

Publishing and Using Linked Open Data in Public Sector

Share-PSI 2.0 Berlin WS Session

Thursday 26 November, 13:45 - 15:00 Parallel Sessions D, Auditorium 2 (Room 0.036)

Facilitators: Jan Kučera (UEP), Martynas Jusevičius (UABLD)

Scribe: Jan Kučera (UEP)

Contents

[Contents](#)

[Session outline](#)

[Factors facilitating PSI re-use](#)

[Session contents](#)

[Introduction](#)

[Self-describing Fiscal Data](#)

[Publishing Linked Data with reusable declarative templates](#)

[Concluding remarks](#)

[Session schedule](#)

[Session minutes](#)

[Part A: Self-describing Fiscal Data](#)

[Part B: Publishing Linked Open Data with reusable declarative templates](#)

[Part C: Planning a domain-specific ontology](#)

[Part D: Semantics for the long term ecological research](#)

[Best practices](#)

[Make your data self descriptive](#)

[Outline](#)

[Management Summary](#)

[Challenge](#)

[Solution](#)

[Best Practice Identification](#)

[Why is this a Best Practice?](#)

[Links to the Revised PSI Directive](#)

[Why is there a need for this Best Practice?](#)

[What do you need for this Best Practice?](#)

[Applicability by other Member States](#)

[Contact Information](#)

[Related Best Practices](#)

[Increase interoperability between Linked Data software systems by using Linked Data templates](#)

[Outline](#)

[Management Summary](#)

[Challenge](#)

- [Solution](#)
- [Best Practice Identification](#)
- [Why is this a Best Practice?](#)
- [Links to the Revised PSI Directive](#)
- [Why is there a need for this Best Practice?](#)
- [What do you need for this Best Practice?](#)
- [Applicability by other Member States](#)
- [Contact Information](#)
- [Related Best Practices](#)

Session outline

Titles of the joined session proposals:

- [Self-describing Fiscal Data](#) (UEP)
- [Publishing Linked Data with reusable declarative templates](#) (UABLD)

Issues proposed for discussion in each of the proposals are summarised in the following table.

Session proposal	Issues to discuss
Self-describing Fiscal Data (UEP)	<ul style="list-style-type: none"> • A set of key elements that make up core of the fiscal datasets and • A set of recommendations for publication of self-describing fiscal data
Publishing Linked Data with reusable declarative templates (UABLD)	<ul style="list-style-type: none"> • benefits of the declarative approach • standardizing Linked Data templates as a W3C specification

Session is going to be done as a split session, i.e. two mini sessions, each based on one of the original proposals.

Factors facilitating PSI re-use

Common session questions are:

1. What X is the thing that should be done to publish or reuse PSI?
2. Why does X facilitate the publication or reuse of PSI?
3. How can one achieve X and how can you measure or test it?

Both session proposals deal with factors that facilitate PSI reuse by increasing the degree up to which data can be automatically processed. These factors stack as follows:

1. Publish self-describing data
2. Use Linked Data principles and reuse appropriate vocabularies to provide rich documentation
3. Use Linked Data templates

Session contents

Introduction

Very short, basically welcome to the audience and overview of the agenda.

Self-describing Fiscal Data

Section agenda:

- Anecdotal examples underlining importance of self-describing data
- Confirming that publication of self-describing data is of the factors facilitating PSI re-use
- What is the best practice in publishing self-describing data?
- How to achieve a good balance between data description that gives leverage to data reusers and description that is burdensome to produce for data publishers?

Publishing Linked Data with reusable declarative templates

Section agenda

- Short presentation about motivation behind LD templates
- Graphity Linked Data platform demo
- Discussion emphasising the LD templates which increase interoperability

Concluding remarks

Short summary - basically what should be said during the parallel session summary, which means it must be up to 3 minutes in length.

Session schedule

Time slot is 75 minutes long, including the “Come to my session”, so let’s plan for 60 minutes. We can adjust the exact timing on stage. If possible, facilitators introduce the session during the “Come to my session” round, while the rest prepares the room (computers, presentations etc.) for the session. Time is short so we do not want to lose time setting up things

Activity	Time
Welcoming and introductory notes	2 min.
<i>Part A: Self-describing Fiscal Data</i>	
Short introduction and motivation for self-describing data	5 min.
Discussion	17,5 min.
Summary - best practices for publication of self-describing fiscal data	5 min.
<i>Part B: Publishing Linked Data with reusable declarative templates</i>	
Introductory presentation	5 min.

Video/demo of LD templates in action (using Graphity Platform)	5 min
Discussion	12,5 min.
Summary - benefits of the declarative approach, best practices in using the declarative approach	5 min.
Closing notes	3 min.

Session minutes

Facilitators: Jindřich Mynarz (UEP), Jakub Klímek (UEP), Džiugas Tornau (UABLD/Graphity), Martynas Jusevičius (UABLD/Graphity)

Scribe: Jan Kučera (UEP)

Attendees: Peter Krantz, Hanna Niemi-Hugaerts, José-Luis Roda, Dolores Hernandez, Besjana Hysa, Fatemeh Ahmadi Zeleti, Muriel Foulonneau, Hannes Kiiwet, Michael Predeschly, Georg Hittmair, Renārs Liepiņš, Giuseppe Futia, Nicky van Oorschot, Martin Alvarez-Espinar, Miska Knapek, Andreas Harth, Joonas Dukpa, Bernhard Krabina, Noël Van Herreweghe, Emma Beer, Stuart Lester, Makx Dekkers, Peter Winstanley, Daniel Pop, Agnieszka Zajac, Federico Piovesan, Phil Archer

Session joined with Herbert Schentz and András Micsik (Semantics for the "Long Term Ecological Researchers")

Part A: Self-describing Fiscal Data

Slides:

<https://docs.google.com/presentation/d/1pzqbEYfLk7ZjYcmG7UdKWFB0987cSJiqsKWlcQcVGLo/pub>

More data users than publishers in the room.

Anecdotal evidence and introduction (J. Mynarz, J. Klímek)

- For details please see slides
- Highlights:
 - Fiscal data that is available is often difficult to use
 - No schema for EU budget in EU data portal (only hidden deep in some PDF)
 - Fiscal data is difficult to use even for policy makers
 - Self describing data - data that you can use without any prior knowledge.
 - Open data relieves you from asking for permission for reuse, however you need to ask about the meaning of the data.
 - There is a tradeoff between effort of making data self-describing and easing the re-use.
 - Data often disconnected from real-world phenomena
 - Examples of approaches making fiscal data more self-describing

- publishing data using CSV + JSON descriptor - better but there are still some issues
- publishing data using Linked Data principles and RDF Data Cube Vocabulary - all values except literals which allows interlinking, e.g. with definition of public entities

Agnieszka: EU budget data converted to RDF in an ISA project. Collaboration with Publications Office

- JiMy: We are not aware of this initiative. Are you aware of OBEU project?
- Agnieszka: Available vocabularies checked, however many specific elements so new vocabularies were created. It will be available in the metadata registry on Joinup.
- Makx: We defined not the RDF vocabulary itself, but what the vocabulary needs to cover. It is a conceptual model

Peter W: It takes a lot of effort to convert data into RDF. There is an XML markup for tabular data. Why do not use some existing solutions like XBRL.

- JiMy: I would not agree that is is difficult to reconstruct tabular data from RDF. XBRL is not linked to any standardized vocabulary and uses no
- PW: Could be added.
- JiMy: We can point to dissertation of B. Kämpgen
- Makx: I do not try to see Linked Data as a hammer. Sometimes is better to use different approaches. There are closed communities that know best how to exchange data within their community. However should be looking for way how to cross closed and open world
 - JiMy: We do not say Linked Data is the only solution. That is why we mention CSV+JSON descriptor

Andreas: How to present indicators in a way it can be linked to other data? Bunch of RDF is not always enough.

- JiMy: We can model data for example using SKOS
- Martynas: You can use tools like CubeViz
- Peter W: In financial world you do not have natural units. We need some equivalent to physical units.

Herbert Schentz: In one case we used quantity ontology as base to develop ontology for quantities

B. Krabina: Why nobody mentions international accounting standards?

- JiMy: There is a distinction between fiscal data na accounting data. However these domains are related.

JiMy: Proposes 5 star schema for self-describing data. See slides.

Part B: Publishing Linked Open Data with reusable declarative templates

Slides:

Introduction and demo (Martynas, Džiugas):

- See slides for details
- Declarative vs. imperative approach
- Motivation: replacing program code with Linked Data
 - Linked Data templates close the current gap
- Graphity Processor vocabulary example presented
- Video demo - data from Copenhagen municipality combined with data collected from the web

What is the best practice in using declarative templates for Linked Data publishing?

How can we get Linked Data implementers and publishers to adopt the LD templates?

How can LD templates be applied for fiscal Linked Data?

How can we standardize Linked Data templates as a W3C specification? Using Declarative Apps Community Group?

Phil: CSV on the web just finished its work. It basically provides what you need for tabular data.

- Martynas: Our platform is able to do the things in very generic way
- JiMy: Have you tried your approach with other solutions in terms of effort?
 - We used it mostly internally for our customers. However the core is open source. We do not have external feedback, however we see the benefits internally (compared to the imperative programming). With the declarative approach the flow is simple: create data -> model data -> plot data

Part C: Planning a domain-specific ontology

Slides:

Introduction (András Micsik)

- See slides for details
- Modelling ontology for university lectures
 - Started with existing ontologies. However there are various issues in existing ontologies - they do not always fit well, sometimes obsolete and no longer maintained
- Templating mechanism is needed - ordinary staff cannot provide data in RDF using existing ontologies
- I would be glad to hear your feedback

Part D: Semantics for the long term ecological research

Slides:

Introduction (Herbert Schentz):

- See slides for details
- LTER Europe project
- Relationships between concepts developed in the project and concepts developed in other domains, e.g. EEA vocabulary
- Open issue ontology
- Key issue: how to reuse and avoid reinventing existing ontologies
- It makes no sense to replace data silos with semantic silos
- We want to bring the vocabularies to data

Key points

Key points of the whole session:

- It is important for data to be self-describing. 5 star schema for self-describing data was proposed (see slides)
- RDF is not always the hammer. Take existing solutions (such as XBRL for financial data) into account when designing solution to a problem.
- Linked Data templates allows quick development of applications with very little code. Graphity Platform vocabulary allows development of such templates in any data domain
- Tool for development and management of ontologies are needed
- In order to avoid reinventing existing ontologies we need a better communication between those who develop the ontologies

Proposed best practices

Based on the expected factors facilitating publication and reuse of PSI as well as on the topics of the original session papers we can draft the following best practices:

- Make your data self descriptive
- Increase interoperability between Linked Data software systems by using Linked Data templates

Make your data self descriptive

Outline

Poorly documented data and data lacking schema is difficult to use and might be easily misunderstood. Therefore machine-readable description of structure and semantics of the data

should be provided to guide human re-users and to increase the degree to which the data can be automatically processed, i.e. data should be self-descriptive.

Management Summary

Challenge

Some data is complex and requires significant amount of domain-specific knowledge to interpret it. However it might be poorly documented and lacking explanation of its semantics at the same time. This makes the data difficult to use and might cause misinterpretation of the data.

Solution

Provide machine-readable description of the data which would explain its schema and semantics.

Best Practice Identification

Why is this a Best Practice?

Providing machine readable description of data improves usability of the data, guides the human re users and increases the degree to which the data can be automatically processed. It also helps to mitigate the risk of misinterpreting the data.

Links to the Revised PSI Directive

Techniques

Why is there a need for this Best Practice?

Making data available under an open licence and in machine-readable format is often not enough because without proper understanding of the data it might be difficult to use it. Therefore missing or poor documentation of the data hinders its re-use.

What do you need for this Best Practice?

Machine-readable descriptors of data such as vocabularies in RDF format or JSON descriptor that is part of the Data Package. There are domain specific specifications that should be taken into consideration when choosing an approach to make data self-describing. XBRL in the financial data domain is an example of such a domain specific solution.

Applicability by other Member States

Best practice is applicable across the member states.

Contact Information

Jindřich Mynarz (jindrich.mynarz [at] vse.cz), Jakub Klímek (klimek [at] ksi.mff.cuni.cz)

Related Best Practices

Increase interoperability between Linked Data software systems by using Linked Data templates

Outline

Interoperability of the Linked Data systems might be increased if declarative templates are used when publishing Linked Data. With such templates applications can share for example information about the URI patterns or RDF description patterns and therefore these internals no longer need to be hardcoded in the publishing software.

Management Summary

Challenge

Linked Data is standards-based framework for publishing and linking RDF data on the Web. There are many domain specific vocabularies that can be used to describe documents, entities and concepts in structured data. However, shape of the published Linked Data: the URI patterns, RDF description patterns, templates for new resource instances etc. is often hardcoded in publishing software, which makes it hard to share and reuse.

Solution

Use low-level vocabularies and provide information about the internals of your Linked Data. Use these vocabularies to define templates.

Best Practice Identification

Why is this a Best Practice?

Linked Data allows clear separation of data and applications that use it. However if for example it is necessary to hardcode URI patterns of a specific dataset in an application, the application is not fully separated from the dataset. Using declarative templates and vocabularies allowing description of Linked Data internals increases interoperability of Linked Data systems.

Links to the Revised PSI Directive

Techniques

Why is there a need for this Best Practice?

Following this best practice makes development of systems using Linked Data easier and thus it might foster adoption of Linked Data.

What do you need for this Best Practice?

Low-level vocabularies allowing description of the shape of the published Linked Data: the URI patterns, RDF description patterns, templates for new resource instances etc.

Applicability by other Member States

Best practice is applicable across the member states.

Contact Information

Martynas Jusevičius (martynas [at] graphity.org), Džiugas Tornau (dziugas [at] graphity.org)

Related Best Practices