FAIR Digital Objects in Official Statistics

Olav ten Bosch[‡], Edwin de Jonge[‡], Henk Laloli[‡], Christine Laaboudi-Spoiden[§]

‡ Statistics Netherlands, The Hague, Netherlands
§ Eurostat, Luxembourg, Luxembourg

Corresponding author: Olav ten Bosch (o.tenbosch@cbs.nl)

Abstract

Introduction*[1]

Statistical offices on national and international scale provide statistics on demography, labour, income, society, economy, environment and other domains. Their collective output is usually referred to as 'Official Statistics'. These offices have a long tradition of publishing data fairly and open, which is often part of their mission statement. For decades they have been providing websites with articles, press releases, graphs and tables of data for free, for research, for policy-making, and for common understanding. However, for users it often is not so easy to find the data needed, to (re-)use it in data-driven work or to refer to the right (sub)set of data in a sustainable way. Therefore, in this article we take a closer look at Official Statistics from a findable, accessibility, interoperability, and reusability (FAIR) perspective.

Digital Objects in Statistics

Digital objects in official statistics can be identified on multiple levels. The core concept is the *statistical fact*: a number describing a certain estimate on a certain phenomenon in a certain population over a certain period of time. For example the estimated number of elderly inhabitants in Province Friesland (the Netherlands) on Jan 1, 2020, or the inflation in Belgium for fruits in 2021 are both statistical facts. Each of these statistical facts is uniquely defined and published as a digital object in the online statistical databases of Statistics Netherlands and Eurostat respectively. Statistical facts may have a production status (provisionary, final, revised) and are typically visualized as a number in a table cell or in a chart.

Data without metadata are without meaning. A statistical fact refers to metadata (region, time, subject, population, uncertainty, quality etc.) which are essential to understand the context of the fact. We make a distinction here between *structural* or *conceptual metadata*, i.e. the structure and definitions of concepts, dimensions and types of data used, and *referential metadata*, i.e. descriptive information on the dataset. The metadata are of utmost importance to the data consumer to understand the data. Metadata have their own

dynamics, e.g. classifications change over time. They are published as digital objects too, for example the statistical classification of economic activities (NACE).

Statistical facts and their metadata form the foundation for higher level statistics products. News releases and thematic articles that explain statistics in a broader context are examples. This higher level content can be seen as digital objects too as it is usually the main entry level for the general public and search engines and enables their findability and accessibility.

Standards and FAIR

Each digital object in official statistics has its own structure, dynamics, dissemination channels and standards. This can make it sometimes hard to work with data from official statistics.

Statistical databases differ among statistical organizations, both technically as well as in metadata and the API's that they offer for automated access. Main standards in this field are the Statistical Data and Metadata eXchange (SDMX), JSON-stat, OData, or simple formats such as CSV. Commonly agreed structural metadata is organized into SDMX registries (global registry, Eurostat registry), which provide automated access to statistical metadata, which is good for accessibility.

The SDMX standard is actually targeted to statistical and financial data which may hinder wider reusability. Therefore some statistical offices are moving to semantic standards. An an example are the vocabularies and classifications published as linked open data by Statistics Netherlands. Publishing metadata this way makes it possible to reuse and link data across organizations and gives semantic structure that is machine readable. Another example is from the statistical office of the European Union, Eurostat, that is converting the statistical classifications and correspondence tables from their current metadata system into Linked Open Data in the EU Vocabularies website. The representation is based on XKOS, an ontology for modelling statistical classifications, offering machine-readable access for reusing objects as well as facilitating linking among classifications on national, EU or international level. Yet another initiative is from the United Nations Economic Commission for Europe (UNECE), where statistical organizations collectively develop a Core Ontology for Official Statistics (COOS) describing the statistical production process. All in all for structural metadata, statistical organizations are increasingly moving towards linked data standards to better align to non-statistical communities.

In the field of referential metadata the Single Integrated Metadata Structure (SIMS) is used. It offers machine-readable descriptive metadata such as unit of measure, reference period, confidentiality, quality, accuracy etc. Some of the elements are also covered in the widely used RDF-based Data Catalog Vocabulary (DCAT) and the statistical variant (STAT-DCAT), which raises the question whether a further integration of these could improve FAIR-ness of statistical referential metadata.

With respect to higher level digital objects, such as statistical articles, the use of semantic web ontologies such as schema.org and Dublin Core for annotating statistical output in common terms are increasingly being used. The use of Digital Object Identifiers (DOIs) where applicable makes it easier to refer to statistical output.

From the above we can see that the use of different standards at different levels creates various ways to identify statistical content, such as Uniform Resource Names (URNs), SDMX identifiers, Digital Object Odentifiers (DOIs), Uniform Resource Identifiers (URIs) or organization specific identifiers. Although they probably all satisfy FAIR principle A1, from a user perspective it would be good to minimize variety here.

Wrap-up

Although official statistics have a long tradition and experience in publishing open data, the FAIR principles are an excellent vehicle to further improve findability and enable data-driven work. Openness is not enough, the facts, structural and referential metadata and higher level statistical digital objects should ideally all be optimized from a FAIR point of view. The mix of standards being used at various levels and the distributed statistical system in official statistics may hinder reusability. Moving to semantic-interoperability via generally accepted linked data standards is ongoing and has the promise to increase the reusability of statistics into a broader web of (meta)data. This makes trustful statistics more FAIR, better searchable, findable and interpretable which is necessary for a further integration of official statistics into wider communities.

Keywords

statistical (meta)data, semantic web, findability, interoperability, SDMX, ontologies, classifications

Presenting author

Olav ten Bosch

Presented at

First International Conference on FAIR Digital Objects, FDO2022, presentation

Conflicts of interest

Endnotes

*1

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of their institutes.