC2[32].

[22] `http://dbpedia.org/`
[23] `http://worldbank.270a.info/`
[24] `http://transparency.270a.info/`
[25] `http://eunis.eea.europa.eu/`
[26] `http://incubator.apache.org/jena/documentation/tdb/`
[27] `http://www.w3.org/TR/void/`
[28] http://{ oecd | bfs | fao | ecb | imf}.270a.info/.well-known/void
[29] `https://github.com/csarven/linked-data-pages`
[30] `http://code.google.com/p/moriarty/`
[31] `http://code.google.com/p/paget/`
[32] `https://github.com/semsol/arc2`

Table 5
Links between datasets

Source	**Target**	**Entity type**	**Link relation**	**Count**
OECD	WB	`skos:Concept`	`skos:exactMatch`	3487
OECD	TI	`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	WB	`skos:Concept`	`skos:exactMatch`	185
BFS	DBpedia	`skos:Concept`	`skos:exactMatch`	261
FAO	WB	`skos:Concept`	`skos:exactMatch`	178
FAO	TI	`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	WB	`skos:Concept`	`skos:exactMatch`	188
ECB	TI	`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	WB	`skos:Concept`	`skos:exactMatch`	26
IMF	TI	`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

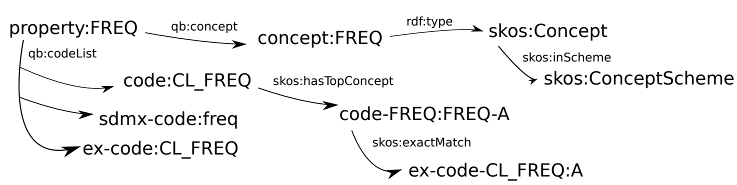

Fig. 2. SDMX Concept links

SPARQL Endpoint Apache Jena Fuseki[33] is used to run the SPARQL server for the datasets. SPARQL endpoints are publicly accessible and read only at their respective `/sparql` and `/query` locations for OECD, BFS, FAO, ECB, and IMF[34]. Currently, 12 GB of memory is allocated for the single Fuseki instance serving all datasets.

Data Dumps The data dumps for the datasets are available from their respective `/data/` directories: OECD, BFS, FAO, ECB, and IMF[35]. Additionally, they are referenced in the VoID files and from the Data Hub[36] entries.

Source Code The Linked SDMX toolkit and for retrieval and data loading to the RDF store for OECD,

[33] `http://incubator.apache.org/jena/documentation/serving_data/`
[34] http://{ oecd | bfs | fao | ecb | imf }.270a.info/sparql
[35] http://{ oecd | bfs | fao | ecb | imf }.270a.info/data/
[36] `http://datahub.io/`

BFS, FAO, ECB, and for IMF[37] is available at GitHub[38] using the Apache License 2.0[39].

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[40].

9. 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.

10. Acknowledgements

We thank Richard Cyganiak[41] for his ongoing support, as well as graciously offering to host the dataspaces on a server at Digital Enterprise Research Institute[42]. We also acknowledge the support of Bern University of Applied Sciences[43] for partially funding the transformation effort for the pilot Swiss Statistics Linked Data project and thank Swiss Federal Statistical Office[44] for the excellent collaboration from the very beginning.

References

[1] S. Auer, J. Demter, M. Martin, and J. Lehmann. Lodstats – an extensible framework for high-performance dataset analytics. In A. Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Acquin, A. Nikolov, N. Aussenac-Gilles, and N. Hernandez, editors, *Knowledge Engineering and Knowledge Management*, volume 7603 of *Lecture Notes in Computer Science*, pages 353–362. Springer Berlin Heidelberg, 2012.

[2] S. Capadisli. Statistical linked dataspaces. Master's thesis, National University of Ireland, Galway, College of Engineering and Informatics, Digital Enterprise Research Institute, 2012.

[3] R. Cyganiak, M. Hausenblas, and E. McCuirc. Official statistics and the practice of data fidelity. In D. Wood, editor, *Linking Government Data*, pages 135–151. Springer, New York, 2011.

[4] R. Cyganiak and D. Reynolds. The RDF data cube vocabulary. Candidate Recommendation, W3C, 25 June 2013.

[5] M. Hausenblas, B. Villazón-Terrazas, and R. Cyganiak. Data shapes and data transformations. *CoRR*, abs/1211.1565, 2012.

[6] A.-C. Ngonga Ngomo and S. Auer. LIMES – a time-efficient approach for large-scale link discovery on the web of data. In T. Walsh, editor, *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, 16–22 July 2011*, volume 11, pages 2312–2317, 2011.

[7] J. Sheridan and J. Tennison. Linking UK government data. In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, *Proceedings of the Linked Data on the Web Workshop (LDOW2010)*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.

[37] https://github.com/csarven/{ oecd | bfs | fao | ecb | imf }-linked-data
[38] https://github.com/csarven/linked-sdmx
[39] http://www.apache.org/licenses/LICENSE-2.0.html
[40] http://creativecommons.org/publicdomain/zero/1.0/
[41] http://richard.cyganiak.de/
[42] http://deri.ie/
[43] http://bfh.ch/
[44] http://www.bfs.admin.ch/