

# Best Practice: Publish Statistical Data In Linked Data Format

25 July 2016

This version

<http://www.w3.org/2013/share-psi/bp/stats-20160725/>

Latest version

<http://www.w3.org/2013/share-psi/bp/stats/>

Previous version

<http://www.w3.org/2013/share-psi/bp/stats-20160627/>

This is one of [a set of Best Practices](#) for implementing the [\(Revised\) PSI Directive](#) developed by the [Share-PSI 2.0 Thematic Network](#).

Share-PSI Best Practice: Publish Statistical Data In Linked Data Format by [Share-PSI 2.0](#) is licensed under a [Creative Commons Attribution 4.0 International License](#).

---

## Outline

Publishing statistical data as Linked Data on the basis of W3C's [RDF Data Cube vocabulary](#) which specifies an approach for the expression of the data in a standardised machine-readable way as well as identifying a recommended set of metadata terms to describe the datasets.

## Links to the Revised PSI Directive

[Techniques](#)

## Challenge

Statistical data is currently published in a range of formats and standards that do not allow linking across datasets. It is used as the foundations for policy prediction, planning and adjustments, and therefore has a significant impact on the society (from citizens to businesses to governments). The process of collecting and monitoring socio-economic indicators can be considerably improved if the data produced by government organizations such as Statistical Offices, National Banks, Employment services, etc. are published in Linked Data Format.

## Solution

Linked Data paradigm has opened new possibilities and perspectives for government organisations to open data and interchange information. Data is open if it is technically open (available in a machine-readable standard format, which means it can be retrieved and meaningfully processed by a computer application) and legally open (explicitly licensed in a way that permits commercial and non-commercial use and re-use without restrictions), see the [World Bank Open Data Essentials](#).

The Linked Data approach enables datasets to be linked together through references to common

concepts. A dataset is represented in the form of a graph, using the Resource Description Framework (RDF) as a general-purpose language. Linked Data publication process refers to a set of activities related to extraction, transformation, validation, exploration and publication of RDF datasets originating from different sources (e.g., databases) on the Web. The ready for use RDF datasets can be either stored locally or registered at a metadata catalog e.g. build with [CKAN](#) open-source tool.

In 2014, The [RDF Data Cube Vocabulary](#) was published by the W3C [Government Linked Data Working Group](#) as a Recommendation for publishing multi-dimensional data on the Web.

## Why is this a Best Practice?

The approach contributes to the standardisation of the process of publishing and re-use of multi-dimensional data on the Web. The approach is based on RDF Data Cube vocabulary that is mature enough to be used for publishing statistical data as it improves interoperability and allows comparison of data from different statistical sources. The vocabulary underlies SDMX (Statistical Data and Metadata eXchange), an ISO standard for exchanging and sharing statistical data and metadata among organizations and provides a layer on top of data to describe domain semantics, dataset's metadata, and other crucial information needed in the process of statistical data exchange.

Cost implication: Costs of publication should be minimised unless there are clear benefits. Public sector body should analyse the current status of data availability, the demand for data and thus avoid unnecessary costs of transformation of data in Linked Data format. Public sector bodies publishing information SHOULD either:

- Publish it in the manner that involves lowest cost, consistent with making it available effectively and openly, or
- Carry out cost-benefit analyses of the possible measures to assess potential use and stimulate take-up, methods of publication, and formats for publication, and select measures, methods and formats in the light of those analyses.

The risk of deciding what publication form will best deliver value (commercial or other value of public information), and the work of converting it to that form, could be left to commercial product and service providers, and other consumers. If due to cost implications it is not possible to publish statistical data in that format, it is important to ensure possible transformations by third parties from the provided format to the RDF Data Cube Vocabulary. The multidimensional data model (with n-dimensional data cubes as datasets with observations, dimensions, measures) used by the RDF Data Cube Vocabulary is sufficiently generic to not restrict publishers.

A possible transformation has been shown for other common data formats for statistical data such as [SDMX](#), [XBRL](#), and the [Dataset Publishing Language](#). If sufficient metadata is provided, transformation scripts are also possible from CSV and spreadsheet (e.g., Microsoft Excel) data.

## How do I implement this Best Practice?

This best practice is based on a set of tools for automating the data extraction and publication process. However the EU research community delivered many open-source tools for publishing the statistical data in Linked Data format, see e.g. the [LOD2 Statistical Workbench](#), the [OpenCube](#) toolkit.

## Where has this best practice been implemented?

Country	Implementation	Contact Point
Italy	<a href="#">LOD ISTAT</a> (residency population)	<a href="#">Istat</a>
Italy	<a href="#">LinkedStat</a> (a project between ISTAT and SpazioDati)	<a href="#">SpazioDati</a> and <a href="#">Istat</a>
UK	<a href="#">Scottish Government Statistics</a>	Scottish Government
Finland	<a href="#">Semangtic hri.fi</a>	Page includes contact information
Czech Republic	<a href="#">Publikace dat statistických ročenek ve standardu otevřených dat</a>	<a href="#">Jan Kučera</a>

## References

- Samos Workshop presentation: [A Methodology for Publishing Linked Open Statistical Data](#) (PDF), George Papastefanatos IMIS / RC Athena, Greece
- Samos Workshop presentation: [Publishing and Consuming Linked Open Data with the LOD Statistical Workbench](#), Valentina Janev, Institut Mihajlo Pupin
- [Publishing Official Classifications in Linked Open Data](#), Agenzia per l'Italia Digitale (Giorgia Lodi, Antonio Maccioni), Istituto Nazionale di Statistica (Monica Scannapieco, Mauro Scanu, Laura Tosco). SemStats 2014.

## Local Guidance

This Best Practice is cited by, or is consistent with, the advice given within the following guides:

- (Belgium) [Open Data Handleiding](#) Open Data Handbook
- (Estonia) [Avaandmete loomise ja avaldamise juhend](#) Open Data Guidelines
- (Finland) [Helsinki Region Infoshare](#)
- (International) [Using Open Public Sector Information](#)
- (International) [Open Data Handbook, Solutions Bank](#)
- (Italy) [Linee Guida Nazionali per la Valorizzazione del Patrimonio Informativo Pubblico](#) National Development Guidelines for Public Sector Information
- (Lithuania) [Viešojo Sektoriaus Informacijos platinimo gerosios praktikos](#) Best Practices for Sharing Public Sector Information
- (Luxembourg) [Recommandations pour l'ouverture des données publiques](#) Recommendations for opening data
- (Serbia) [Open Data Handbook](#)
- (Spain) [Guía de aplicación de la Norma Técnica de Interoperabilidad de reutilización de recursos de información](#) Application Guide for Technical Interoperability Standard on PSI re-use
- (UK) [Open Data Resource Pack](#)

## Contact Info

Original Author & editor: [Valentina Janev](#), [Institute Mihajlo Pupin](#); contributor: [Benedikt Kämpgen](#), [FZI Research Center for Information Technology](#)

## Related Best Practices

- [Use machine-readable standardized data formats](#)
- [Reuse vocabularies](#)

## Issue Tracker

Any matters arising from this BP, including implementation experience, lessons learnt, places where it has been implemented or guides that cite this BP can be recorded and discussed on the project's [GitHub repository](#)