To be FAIR, what is missing in Official Statistics?

Olav ten Bosch[1], Edwin de Jonge[1] and Henk Laloli[1]

[1] *Statistics Netherlands, Henri Faasdreef 312, The Hague, the Netherlands*

Abstract

The FAIRness of the official statistics landscape is reviewed from three perspectives. First of all, a list of open-source software to access official statistics is used to identify common standards, common features offered in this software and how they relate to the FAIR principles. Second, the use of linked data by three official statistics organizations informs the reader about the potential of semantic technologies to further improve statistical metadata on flexibility, modeling power, linking capabilities, and alignment to other communities. Third, interesting ingredients of the emerging FAIR Digital Object (FDO) standard are described and interpreted in the scope of official statistics. The paper concludes with a set of directions for future development for the official statistics community to grow as trusted partners in the ever-evolving digital society.

Keywords

Official statistics, linked data, statistical software, knowledge graph, FAIR digital object

1. Introduction

Statistical organizations have a long tradition of publishing statistical content fairly and open. This is often part of their mission statement and aligns to the European Code of Practice (principle 15: accessibility and clarity) [1]. For decades they have been providing websites with articles, press releases, graphs, and data tables, supplemented with metadata, definitions and explanations to make users use it for research, for policy-making, for data science, for education, for fact checking, and for many other goals. The collective output of all official statistics organizations we call the *official statistics landscape*.

From a user-perspective it can be difficult to navigate and travel this landscape. In need of a statistical figure, it is still a challenge to find it, to access it smoothly, to (re-)use it in data-driven work, and to refer to it in a sustainable way. Moreover, since data without metadata is without meaning a user needs high quality metadata based on "smart" metadata standards that are understandable and interpretable, also by non-statistical user communities. Even more, in an ever-evolving digitalized society with strong needs for trusted content, automated access to statistical data is a necessity. Many of these demands are contained in the FAIR principles [2]. Implementing these principles is a good first step, but not enough. Additional concepts should be explored as outlined in this paper. We look at and into the official statistics landscape from different angles to identify promising directions of development that may improve and complement the official statistics "planet" within its evolving digital society "cosmos". We approach this challenge from three different perspectives.

In chapter 2 we start from the perspective of the open-source software offered by statistical institutes or developed by the open data community for easy access to official statistics. Such software is listed on the *awesome list of official statistics software* [3]. This allows us to identify the standards currently in use on the official statistics landscape. Furthermore, it provides us an insight into the common features offered to end-users, which we confront with the FAIR principles.

Proceedings Acronym: Proceedings Name, Month XX–XX, YYYY, City, Country

o.tenbosch@cbs.nl (O. ten Bosch); e.dejonge@cbs.nl (E. de Jonge); h.laloli@cbs.nl (H. Laloli)

0000-0002-1943-7558 (O. ten Bosch); 0000-0002-6580-4718 (E. de Jonge); 0009-0004-9921-4670 (H. Laloli)

In chapter 3 we turn our attention to the emerging use of *linked data technologies* in official statistics. Linked data has been around for many years but was never fully adopted by statistical organizations. Recently, some of them started providing statistical metadata in linked data format. We look at the progress, try to identify the added value and develop a view on future developments that may benefit the digitalized society.

In chapter 4 we go beyond FAIR principles. Although these principles are commonly accepted for open data policies, there is also room for interpretation in what they mean in practice [4]. The FAIR Digital Objects forum [5] is working on a set of standards to make these principles more practically implementable among all scientific domains using the concept of FAIR Digital Objects (FDO). In chapter 4 we explore the FDO concept and ask ourselves what this concept may mean for future statistical dissemination strategies.

In chapter 5 we wrap up the work with some observations and conclusions on how a future smart and actionable official statistics landscape might look.

2. The official statistics landscape from a software perspective

Providing FAIR access to statistical content is an essential task for Statistical Institutes. Apart from content such as press releases, textual and thematic stories they also provide access to the underlying data: the statistical estimates and their rich metadata. Many statistical organizations provide software for easy access to their dissemination database. This is reflected in the *awesome list of official statistics software*. This community-maintained list was created in 2017 during the UNECE workshop on statistical data editing (SDE) and grew over time [6] with contributions from the official statistical community and users of statistical open data. The list is organized according to the Generic Statistical Business Process Model (GSBPM) [7]. Figure 1 gives an indication of the number of software items per GSBPM category, showing that the largest category is "Access to official statistics" within GSBPM 7.4: "Promote dissemination products". This category contains over 30 software packages that help users access official statistics data or metadata from International or National organizations. In this chapter we take a closer look at the software packages on this list with the goal to understand the current state of access to the official statistics landscape, the main standards used and the functionalities offered.

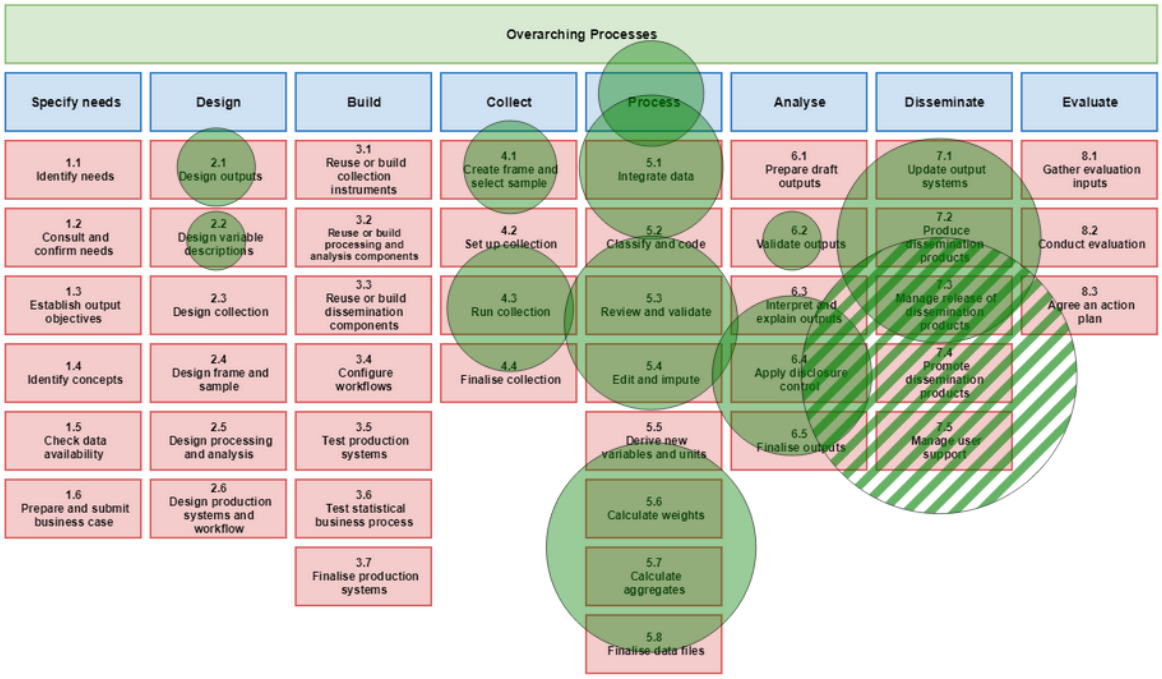

Figure 1: Software on the awesome list of statistical software organised by GSBPM with largest category "access to official statistics" in 7.4 "Promote dissemination products".

The majority of software on the list are R-packages. For the non-R software there is mostly an R equivalent giving access to the same data provider(s). Hence, we limit our analysis to the 28 R-packages in this category. Some are more generic and targeted at standardized access to multiple data providers, others contain detailed and dedicated functionality to access data from just one organization. An overview is contained in Annex I. It shows the dependencies between the packages on the list, the statistical data providers and the standards being used. These relationships were derived from the packages documentation, the pages they link to and running some of the packages that offer a list of pre-configured data-providers. Only main data providers on (inter)national level were considered. We note that this is an abstraction of reality as for example different versions of standards and endpoints are not taken into consideration, which would complicate the figure considerably.

From this table we can identify standards-based packages such as *rsdmx*, *readsdmx* using the SDMX standard [8], *rjstat* offering access to the JSON-STAT data format [9] and various *px** packages targeted at the PC-Axis format [10]. The ODATA [11] standard is used by one organization only: statistics Netherlands. Others packages such as *inegiR*, *readabs* and *statcanR* provide dedicated access to just one specific data provider. Note that it is possible that a package does use a standard internally but that it is not mentioned in the documentation as the package writer tries to hide these details from the user. In that case the use of standards is underestimated. Concluding we can see from this table the dominance of SDMX, JSON-STAT and PX as leading standards in the official statistics landscape.

Looking from the data provider side we can see that certain data providers can be accessed via multiple R-packages. Eurostat supporting SDMX as well as JSON-STAT is an example. Moreover, there are also two dedicated packages *restatapi* and *eurostat* which target Eurostat data exclusively. Another observation is that the set of providers offering JSON-STAT data is mostly disjoint from the set of SDMX providers and that all PX providers provide JSON-STAT as well. Some organisations, such as Eurostat and the World Bank, provide multiple endpoints for specific domains. Some endpoints provide harmonised data on one specific domain via a dedicated R-package. Examples are *rdbnomics* offering access to economic data from many institutions and *ipumsr* providing access to census and survey data integrated across time and space. Although a special category it is useful to note the existence of *official statistics aggregator sites* and their dedicated R-packages in the official statistics landscape.

A category, not covered so far, is access to *statistical metadata*. Many organisations, mostly international, offer access to definitions, classifications and code lists in metadata registries such as the SDMX global registry and the Eurostat registry. These predominantly use SDMX and can be accessed with generic tools. Access from R to registries has proven to be useful for statistical operations, such as checking data against internationally harmonised code list in the validation process [12].

The list of R-packages on the awesome list also provide us some insight into the functionality that is offered. We can see certain features reoccurring, such as:

- *endpoint hiding*: wrapping the preconfigured endpoint(s) in a R function within the package
- *catalogue retrieval*: the ability to list the availability datasets on the endpoint(s)
- *search*: the ability to search for datasets or within datasets on the endpoint(s)
- *endpoint queries*: the ability to query for subsets on the endpoint(s) side
- *local queries*: the ability to easily slice or filter the retrieved data on the client
- *caching*: preventing unnecessary roundtrips to the endpoint(s) by caching results
- *cartographic queries*: retrieve a (cartographic) map to be used with the data
- *registry access:* access to coordinated metadata in registries

This list is a rough inventory of features apparently needed by users in practice. It is interesting to try to map them onto the FAIR principles. Although we realize that this exercise is not an exact endeavor as the FAIR guidelines give quite some room for interpretation, the results are shown in Table 1. Most of the features are targeted at findability and accessibility. Roughly speaking the access to statistics are weak on interoperability and reusability, with the exception of the cartographic functionality which is a typical example where statistics meets another community: the spatial community. Registry access scores a yes on interoperability because it helps relating content that shares some coordinated metadata.

Table 1
Software features supporting FAIR principles

Software feature	Findability	Accessibility	Interoperability	Reusability
endpoint hiding	yes	yes		
catalogue retrieval	yes	yes		
search	yes			
endpoint queries		yes		
local queries		yes		
caching		yes		
cartographic queries	yes		yes	yes
registry access	yes	yes	yes	

We can see that the official statistics landscape grows towards standardisation of data access and software features, but also that it's still common use to develop dedicated software targeted at specific functionality or specific data providers. This results in a software landscape where the R user can choose from at least 28 packages to access official statistics, each offering different functionality and targeted at different parts of the official statistics landscape. There is no 'one-for-all' software that provides access to all official statistics providers. This is understandable from an organisational viewpoint, but from an end-user's viewpoint such software would be convenient. The situation will improve with ongoing standardisation on the data providers side, however from the statistical community it is necessary to also work towards creating a generic interface to all official statistics data providers, notwithstanding the value of the dedicated packages. The analysis presented in this paper could serve as a start.

This chapter we looked at the FAIRness of the official statistics landscape through software to access it. That's only part of the total picture, however we can see weakness on interoperability and reusability. Happily, there is ongoing work on linked data within the statistical community. The next chapter concentrates on some of the potentially fruitful developments in this area.

3. Adding linked data to the landscape

For long, linked data had the promise to make it easier to link internet content together. The idea of a "semantic web of data" was introduced by World Wide Web founder Tim Berners-Lee in 1999 [13]. The term "Linked Data" was brought up a bit later to denote "structured data which is interlinked with other so it becomes more useful through semantic queries" and was a cornerstone of the 5-star model for open data [14]. Over years the concept gave rise to a number of 'semantic' technologies. Many governments adopt linked data techniques as a base layer for open data policies, but the statistical community mainly chose their own standards, causing a weak score on reusability to other domains. There have been international projects pioneering in the use of linked data in official statistics [15], however only recently organizations started adding linked data to their output portfolio more prominently. Examples are Statistics Netherlands, the Scottish government and Eurostat. In this chapter we dive into these cases and ask ourselves what linked data adds and how it could fit best into the official statistics landscape.

3.1. Linked data at Statistics Netherlands

At Statistics Netherlands a knowledge graph has been created that supports the dissemination process in multiple ways. This knowledge graph grew over time by mining metadata spread over many different internal and external data sources, databases, registries, and textual descriptions. It was supplemented with human expertise by interviewing statistical experts, modeling their knowledge in mind maps and translating these into linked data statements. Priority has been given to the use of common standards such as Dublin Core Terms [16], DCAT [17], SKOS [18]. Moreover dedicated ontologies from https://schema.org, http://vocab.getty.edu/aat/ and http://rdf.histograph.io/ are used for modeling geo primitives such as continents, countries, provinces, municipalities, neighborhoods and relationships such as geoWithin, geoEquals, absorbedBy. Support for XKOS [19], an SKOS extension for representing statistical classifications created by the DDI Alliance is planned. We explain the most prominent parts of the knowledge graph in more detail.

The graph contains a taxonomy of terms used in Dutch official statistics which can be referred to in every web publication using such term. The taxonomy uses the SKOS standard to describe over 3000 terms in English and Dutch. It models broader and narrower relationships among concepts and makes all statistical concepts identifiable by a globally uniform resource identifier (URI) which resolves to the corresponding concept page on the CBS website. This taxonomy is also used to make the search engine on the website smarter.

Another part of the Dutch linked data describes hierarchical classifications and their international counterparts. It currently contains the classification of Dutch courses and education levels linked to the International Standard Classification of Education (ISCED) and the Dutch Standard Industrial Classification linked to the international NACE (and indirectly to ISIC). This is an example of using linked data concepts to link national metadata to European linked data in a flexible way. We come back to this in section 3.3.

Moreover, the graph contains a list of all frequencies used in the dissemination database of Statistics Netherlands. Each item links to their SDMX equivalent in the SDMX global registry and to the equivalents in the ISO-8601 standard, where possible. Since SDMX currently doesn't support URIs, the links could only be established through SDMX REST-API calls. The support for URIs is clearly a candidate for improvement on the SDMX side, which would increase reusability of Internationally standardised metadata the official statistics landscape.

Another part of the Dutch linked data is dedicated to modelling geo relationships. First of all, it contains a Dutch geo classification derived from the current contents of the Dutch dissemination database. It describes international countries and areas based on the ISO-3166-1 standard, supplemented with a number of CBS-specific area concepts and connects them to national codes to SDMX codes (CL_AREA/2.0), where possible.

Another geo-vocabulary is dedicated to modelling Dutch provinces and historical municipalities and NUTS. Every municipality from 1812 to date is present with a start and end date and, if no longer existing, a link to the absorbing municipality. This results in a complete historical graph of Dutch geo-changes on municipality level for over two centuries. To facilitate linkability of the content in wider context, links to corresponding Wikidata concepts are provided. This use-case is an example of how slowly changing statistical metadata can be modelled in linked data.

Yet another part describes the Dutch administrative areas and their interrelationships on multiple hierarchical levels. It contains over 18000 terms relating CBS-specific higher area codes to municipalities, districts and to the smallest administrative area: the neighbourhood blocks. There is a vocabulary per year, linked together via geo relationships and supplemented with change notes that provide details on the change where applicable.

The linked data parts described above are accessible from a Skosmos [20] interface generated from the underlying knowledge graph, which is implemented in a semantic graph database (GraphDB). This graph database is also publicly available but not actively advertised as it is still work in progress. It contains more than one million statements which can be explored via SPARQL. Apart from the vocabularies listed above it also connects metadata to data contained in

the statistical dissemination database of Statistics Netherlands. At this moment the connection is on dataset level, meaning that for every public multidimensional dataset it is known what metadata is used and vice versa. Figure 2 gives an impression, showing a tiny part of the huge knowledge graph that connects data to rich metadata.

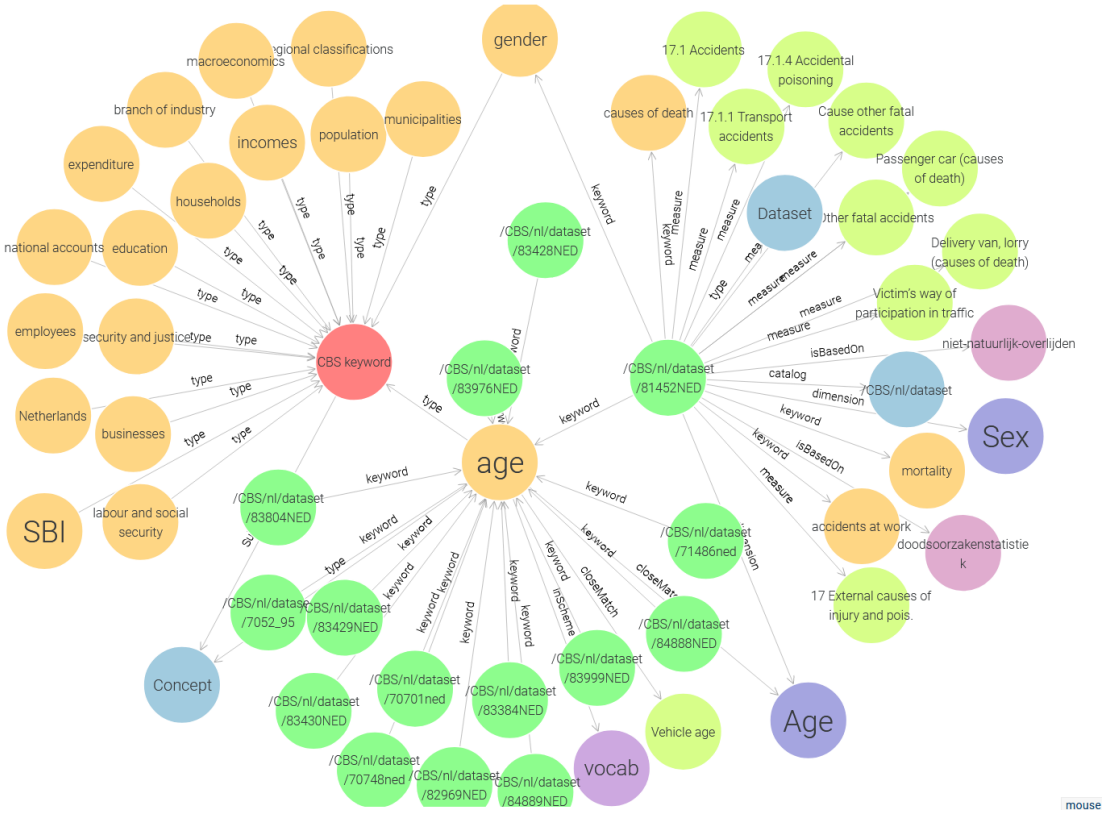

Figure 2: Part of knowledge graph linking datasets to metadata concepts, such as age, sex, mortality, gender, etc.

3.2. Linked data at Scotland's official statistics

The Scottish Government provides official statistics about Scotland from a variety of data producers on https://statistics.gov.scot. Data is published as data cubes organized by theme, organization and geography. It provides a search interface by subject or area, the ability to slice through datasets or to download as CSV or triples. The whole graph can be accessed programmatically via SPARQL. There is also an R package *opendatascot* that provides easy access to R users. Although the system may look a bit complex to the average user and is unfortunately sometimes slow, this example does show the added value of modeling metadata and data in an integrated linked data graph. Every element – from data cubes to slices to observations and their connected metadata - has an URI and links to other (metadata) elements, making advanced cross-dataset or metadata-driven queries possible.

3.3. Linked data at Eurostat

Eurostat linked data is published in the EU Vocabularies website. It contains statistical classifications such as NACE, CPA, and ECOICOP and regional classifications such as NUTS. The primary model used for describing the statistical classifications is XKOS. The Eurostat vocabularies site also contains content such as the European Business Statistics Manuals, the Catalogue of ESS standards and the ESS quality Glossary. They can all be explored via the ShowVoc system, a visual interface to browse the content manually that also offers a SPARQL interface. The content is supported in many different languages.

3.4. What to conclude?

Although we only described three statistical institutes offering linked data, the examples do show some interesting insights. In all cases linked data was used to describe (semi-)structured content such as classifications, concepts, terms, definitions, glossaries, in such a way that they can be linked to from other (web) content and across organisations. This is actually the core use-case for linked data, and applicable to the ESS to link metadata from ESS institutes together.

We see different approaches to the use of linked data techniques on data level. Although standards exists such as RDF Datacube [21] (partly based on SDMX) and CSV on the web [22], and large data volumes seem to be manageable as modern triple stores can easily store billions of triples, only the Scottish apply linked techniques up to the data level. In the Dutch example metadata is linked to data on dataset level.

Another observation we make is that the linked data model offers quite some flexibility. In the Dutch example it was possible to add new concepts and relationships whenever necessary to the existing knowledge graph that therefore became more powerful over time. This shows a degree of flexibility that is not visible in other statistical metadata standards that may take years of standardisation meetings to add some new features.

Furthermore, the Dutch example of modelling detailed geographical interdependencies among regions and their development over time shows the use of linked data for modelling complex statistical metadata dependencies. And finally, SPARQL offers the user a machine-readable access to all such linked data content.

The flexibility of linked data also has challenges. The models in the examples typically use the SKOS base layer for standard properties including exact matches between content. Differences occur in the use of extensions such as XKOS properties at Eurostat and extra classes from schema.org and others for geo modelling at the Dutch knowledge graph. Although via SPARQL federation [23] one can use both extensions in distributed queries, it would be good to harmonise approaches as much as possible. Another practical aspect is that knowledge graphs can easily become huge and difficult to understand. The creation of so-called 'data stories' that show how the rich content can be retrieved for practical purposes may help and would make official statistics landscape stronger on the FAIR interoperability and reusability dimensions.

From the perspective of the evolving digital society, we can see that AI companies and other content consumers scrape the web in their hunger for training data. The use of dedicated statistical standards makes the chance of being picked up smaller than if widely accepted models and formats are being used. On the long run an official statistics landscape that is easily and correctly interpreted could steer AI engines and other data consumers in the right direction, with the net effect of providing the public with trusted data instead of fake alternatives found in other places. Although this is highly speculative, we expect that the growing use of linked (meta)data in official statistics will positively influence standardisation and might open up new possibilities to spread statistical knowledge and content in society that needs trusted official statistics.

4. FAIR Digital Objects

The emerging FAIR Digital Objects (FDO) standard is an attempt from a community of scientists and leading data standardization organizations [5] to turn the internet into a meaningful data space. It follows-up on the original idea of the Digital Object introduced by Robert Kahn in the 1990s as a basic entity of a digital system [24]. From there we have seen many refinements to or implementations of the original idea. With the advent of the FAIR principles and the rise of linked data the concept was further refined and is currently defined as "*An FDO is a unit composed of data and/or metadata regulated by structures or schemas, and with an assigned globally unique and persistent identifier (PID), which is findable, accessible, interoperable and reusable both by humans and computers for the reliable interpretation and processing of the data represented by the object*" [25].

An FDO has four layers: 1) a PID System which resolves a global unique resolvable and persistent identifier (PID) into its PID Record, 2) an FDO record which defines a machine-readable, interpretable, and actionable structure that specifies the FDO's type, its metadata, its machine interpretable FDO kernel attributes, and its sets of bit sequences or references to bit sequences where the data is, 3) the FDO resource layer provides access to external resources in a distributed landscape and 4) a set of mandatory and recommended attributes, including a mandatory type which is a short hand for the characteristics of the FDO.

An important aspect of the FDO concept is *machine actionability*, which is defined (simplified from [26]) in 3 steps: 1) "machine readable" are those elements that are clearly defined by structural specifications, 2) "machine interpretable" are those elements that are machine readable and can be related with semantic artefacts in a given context 3) "machine actionable" are those elements that are machine interpretable and belong to a type for which operations have been specified in symbolic grammar.

For further details we refer to the full FDO specification. It is too early to implement the emerging FDO standard in the official statistics landscape, but it is certainly useful to view the official statistics landscape through the eyes of the FDO model and see how it relates to basic FDO concepts such as a global PID system and machine-actionability.

4.1. The Official statistics landscape from an FDO perspective

From an FDO perspective we can identify FDOs on multiple levels in the landscape [27]. Abstracting away from specific implementations, at the core of every official statistics provider we find the *statistical estimate*: 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 Limburg (NL) on Jan 1, 2022, or the inflation in Belgium for energy in 2021 are both statistical estimates. Statistical estimates typically have a production status (provisionary, final, revised) and are usually presented in – possibly large - multi-dimensional tables that offer selection, filtering and drill-down functionality. The underlying statistical database might be accessible by an open API and in chapter 2 we have seen that there are numerous software packages that help access it.

To provide meaning every statistical estimate links to metadata essential to understand its context. We often make a distinction into *structural* or *conceptual metadata*, i.e. the structure and definitions of concepts, dimensions and types of data used, (for example region, time, subject, population,) and *referential metadata*, i.e. descriptive information on the dataset (for example uncertainty, unit of measure, reference period, confidentiality, quality, accuracy etc.), see the Single Integrated Metadata Structure (SIMS) standard). Commonly agreed structural metadata is in the official statistics landscape currently organized into SDMX registries (global registry, Eurostat registry), which are accessible via their SDMX APIs.

Metadata have their own dynamics. Classifications and definitions change over time. For example, the Dutch administrative regions change almost every year and even relatively stable metadata such as a gender classification have to be adapted to modern times.

Statistical estimates and their dynamic metadata form the foundation for higher levels statistical output. News releases and thematic articles highlight specific aspects in a broader context. Infographics and data visualizations make it understandable to the larger public. This higher-level content can be viewed as digital objects too as it is usually the main entry level for the general public and search engines making FAIR aspects crucial as well.

A special category in the official statistics landscape is microdata. Not open to public, but accessible for research and policy making under strict conditions. In general, the metadata of the microdata is open in a catalogue.

The layered architecture of the official statistics landscape and the most common standards used on each layer is shown in Figure 3.

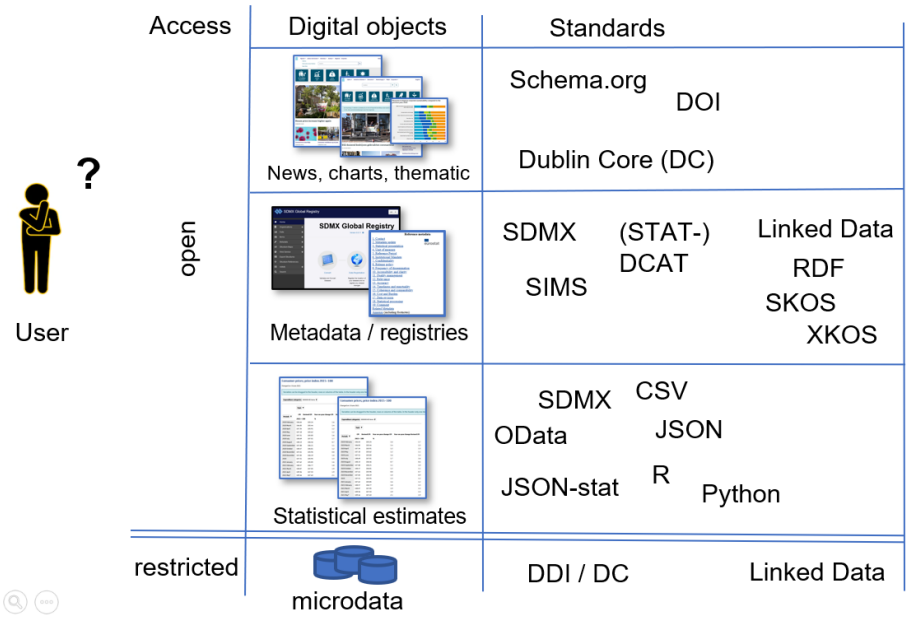

Figure 3: Layered official statistics architecture and standards used.

When we apply the FDO PID system to the official statistics landscape we see many different standards at different levels which have their own ways to identify elements, such as URNs, SDMX ID's, DOIs[28], URI's or organization specific identifiers. Although they probably all satisfy FAIR principle A1, from a user perspective it is better to minimize variety. Taking it to the level of API endpoints at this very moment there is no centrally maintained list of API endpoints for official statistics. Taking it further down, individual statistical estimates typically do not satisfy the PID concept. The multidimensional tables and their slices are typically identifiable, but at the statistical estimate level there is usually no globally unique and persistent ID. For higher level content, DOIs are used to provide stable access to scientific publications or standardization documents. For metadata that is already described via linked data (see chapter 3) each item has an URI by definition, which satisfies the FDO PID system. Internationally harmonized classifications maintained in international registry also fulfil this requirement, as long as they are in SDMX or in linked data, or both.

Machine-actionability of content in the official statistics landscape is more difficult to grasp. Statistical facts and metadata in whatever format are mostly machine-readable and, in some cases, also machine-interpretable, especially if linked data techniques are used. Reaching the level of machine-actionability would require a symbolic grammar to be defined on the content. Taken literally the statistical community would define what is allowed to do with the content, which we don't think is in line with core primitives in free and open statistics. Therefore, we can imagine that it could be interpreted in a more relaxed and helpful way as a set of readily available operations that have been defined and implemented to help the end-user understand and process official statistics. This include providing pointers to services for intelligent semantic context-aware interpretations of statistical content. Luckily, this is exactly what happened with the community generated software for access to official statistics described in chapter 2. In other words: providing good software in that category makes official statistics more actionable, which could be a driver for future improvements.

4.1. Future perspective

The FDO standard is not fully defined and at this moment not implementable. However, it is useful to view the official statistics landscape in terms of FDO concepts. Doing so might help improving on FAIR dimensions and aligning to current and future scientific communities. As we do, we can see that openness of statistics is not enough, the estimates, structural and referential metadata and higher-level statistical content should ideally all be optimized from a FAIR point of view.

Global unique resolvable and persistent identifiers should be available on all layers and to make the content actionable it should be aligned and supplemented with ready-available and easy to use software to access, understand and process statistics content.

Confronting the FDO model to the official statistics landscape also highlights official statistics elements that seem ahead of the FDO model. Standardization and coordination of metadata is strong in statistics, but at this time, absent in the FDO thinking. This is an essential element that should be maintained and even made stronger in future. Positively thinking, the official statistics community could maybe help the FDO model develop on this aspect to arrive at a situation where both become an integral part of a future society where internet turned into a meaningful data space.

5. Conclusion

In this paper we reviewed the official statistics landscape from different perspectives.

First of all, via the awesome list of official statistics software we looked at the software available to access official statistics. It shows that the official statistics landscape grows towards standardization on SDMX, JSON-STAT and PX. The most common features offered to users in such software are: *endpoint hiding, catalogue retrieval, search, endpoint queries, local queries, caching, cartographic queries* and *registry access*. It is still common use to develop dedicated software targeted at specific functionality or specific data providers, which results in many different software solutions to access official statistics, each offering different functionality and targeted at different parts of the official statistics landscape. It would be good if the official statistics community works towards a generic interface and software to *all* official statistics data providers, notwithstanding the value of the dedicated software. The analysis in this paper could serve as a start. Finally, a confrontation from software features to FAIR principles showed evidence that there is weakness on interoperability and reusability.

The second perspective is the ongoing work on linked data within the statistical community. At Statistics Netherlands a knowledge graph has been created by mining internal (metadata) systems and translating expert knowledge. This graph contains a hierarchical taxonomy that supports the website and the search engine, national classifications on education and economic activities with links to their international counterparts on the Eurostat linked data, all frequencies used at Statistics Netherlands with links to official codes, and multiple complex geo classifications one of which is a complete historical graph of Dutch municipality changes spanning two centuries up to present. Scottland provides official statistics as data cubes up to the data level which makes advanced cross-dataset or metadata-driven queries possible. Eurostat makes statistical classifications, manuals, a catalogue of standards and a glossary available as linked data on the EU vocabularies site.

From these examples we learn that using linked data creates flexibility, opens up modeling power, and adds the ability to link content across statistical domains, organizations and to wider communities. It can well be used to link national to international ESS metadata and to strengthen the interoperability and reusability in wider communities. Machine-readability is guaranteed through SPARQL. However, the flexible model makes it possible to model content differently, which creates the need to harmonise approaches across statistical organisations. Also, knowledge graphs may become large and difficult to understand. The creation of 'data stories' showing how to use the linked data may help. All in all, in an evolving digital society where AI companies and other content consumers scrape the web for training data, the use of commonly accepted linked data standards makes the chance of trusted data being picked up higher. Hence, we expect that the growing use of linked (meta)data in official statistics will positively influence standardisation and might open up new possibilities to spread statistical knowledge and content in society that needs trusted official statistics.

The third perspective is through the emerging FDO standard. This brings the insight that the estimates, structural and referential metadata and higher-level statistical content should ideally all be optimized from a FAIR point of view. Global unique resolvable and persistent identifiers

should be available on all layers and to make the content actionable it should be aligned and supplemented with ready-available and easy to use software to access, understand and process statistics content. On the aspect of standardization and coordination the official statistics community could complement the FDO thinking.

To the question raised in the title we conclude that for the official statistics landscape to be FAIR, there is work to do on all perspectives: aligning data access standards, improving (community) software, a stronger adoption of linked data standards, and keeping up with emerging concepts such as FDOs to prepare for the future meaningful data space. Only this way official statistics can play the role of trusted data partner also in future society.

Acknowledgements

The authors want to thank all colleagues from Statistics Netherlands and other statistical institutes that we worked with over many, many years on subjects related to this paper. We could not have written this paper without the international projects we were involved in, the experiments we carried out together and the discussions we had over time.

We also stress that the views expressed in this paper are those of the authors and do not necessarily reflect the policies of their institutes.

References

[1] European Statistical System Committee (ESSC), "European Statistics Code of Practice", revised edition 2017, doi: 10.2785/798269.

[2] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[3] Awesome list of statistical software for creating and accessing official statistics, https://github.com/SNStatComp/awesome-official-statistics-software

[4] Jacobsen A, et al; FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2020; 2 (1-2): 10–29. https://doi.org/10.1162/dint_r_00024

[5] FAIR Digital Objects Forum, https://fairdo.org

[6] ten Bosch, O., van der Loo, M., Kowarik, A., (2020), The awesome list of official statistical software: 100 ... and counting, 7Th international uRos conference, 2020

[7] Generic Statistical Business Process Model (GSBPM). v.5.1 – Jan.2019. UNECE. https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1

[8] Statistical Data and Metadata eXchange (SDMX), https://sdmx.org/

[9] JSON-stat: A simple lightweight standard for data dissemination, https://json-stat.org/

[10] PC-Axis software family, https://www.scb.se/en/services/statistical-programs-for-px-files

[11] Open Data Protocol (OData), https://www.odata.org/

[12] ten Bosch, O, van der Loo, M., Quality assurance from an internationally standardized and generic data validation ecosystem, Q2022, Vilnius, Lithuania, June 2022

[13] Berners-Lee, Tim; Fischetti, Mark (1999). Weaving the Web. HarperSanFrancisco. chapter 12. ISBN 978-0-06-251587-2.

[14] Berners-Lee T. "Is your linked open data 5 star". Berners-Lee, T. Linked Data. Cambridge: W3C. 2010.

[15] ESSnet on "Linked open statistics". 2017-2019. https://cros-legacy.ec.europa.eu/content/essnet-linked-open-statistics

[16] DCMI Usage Board (2006). DCMI Metadata Terms. Dublin Core Metadata Initiative. http://dublincore.org/documents/2006/12/18/dcmi-terms

[17] Albertoni, R., et. al. , P. (2014). Data Catalog Vocabulary (DCAT) (W3C Working Draft, 07 March 2023). https://www.w3.org/TR/vocab-dcat-3

[18] Isaac, A., Summers, E., SKOS Simple Knowledge Organization System Primer, Editors W3C Working Group Note 2009. Latest version http://www.w3.org/TR/skos-primer
[19] XKOS – An SKOS extension for representing statistical classifications. W3C 01 May 2019. Cotton, F., et. Al. . https://rdf-vocabulary.ddialliance.org/xkos.html
[20] Suominen, O., Ylikotila, H., Pessala, S., et. Al. (2015). Publishing SKOS vocabularies with Skosmos. June 2015
[21] RDF Data Cube Vocabulary, W3C Recommendation 16 Jan 2014, https://www.w3.org/TR/vocab-data-cube/
[22] CSV on the Web: A Primer, W3C Working Group Note 25 Feb 2016, https://www.w3.org/TR/tabular-data-primer/
[23] SPARQL 1.1 Federated Query, W3C Recommendation 21 March 2013, https://www.w3.org/TR/sparql11-federated-query
[24] Kahn, R., & Wilensky, R. (2006). A framework for distributed digital object services. International Journal on Digital Libraries, 6(2), 115–123. https://doi.org/10.1007/s00799-005-0128-x
[25] Ivonne, Anders et al., 2023, FAIR Digital Object Technical Overview: doi:10.5281/zenodo.7824714
[26] Claus, Weiland, Sharif, Islam, Daan, Broder (2022) FDO Machine Actionability. https://doi.org/10.5281/zenodo.7825650
[27] ten Bosch O, de Jonge E, Laloli H, Laaboudi-Spoiden C (2022) FAIR Digital Objects in Official Statistics. Research Ideas and Outcomes 8: e94485. https://doi.org/10.3897/rio.8.e94485
[28] ISO 26324:2012, Information and documentation – Digital object identifier system.. 2016.

Annex I

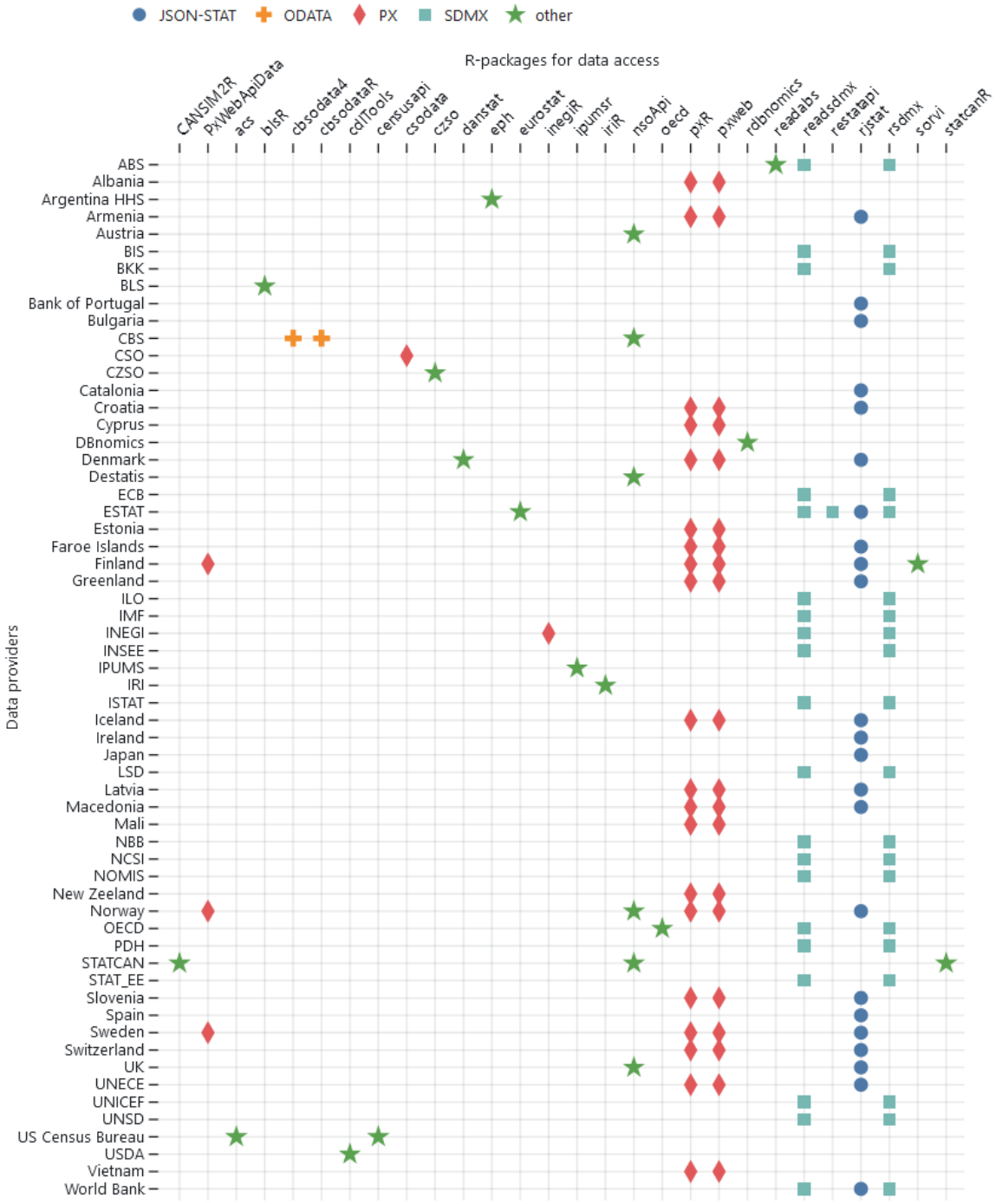

Figure Annex-1: R-packages for data access, statistical data providers and standards

Online version of this table, which will be updated regularly:
https://observablehq.com/@olavtenbosch/access-to-official-statistics-from-r