LinkedSpending: OpenSpending becomes Linked Open Data

Extended Abstract

Anonymous Author(s)

ABSTRACT

We present the conversion of financial transactions from OpenSpending.org to Linked Open Data, which includes more than five million transactions in 955 RDF Data Cubes.

CCS CONCEPTS

• **Information systems → Resource Description Framework (RDF)**; • **Computing methodologies → Semantic networks**; • **Applied computing → Economics**;

KEYWORDS

fiscal data, RDF, data cube, Linked Open Data

ACM Reference Format:
Anonymous Author(s). 2018. LinkedSpending: OpenSpending becomes Linked Open Data: Extended Abstract. In *Proceedings of The Web Conference.* ACM, New York, NY, USA, 5 pages. https://doi.org/unkownDOI

1 INTRODUCTION

Increased transparency of government spending is benefical to accountability and efficiency [5], leads to a larger size of government [2] and is in high demand from the public [16]. Several States and Unions are bound to financial transparency by law, such as the European Union [1]. Public spending services satisfy basic information needs, but in their current form they do not allow queries which go further than simple keyword search or which cannot be answered with data from one system alone. Linked Data technologies provide a the unified data model of RDF, a powerful query language and the possibility of integration with linked data sets from other services. Our contribution is an RDF transformation of the OpenSpending[1] project which provides government spending financial transactions from all over the world and is thus suitable as a core knowledge base that can be enriched and integrated with other, more focused data sets. Transforming OpenSpending to Linked Data and publishing it adds to and profits from the Semantic Web which offers benefits including a standardized interface, easier data integration and complex queries over multiple knowledge bases.

As LinkedSpending is represented in Linked Open Data, it facilitates data integration, which enables *economic analysis*. Currencies from DBpedia and countries from LinkedGeoData are already integrated. Financial data offers further integration candidates, such as political or other statistical, policy-influencing data such as health care. This allows queries such as the following, which asks for data sets with currencies whose inflation rates are greater than 10 %:

```
select distinct ?d ?c ?r
{?o qb:dataSet ?d. ?o  dbo:currency ?c. ?c
    dbp:inflationRate ?r. filter(?r > 10)}
```

[1] http://openspending.org

The Web Conference, April 2018, Lyon, France
2017. ACM ISBN unknown ISBN…$15.00
https://doi.org/unkownDOI

Table 1: Namespaces and prefixes used in the paper

prefix	URL
os	http://openspending.org/
owl	http://www.w3.org/2002/07/owl#
ls	http://linkedspending.aksw.org/instance/
lso	http://linkedspending.aksw.org/ontology/
qb	http://purl.org/linked-data/cube#
dbpedia	http://dbpedia.org/resource/
dbp	http://dbpedia.org/property/

LinkedSpending can also be used to compute economic indicators across several data sets. A possible indicator is a country's spending on education per person where the population size can be taken from the LinkedGeoData countries linked from one or more budget data sets. One such data set is `ugandabudget`, which contains the Uganda Budget and Aid to Uganda, 2003–2006. LinkedSpending serves as a hub for the integration of those data sets and their provenance information. More data sets can be integrated with similarity-based interlinking tools such as LIMES [15].

Another use case is *finding and comparing relevant data sets*. Government spending amounts are often much higher than the sums ordinary people are used to dealing with but even for policy makers it is hard to understand whether a certain amount of money spent is too high or normal. Comparing data sets and finding those which are similar to another one helps separating common values from outliers which should be further investigated. For example, if another country has a similar budget structure but spends way less on health care with a similar health level, it should be investigated whether that discrepancy is caused by inherent differences such as different minimum wages or a different climate or if it is due to preventable factors such as inefficiencies or corruption. While OpenSpending provides several hundreds of data sets which can be searched and it allows browsing and visualization of any single one, it does not provide a comparison function between data sets. Because of the mechanism to identify equivalent properties (see Section 3), SPARQL queries can compare different data sets, e.g. between similar structures in different countries. Another examplary SPARQL query finds data sets that are the most similar to any given data set:

```
select ?d (count(?c) as ?count)
{ls:2012_tax qb:structure ?s. ?s qb:
    component  ?c.
 ?d qb:structure ?s2.          ?s2 qb:
     component ?c.
 FILTER(?d!=ls:2012_tax)}
 group by ?d order by desc(?count)
```

This is done by calculating the number of common measures, attributes and dimensions.

2 OPENSPENDING SOURCE DATA

The OpenSpending project aims to track and analyze public spending worldwide. Data Sets can be submitted and modified by anyone but they have to pass a sanity check from the OpenSpending Data Team which also cleans the data before publishing.[2] OpenSpending hosts transactional as well as budgetary data with a focus on government finance.[3] It contains this data in structured form stored in database tables and provides searching and filtering as well as visualizations and a JSON REST interface. The data sets differ in granularity and type of accompanying information, but they share the same meta model.

2.1 The Data Cube Model

The domain model of OpenSpending is a *data cube* (also *OLAP cube*, *hypercube*), which represents multi-dimensional statistical observations. Each cell corresponds to an observation (an instance of spending or revenue) that contains measurements (e.g. the amount of money spent or received). The context of the measurement is provided by the *dimensions* like the purpose, department and time of a spending item and optionally by *attributes*, which further describe the measured value, e.g., the unit of the measurement.

```
"sub-programme": {
 "label": "Sub-programme",
 "type": "compound",
},
"amount": {
    "datatype": "float",
    "label": "Total",
    "type": "measure",
}
```

Figure 1: simplified excerpt of an OpenSpending *model*

```
"sub-programme": {
 "label":"Security and safeguarding liberties",
 "html_url":"http://openspending.org/eu-budget/sub-
      programme/security-and-safeguarding-liberties",
 "name":"security-and-safeguarding-liberties"
},
"html_url": "http://openspending.org/eu-budget/entries
     /017dfcb58d05671ef9eb5a9f77fef39c8b14150c",
"amount": 41.2
```

Figure 2: simplified excerpt from an OpenSpending *entry*

Figure 1 shows an excerpt from the model of the OpenSpending data set *eu-budget* with the dimension *sub-programme* and the measure *amount*. Figure 2 shows the an entry that contains the actual values for the dimension and the measure of the observation.

[2]http://community.openspending.org/contribute/data/
[3]http://community.openspending.org/help/guide/en/financial-data-types/

2.2 Problems

While the data is well-structured and thus suitable for conversion without data cleaning or extensive preprocessing, it still poses problems that need to be taken into account: (1) New data sets are frequently added (approximately 50 per month) and, less often, existing data sets are modified. (2) Some data sets do not specify a value for all properties in all observations. (3) There are properties with the same name in different data sets where it is unknown if they specify the same property. (4) Data Cube is a meta model. The deep structure of the data sets is heterogeneous and described only shallowly. (5) The language of literals is varying between and even within data sets but the language used is not specified. Points 1 to 3 are addressed in the next section while points 4 and 5 are discussed in Section 7.

3 CONVERSION OF OPENSPENDING TO RDF

The RDF DataCube vocabulary [6], i.e. an RDF variant of the previously explained data cube model, is an ideal fit for the transformed data.

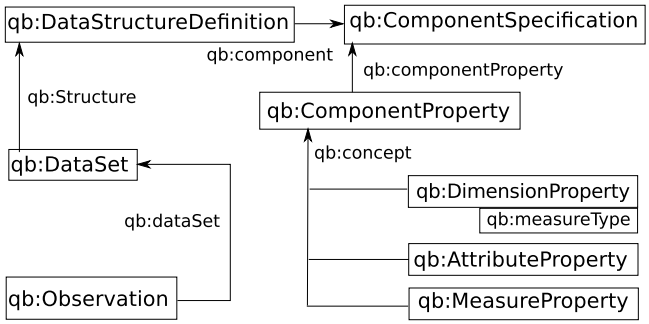

Figure 3: Used RDF DataCube concepts and their relationships[4]

First and foremost, this vocabulary provides the backbone structure for every LinkedSpending data set, see Figure 3. Each data set is represented by an instance of `qb:DataSet` and an associated instance of `qb:DataStructureDefinition` which includes *component specifications* (see Figure 4 for an example). Each component specification is associated to a *component property* which can be either a *dimension*, an *attribute* or a *measure*. Commonly used concepts are specified in the model of the *Statistical Data and Metadata eXchange (SDMX)* initiative[5]. The RDF Data Cube vocabulary is supported by the LOD2 Statistical Office Workbench[6] which is part of the Linked Data Stack (an advanced version of the the LOD2 Stack [3]). The workbench includes a DataCube validator, a split and merge component and a CKAN Publisher. The OntoWiki [4], which manages several parts of the the Linked Data Lifecycle [3], such as Storage/Querying and Search/Browsing/Exploration offers a CSV import plugin for the format as well as a faceted RDF Data Cube browser, CubeViz. Data cubes may contain slices, which are presets for certain dimension values, effectively selecting a subset

[4]Simplified version of the structure described in [6].
[5]http://sdmx.org
[6]http://demo.lod2.eu/lod2statworkbench

```
ls:berlin_de
 rdf:type          qb:DataSet;
 rdfs:label        "Berlin_Budget";
 dc:source         os:berlin_de;
 qb:structure      ls:berlin_de/model;
 qb:slice          ls:berlin_de/views/nach-einzelplan.

ls:berlin_de/model
 rdf:type qb:DataStructureDefinition;
 qb:component
    lso:CountryComponentSpecification,
    lso:DateComponentSpecification,
    lso:Einzelplan-spec,
    lso:amount-spec.

lso:CountryComponentSpecification
 rdf:type qb:ComponentSpecification;
 rdfs:label "country";
 qb:attribute sdmxa:refArea;
 qb:componentRequired "false"^^xsd:boolean;
 qb:componentAttachment
    qb:DataSet,qb:Observation.
```

Figure 4: RDF DataCube vocabulary modelling excerpt of data set `berlin_de` (some properties and values omitted).

of a cube. Users may create and visualize their own slices using the OntoWiki CubeViz plugin. Furthermore, the RDF DataCube vocabulary allows the persistence of slices which is used to represent preconfigured slices from OpenSpending.

Transformation. All of the OpenSpending data sets describe observations referring to a specific point or period in time and thus undergo only minor changes. New data sets however, are frequently added. Because of this, the huge number of data sets and their size, an automatic, repeatable transformation is required. This is realized by a program[7] which fetches a list of data sets on execution and only transforms the ones who are not transformed yet. Each data set is transformed separately. In case there are multiple instances described at one URL, a *JSON path*[8] expression is given, that locates the corresponding subnodes. Finally, the table contains the patterns that describe resulting LinkedSpending URLs. For example, the OpenSpending URL `os:berlin_de/model` contains the node `$.mapping.amount` which has a type value of "attribute" and is, thus, transformed to the OpenSpending instance `lso:amount` of the class `qb:AttributeProperty`.

Equivalent component properties (dimensions, attributes and measures) are identified as follows: A configuration file optionally specifies the mapping of data set and property name to an entity in the LinkedSpending ontology. By default, the property URI is derived from the property name. Properties with the same name in different data sets not having a mapping entry that states otherwise are assumed to represent the same concept and thus given the same URL.[9]

[7] written in Java, available as open source at https://github.com/AKSW/openspending2rdf

[8] *JSON path* (http://code.google.com/p/json-path/) is a a query language for selecting nodes from a JSON documents, similar to XPath for XML

[9] Although that has the possibility of mismatches, such a mismatch has not been spotted yet. Still, evaluating and, if necessary, improving the automatic matching is part of future work.

Use of Established Vocabularies. In addition to the standard vocabularies, RDF, RDFS, OWL and XSD, the DCMI vocabulary is used for source and generation time metadata. The data sets are modelled, first and foremost, according to the RDF Data Cube vocabulary, which specifies the structure of a data cube. LinkedSpending follows the RDF Data Cube recommendation to make heavy use of the SDMX model for measures, attributes and dimensions.The data sets are very heterogeneous but there are some properties which are commonly specified and thus modelled with established vocabularies. The year and date, a data set and an observation refers to, respectively, is expressed by `sdmx-dimension:refPeriod` and XSD. Currencies are taken from DBpedia [12] and countries are represented using the vocabulary of LinkedGeoData [18], a hub for spatial linked data. Some amount of data is imported from LinkedGeoData countries and DBpedia currencies. Because of the limited number of countries and currencies, and properties values imported per country and currency, the amount of data is too small to consider federated querying. As most countries and currencies are stable in the medium term, this data needs to be updated only infrequently.

Interlinking. There are two possibilities to align entities to another vocabulary: 1) to use the entities directly and 2) to create an own RDF resource with interlinks, like `owl:sameAs`, to that vocabulary. We generally preferred the first approach because a higher amount of reuse provides easier integration, better understandability and tool support. While we did not find *sameAs* link targets on observation level, i.e. exactly the same statistical observations described in other data sets, there are many possibilities for interlinks between data sets or dimension values and concepts they refer to. Using the labels of those data sets and dimension values, it is possible, for example, to link values of the dimension "region" of a federal budget, and thus indirectly also the observations which use those values, to the cities in DBpedia or LinkedGeoData whose labels are contained in the label of the region value URI.

4 PUBLISHING

The data is published using OntoWiki [4] under the PDDL 1.0 license[10]. Depending on the actor and needs, several services are given to access the RDF data, see Table 2. It can be explored by viewing the properties of a resource, its values and by following links to other resources (see Figure 5). Using the SPARQL endpoint provided by the underlying *Virtuoso Triple Store*[11], actors are able to satisfy complex information needs.

Faceted search offers a selection of values for certain properties and thus slice and dice of the data set according to the interests on the fly. For example, depicted in Figure 6 is all Greek police spending in a certain region. Visualization supports discovery of underlying patterns and gain of new insights about the data, for example about the relative proportions of a budget. We set up the RDF DataCube Browser CubeViz [17] as part of the human consumption interface.

5 OVERVIEW OVER THE DATA SETS

LinkedSpending consists of 955 data sets with more than five million observations total. Observations describe quantities using measures.

[10] http://opendatacommons.org/licenses/pddl/1.0/

[11] http://virtuoso.openlinksw.com

Table 2: LinkedSpending Services

OntoWiki Instance	http://linkedspending.aksw.org
SPARQL endpoint	http://linkedspending.aksw.org/sparql
datahub entry	http://datahub.io/dataset /linkedspending

All data sets have at least one measure which is the amount of money spent or received. Attributes give further context to the measurement. All data sets have at least two measures, a currency and a country, and most also have a date attribute. While the number of dimensions ranges from one to 32, almost all of the data sets have between 1 and 6 dimensions, the most common ones being the year and the time the data set and the observations refers to, respectively.

6 RELATED WORK

This work is an extended abstract of [11]. LinkedSpending is used both as the default knowledge base in the natural language Question Answering system CubeQA [9, 10] and as the target dataset of the statistical Question Answering task of the QALD 6 challenge. The TWC Data-Gov Corpus [7, 8] consists of linked government data from the Data-gov project. However, it only contains transactions made in the US and does not overlap with OpenSpending. The publicspending.gr project generates and publishes [19] public spending data from Greece based on the UK payment ontology and without using statistical data cubes. The UK government expenditure dataset COINS[12] is available as Linked Data[13]. *LOD Around-The-Clock (LATC)*[14] is a project, which was funded by the European Union (EU) and converted European open government data into RDF. One of its outcomes is the FTS[15] [13] project, which transforms and publishes financial transparency data on EU spending. In comparison with LinkedSpending, those projects also contribute linked government data but with a different or more limited scope. The Digital Agenda Scoreboard [14] is an EU project which keeps track of the transformation of statistical data to RDF.

[12]http://data.gov.uk/dataset/coins
[13]http://openuplabs.tso.co.uk/sparql/gov-coins, in a beta version
[14]http://latc-project.eu

ns0:created	2014-04-09T11:24:56.571Z
ns0:source	berlin_de
ns1:completeness	1.0
ns1:refYear	2012-01-01T00:00:00+02:00 2013-01-01T00:00:00+02:00 2014-01-01T00:00:00+02:00 2015-01-01T00:00:00+02:00
ns2:slice	Nach Einzelplan
ns2:structure	model
ns3:refArea	node424310500
rdf:type	ns2:DataSet
rdfs:comment	Berlin Budget 2009-2013 Original data: 20122013
rdfs:label	Berlin Budget

Figure 5: View of the data set `berlin_de` in the OntoWiki

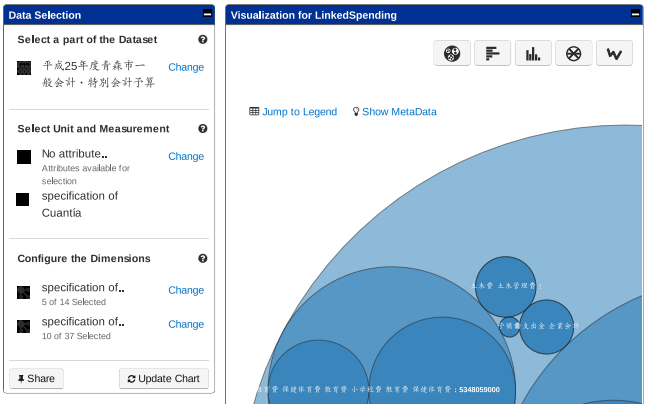

Figure 6: Faceted browsing in CubeViz by restricting values of dimensions

7 CONCLUSIONS AND FUTURE WORK

We converted several hundred financial data sets to RDF and we published them as Linked Open Data in several ways. Planned Future Work centers on the following areas: (1) RDF itself provides support for *multilingualism*, which is one of its key advantages to other representation formats. We plan statistical examinations of the relations between labels of different entities and more complex schemes based on those examinations, which can achieve language detection with a higher precision than automatic language detection. (2) Because the source data is already structured, the transformation of all the data sets without the need of text extraction and in an automatic way was feasible. On a deep level however, there is much data set specific *unmodelled structure*, that requires either a large-scale cooperation or a crowd-driven approach. (3) The hierarchical organization of the different coded properties "groups" and "functions" permits visualizations "zooming" (drilldown) in and out of the different levels of the data. The RDF Data Cube vocabulary specifies the use of skos:ConceptScheme or qb:HierarchicalCodeList but neither variant is fully implemented and it is not clear, which of those modelling possibilities will become standard. (4) Extensive *interlinking* of referenced entities to the all-purpose knowledge base of DBpedia provides additional context. Coded property values, such as the budget areas healthcare and public transportation, can be interlinked with their respective DBpedia concepts. This enables the usage of type hierarchies and thus new ways of structuring the data and provides more meaningful aggregations and new insights.

[15]http://ec.europa.eu/budget/fts

REFERENCES

[1] Regulation (EU, Euratom) no 966/2012, 2012. Article 35: Publication of information on recipients and other information.

[2] J. E. Alt, D. D. Lassen, and D. Skilling. Fiscal transparency, gubernatorial popularity, and the scale of government: Evidence from the states. Technical report, Economic Policy Research Unit (EPRU), University of Copenhagen, 2001.

[3] S. Auer, L. Bühmann, C. Dirschl, O. Erling, M. Hausenblas, R. Isele, J. Lehmann, M. Martin, P. N. Mendes, B. van Nuffelen, C. Stadler, S. Tramp, and H. Williams. Managing the life-cycle of Linked Data with the LOD2 stack. In P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. Parreira, J. Hendler, G. Schreiber, A. Bernstein, and E. Blomqvist, editors, *The Semantic Web—ISWC 2012*, pages 1–16, Berlin Heidelberg, 2012. Springer-Verlag.

[4] S. Auer, S. Dietzold, J. Lehmann, and T. Riechert. OntoWiki: A tool for social, semantic collaboration. In N. F. Noy, H. Alani, G. Stumme, P. Mika, Y. Sure, and D. Vrandecic, editors, *Proceedings of the Workshop on Social and Collaborative Construction of Structured Knowledge (CKC 2007) at the 16th International World Wide Web Conference (WWW2007) Banff, Canada, May 8, 2007*, volume 273 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.

[5] T. Berners-Lee. Putting government data online—design issues, 2009. W3C design issue.

[6] R. Cyganiak and D. Reynolds. The RDF Data Cube vocabulary. W3C recommendation, 2014.

[7] L. Ding, D. DiFranzo, A. Graves, J. Michaelis, X. Li, D. L. McGuinness, and J. Hendler. Data-gov wiki: Towards linking government data. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, volume 10, page 1, Atlanta, Georgia, 2010. AAAI Press.

[8] L. Ding, D. DiFranzo, A. Graves, J. R. Michaelis, X. Li, D. L. McGuinness, and J. A. Hendler. TWC data-gov corpus: incrementally generating linked government data from data.gov. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, New York, NY, USA, 2010. ACM.

[9] K. Höffner and J. Lehmann. Towards Question Answering on statistical Linked Data. In *Proceedings of the 10th International Conference on Semantic Systems*, pages 61–64, New York, USA, 2014. Association for Computing Machinery. doi:10.1145/2660517.2660521.

[10] K. Höffner, J. Lehmann, and R. Usbeck. CubeQA—Question Answering on RDF Data Cubes. In *Proceedings of the 15th International Semantic Web Conference (ISWC2016)*, 2016.

[11] K. Höffner, M. Martin, and J. Lehmann. LinkedSpending: OpenSpending becomes Linked Open Data. *Semantic Web Journal*, 2015.

[12] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia—A large-scale, multilingual Knowledge Base extracted from Wikipedia. *Semantic Web Journal*, 2014.

[13] M. Martin, C. Stadler, P. Frischmuth, and J. Lehmann. Increasing the financial transparency of european commission project funding. *Semantic Web Journal*, Special Call for Linked Dataset descriptions(2):157–164, 2013.

[14] M. Martin, B. van Nuffelen, S. Abruzzini, and S. Auer. The Digital Agenda Scoreboard: A statistical anatomy of Europe's way into the information age. Technical report, University of Leipzig, 2012.

[15] A.-C. Ngonga Ngomo. A time-efficient hybrid approach to link discovery. In P. Shvaiko, J. Euzenat, T. Heath, C. Quix, M. Mao, and I. Cruz, editors, *Ontology Matching, OM-2011, Proceedings of the ISWC Workshop*, pages 1–12, 2011.

[16] S. J. Piotrowski and G. G. Van Ryzin. Citizen Attitudes Toward Transparency in Local Government. *The American Review of Public Administration*, 37(3):306–323, 2007.

[17] P. E. Salas, F. Maia Da Mota, K. Breitman, M. A. Casanova, M. Martin, and S. Auer. Publishing statistical data on the web. *International Journal of Semantic Computing*, 06(04):373–388, 2012.

[18] C. Stadler, J. Lehmann, K. Höffner, and S. Auer. LinkedGeoData: A core for a web of spatial open data. *Semantic Web Journal*, 3(4):333–354, 2012.

[19] M. Vafopoulos, M. Meimaris, I. Anagnostopoulos, A. Papantoniou, I. Xidias, G. Alexiou, G. Vafeiadis, M. Klonaras, and V. Loumos. Public spending as LOD: the case of Greece. *Semantic Web Journal*, 2013.