Statistical Journal of the IAOS 40 (2024) 91–96
DOI 10.3233/SJI-240020
IOS Press

Open and FAIR: Trends in scientific publishing and the implications for official statistics

Arofan Gregory*

Abstract. The FAIR data principles have emerged as a major focus in the world of scientific research data, but have not had as large an impact on official statistics. While there are good reasons for this, FAIR developments within the research community may be of interest to official statistical organizations. These include the increased availability of research data, improvements in the area of machine-actionable metadata, and a focus on provenance information which could lead to increased transparency and data quality. Some activities of interest are described as a starting point for those in official statistics who may wish to follow these developments.

1. Introduction

The FAIR data principles have had a major impact on how the scientific research community views the role of data. Instead of being a supporting asset, made available when required to validate research publications, data has become a primary output of the research process. Data is increasingly viewed as a potentially reusable resource – an asset resulting from the work of scientific research.

The same cannot be said of the impact of the FAIR data principles in the world of official statistics. As a result of the mission and motivation driving the work in these communities, FAIR may not seem to be as relevant in official statistics, and thus does not attract the same amount of attention. There are good reasons why official statistical organizations should pay attention to the FAIR data principles, however, even if they do not have the primacy that they do in the world of research. "FAIR data" and attendant developments in the research community have much to offer official statistics in the pursuit of its own missions, different as they are. This paper presents some of the arguments why FAIR and the attendant developments in the research community around data are something that may be of interest to official statistical organizations.

2. FAIR and the official statistics community

The FAIR Data Principles[1] were published almost a decade ago and since that time have attracted a huge amount of attention, being massively cited, and driving a great deal of expenditure on data management and dissemination in Europe, the US and elsewhere. The principles were conceived at a workshop at Leiden University, where the GO FAIR Foundation[2] acts as a steward for maintaining and publicizing them. They assert that research data should be "Findable, Accessible, Interoperable, and Reusable." While broad acceptance of these principles exists within the research community, the realization of these principles has taken longer to manifest, and for understandable reasons: it demands a metadata-intensive focus on data management which did not traditionally exist within the scientific com-

*Corresponding author: CODATA, 5 Rue Auguste Vacquerie, Paris, France. E-mail: ilg21@yahoo.com.

[1] Wilkinson et al., The FAIR Guiding Principles for scientific data management and stewardship, Nature: Scientific Data, March 2015, https://doi.org/10.1038/sdata.2016.18.

[2] GO FAIR Foundation [https://www.go-fair.org/] [Accessed 7 February 2024].

munity, and it emphasizes the widespread adoption of standards to support interoperability on many levels.

One reason that FAIR has attracted so much attention is that it presents a shift in attitude toward data within the scientific research community. Science is driven by research findings, emphasizing publications rather than on the data used to support those findings. Researchers and research organizations are rewarded by performing respected, high-impact research, and not primarily on the data they produce. While there are many exceptions, this is the general pattern we can see within the scientific community.

This shift can be explained if we examine some aspects of how data has impacted modern research, however. In a world where many research topics are inherently cross-domain (climate change, urban sustainability, disaster risk and response, infectious disease, etc. – sometimes described as the "grand challenges"), the effort required to prepare data for analysis within large research projects involving many organizations and disciplines can consume as much as 80% of the project budget[3] (EU Publications Office). If the costs could be lowered, the scientific community would benefit enormously. At the same time, many analysis techniques (such as those employing AI) have become "data hungry," presenting research projects with an unacceptable level of expense. In official statistics, these dynamics are different. Data is seen as a primary output, and although the cost of collecting and processing data is high, it does not stand as a barrier to the perceived mission of the organization.

We see this in the relative maturity of official statistics in these areas: the level of attention and resources given to data management, data dissemination, and the metadata needed to support these activities is higher. Further, because official statistical organizations often have major data reporting responsibilities, and a broad user base (policymakers, students, journalists, businesses, etc.), there has been a focus on standardization, as we can see in such collaborations as the Statistical Data and Metadata Exchange Initiative (SDMX).

Although not literally true, it would be understandable to think that official statistics has "always been FAIR," as the idea of making data broadly available in a useful form is not a new one. It is also true, however, that the FAIR phenomenon in the world of scientific research is something that can benefit the official statistics community, and which is worth paying attention to and participating in.

3. Machine-actionable metadata

Many aspects of the data landscape have changed in the recent past, both for scientific researchers in some disciplines, and for official statistics. The sources of data have assumed a greater variety, with survey data being supplemented with data from administrative registers, business transactions, social media, and an increasing array of automated systems that collect data to function. We can see this in the social sciences, for example, where "computational social science" has received a lot of attention, and in the official statistics world, where the HighLevel Group on the Modernization of Official Statistics (HLGMOS), based in the United Nations Economic Commission for Europe (UN/ECE), has addressed this theme.[4]

At the same time, the demand for data has grown: this pressure is felt by official statistics and scientific research infrastructures alike. "Big data" technology, based on massively scalable no-SQL databases and similar technologies, has enabled the development of analysis methods that can consume huge amounts of data, without requiring the use of super-computers. Various developments in artificial intelligence such as the use of large language models demand large bodies of data to function effectively. In many cases, official statistics are used to provide context for more specific data in scientific research or are used to identify causal relationships to inform broad-based analysis. Ideally, these techniques consume not only the data, but also the attendant metadata.

This presents the producers and disseminators of data with a challenge: traditional methods of data documentation – often largely manual – are insufficient to support the growing amount of data and the demand for it. To provide sufficient metadata, it is necessary to use production systems and software which can capture or mine the metadata programmatically. In general terms, machines want fine-grained metadata to support the new analysis methods: documents describing data at a general level are not sufficient for direct consumption

[3] EU Publications Office. "Cost of Not Having FAIR Research Data", March 2018, [https://publications.europa.eu/resource/cellar/d375368c-1a0a-11e9-8d04-01aa75ed71a1.0001.01/DOC_1] [Accessed 7 February 2024].

[4] High-Level Group for the Modernization of Official Statistics (HLGMOS) [https://unece.org/statistics/networks-of-experts/high-level-group-modernisation-statistical-production-and-services] [Accessed 7 February 2024].

by machines. One example of this is how statistical classifications are published: where a human researcher could use a PDF detailing the classification, a machine demands a format that is both machine-actionable and standard. In the scientific world, we see a parallel phenomenon around ontologies and controlled vocabularies of different types.

This topic – the creation of granular, machine-actionable metadata – is of great interest within the FAIR community, and there are many ongoing developments that promise to help address the need. Among these are the work being done by CODATA and the International Union for the Scientific Study of Population (IUSSP) around FAIR vocabularies, and publications like the "Ten Simple Rules for making a Vocabulary FAIR".[5] This work has even led to recommendations to the SDMX Initiative[6] about how they disseminate harmonized metadata for reuse, and some budding collaboration between the scientific community and official statistical organizations, although these are in the early stages. Notable here is the work on EuroVoc[7] by Eurostat and Caliper by the FAO,[8] where approaches to the use of common metadata standards are being explored with organizations from the FAIR community.

There is no easy solution to scaling up the documentation of metadata at a sufficiently granular level to meet increased demand, but there are many approaches being explored in the scientific research community which could be of benefit to the producers and users of official statistics as well. Collaboration here is in the interest of both communities.

4. FAIR data as a potential resource

One promise of the focus on FAIR in the scientific research community is the availability of significant amounts of research data in a more easily accessible form across many different domains. The emphasis within the FAIR community is on data and resource sharing, with secondary use of data and reproducibility of findings both being considered. But for the official statistical community, this may offer something different: a new source of data which can be used to support traditional production. There are some specific places where the increased availability of scientific research data might be useful, but they are not necessarily obvious, and there are some barriers to doing so.

First, the technical standards used by the official statistical community are not always the same as those used within the FAIR implementations in the scientific community, although there are some connections. SDMX is probably the most widely used technical standard for official statistics, but it is not used within research infrastructures, FAIR emphasizes RDF technologies, and SDMX does not, although the RDF Data Cube Vocabulary from W3C[9] is directly based on the SDMX Information Model. For classifications, we are beginning to see some use of SKOS[10] and XKOS[11] – RDF specifications – for describing statistical classifications, but these standards are used far more in the FAIR community than they are in official statistics. These FAIR standards would need to be supported if easy reuse of research data in official statistics in to be contemplated. The barrier here is only a technical one, and should not be difficult to overcome.

While finding and accessing FAIR data from the scientific community promises to become easier, with a low cost in resources, the coverage of such data, and the methods used to produce it, also present some barriers, and impact how it can be used appropriately. Unlike official statistics, research data is often geographically localized, with a strong depth of focus on a particular phenomenon of interest. Methods are likewise oriented toward the research question under consideration, which may not align with the purposes of typical official statistical data collection. It should be noted that there are some exceptions to this, however: as an example, the European Social Survey covers the whole of Europe and is conducted as a repeated series, looking at social attitudes and behaviors. The metadata standard used to document the data – DDI Lifecycle[12] – is also used within some national statistical agencies, and even such

[5] Cox et al., "Ten Simple Rules for making a Vocabulary FAIR", PLOS Computational Biology, April 2021, doi: 10.1371/journal.pcbi.1009041.

[6] IUSSP-CODATA Working group on FAIR Vocabularies, "FAIR Vocabulries in Population Research", April 2023 doi: 10.5281/zenodo.7818156.

[7] EuroVoc [https://op.europa.eu/en/web/eu-vocabularies] [Accessed 7 February 2024].

[8] Food and Agriculture organization, Caliper [https://www.fao.org/statistics/caliper/en#:∼:text=Caliper%20is%20the%20platform%20for,the%20dissemination%20of%20statistical%20classifications] [Accessed 7 February 2024].

[9] Data Cube Vocabulary [https://www.w3.org/2011/gld/wiki/Data_Cube_Vocabulary] [Accessed 7 February 2024].

[10] Simple Knowledge organization System [https://www.w3.org/2004/02/skos/] [Accessed 7 February 2024].

[11] Extended Knowledge organization System [https://rdf-vocabulary.ddialliance.org/xkos.html] [Accessed 7 February 2024].

[12] DDI Lifecycle [https://ddialliance.org/Specification/DDI-Lifecycle/] [Accessed 7 February 2024].

things as the geographic classification it uses is consistent with that of many European statistical agencies (NUTS, from Eurostat).[13]

There is a role for more localized research data in official statistics, however, although it may not be straightforward. Localized data can be employed as a way of supporting quality checks, for example. In one case in Malawi – where the national data has been insufficient for understanding the impact and organizing response to natural disasters at a local level – research data conducted by scientists studying public health can serve to baseline small area estimates, helping to improve the quality of data for some disaster-related purposes. Such techniques for enhancing data quality are not new in the official statistics world, but the availability of detailed microdata for employing them is decreasing as a result of FAIR, and the technologies needed to make better use of this data source are improving, as a result of AI techniques. (See Sam Clark's summary of his work at https://samclark.net/site/projects/small-area-estimates.shtml for a good example.)

It is not yet clear how the increased availability of scientific research data could benefit official statistics, but this is an area that is worth paying attention to. The COVID pandemic has given rise to many new data-sharing initiatives and platforms, and in general FAIR has emphasized the need for broad-based research infrastructure. In Europe we see the European Open Science Cloud (EOSC) being heavily resourced; in Africa, we see data-sharing efforts in public health such as VODAN[14] and the INSPIRE Network,[15] which are emerging as potential partners in the broader African Open Science Platform (AOSP).[16] While it is too early to know exactly how such infrastructures will take shape, there is a global movement toward large-scale collaboration in the research world, with a strong emphasis on FAIR data. If this data can be utilized to improve official statistics, especially in a resource-efficient way, then this may prove to help address the growing costs of data production for official statistical organizations.

5. Perspectives on data quality

Perhaps the single biggest impact that FAIR could have on official statistics is in the realm of data quality. The ideas of "data quality" in the two communities are very different, but some of the themes being pursued in FAIR are relevant to both sets of ideas. We will characterize quality as it has been approached in official statistics and in scientific research, and then look at how FAIR can impact these communities.

In the work around data quality, Official statistics has traditionally focused on consistent data production over time. Reporting frameworks such as the IMF's Data Quality Assessment Framework (DQAF),[17] Eurostat's Single Integrated Metadata Structure (SIMS),[18] and the various national frameworks of this type show that assessment of data quality is performed according to a set of criteria that can be applied to official statistics across the board, building confidence in the consistency and comparability of the data, as well as such aspects of quality as timeliness. This is a critical perspective on data from the official perspective.

Scientific researchers have different definitions of quality. In some domains, the possibility of measurement error can be calculated, and this is a very specific metric for data quality – accuracy – which is specific to the methods used within a domain. In more general terms, data quality can be understood as "fitness for purpose", that is, whether the data is useful for answering the research question being investigated. Thus, there is no single set of criteria for data, as data that is useful for one experiment may not be suitable for another: data quality, like beauty, is in the "eye of the beholder." The implication of this is that the amount and granularity of metadata becomes a primary aspect of data quality: you cannot pre-assess the data for any given purpose, but you can provide sufficient information to allow the potential user to perform their own assessment.

This has led to a focus on provenance and data "context" in FAIR, which includes describing the sources of data, and the steps in their processing. A good example of this can be seen in the European Social Survey's "ESS Labs – Climate Neutral and Smart Cities".[19]

[13] NUTS Glossary [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Nomenclature_of_territorial_units_for_statistics_(NUTS)] [Accessed 7 February 2024].

[14] VODAN [https://www.vodan-totafrica.info/] [Accessed 7 February 2024].

[15] INSPIRE Network [https://aphrc.org/project/inspire-implementation-network-for-sharing-population-information-from-research-entities/] [Accessed 7 February 2024].

[16] African Open Science Platform [https://aosp.org.za/] [Accessed 7 February 2024].

[17] Data Quality Assessment Framework [https://www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm] [Accessed 7 February 2024].

[18] Single Integrated Metadata Schema [https://ec.europa.eu/eurostat/documents/64157/4373903/SIMS-2-0-Revised-standards-November-2015-ESSC-final.pdf/47c0b80d-0e19-4777-8f9e-28f89f82ce18] [Accessed 7 February 2024].

[19] European Social Survey Labs [https://www.europeansocialsurvey.org/esslabs] [Accessed 7 February 2024].

In this project, social attitudes about climate change – collected by survey – are integrated with actual measurements of temperature and air quality, coming from other sources (Copernicus and European Environmental Agency). To understand the data, researchers can trace back to the source, and see both a human-readable description and the code used in performing the data preparation and integration. This process metadata compliments the regular variable-level description which is also available for the data at every stage.

We can think of this as a very comprehensive form of data documentation for the end user performing research, and it is, but it is also can be understood from the perspective of transparency. These are not different sets of information: provenance is important both to transparency and to reuse. This notion of comprehensive data description – including rich provenance information – thatcould help official statistics expand their idea of data quality in a similar direction, in line with discussions about this topic within the statistical community. From the perspective of standards/models, technical tools, and metadata there is not a lot of difference in terms of requirements, and it may be possible for the two communities to collaborate on the description of provenance for heightened data quality and transparency. Although they may use different terms for these concepts, there are fundamental similarities in the information they need and how they use it.

6. Summary: Ongoing developments

The case being made above is that developments within the FAIR community may be of interest to those in official statistics. To evaluate the value for any specific organization, it is clear that more investigation is needed. To this end, a list of potentially interesting activities is provided here. Several different developments within the FAIR community are mentioned above, but it can be difficult for people in the official statistics community to know where best to look for information and new developments. Below are some projects and activities which provide a starting point for those who wish to explore further. There is a wide range of activities in this area, so what appears below should not be taken as a comprehensive listing.

WorldFAIR Cross-Domain Interoperability Framework (CDIF): The WorldFAIR initiative is an EU-funded project with a global scope. It looks at 11 case studies in different domains, exploring practical capabilities and requirements for FAIR implementation. FAIR is seen as operating within domains/scientific disciplines and also between and among them. CDIF is a minimum set of profiles of existing domain-neutral metadata standards, and common web-based technology approaches for implementing FAIR to support cross-domain exchange and reuse of resources. It is worth noting that the standards and models used in the official statistical community such as SDMX[20] have been considered as part of this analysis. Work on the use of SKOS and XKOS among official statistical agencies has served as a major input particularly for the description of code lists and classifications.

As of this writing, the first draft of the CDIF guidelines has yet to be published and is scheduled to be made available in the summer of 2024. Further development is anticipated. Among the standards being recommended are DCAT,[21] Schema.org,[22] SKOS/XKOS, DDI Cross-Domain Integration (DDI-CDI),[23] PROV,[24] ODRL,[25] DPV,[26] and the I-ADOPT Framework.[27] There is more information available at the WorldFAIR project site.[28]

FAIR Impact[29] is a European initiative that is focused on many different aspects of FAIR implementation, including interoperability, various domain case studies, etc. They have developed a FAIR Implementation Framework for assessing the "FAIRness" of an organization or infrastructure, and a catalogue of resources.

The European Open Science Cloud (EOSC) is a membership consortium organized to develop and support a pan-European research infrastructure across disciplines. The EOSC Portal[30] has been deployed to provide access to various services, data, and other re-

[20] Statistical Data and Metadata Exchange (SDMX) Initative, [https://sdmx.org/] [Accessed 7 February 2024].

[21] DCAT [https://www.w3.org/TR/vocab-dcat-3/] [Accessed 7 February 2024].

[22] Schema.org [https://schema.org/] [Accessed 7 February 2024].

[23] DDI-CDI [https://ddialliance.org/Specification/ddi-cdi] [Accessed 7 February 2024].

[24] PROC Ontology [https://www.w3.org/TR/prov-o/] [Accessed 7 February 2024].

[25] ODRL [https://www.w3.org/TR/odrl-vocab/] [Accessed 7 February 2024].

[26] Data Privacy Vocabulary [https://w3c.github.io/dpv/dpv/] [Accessed 7 February 2024].

[27] I-ADOPT Framework [https://www.rd-alliance.org/group/interoperable-descriptions-observable-property-terminology-wg-i-adopt-wg/wiki/i-adopt] [Accessed 7 February 2024].

[28] WorldFAIR Project [https://worldfair-project.eu/] [Accessed 7 February 2024].

[29] FAIR Impact [https://fair-impact.eu/] [Accessed 7 February 2024].

[30] EOSC Protal [https://eosc-portal.eu/] [Accessed 7 February 2024].

sources, and to information about the initiative. Some projects of note include the "Climate Neutral and Smart Cities" science project,[31] conducted as part of the now-completed EOSC Futures project; the EOSC Interoperability Framework;[32] and some of the EOSC Core Services.[33]

It should be noted that all of the initiatives above have a degree of cross-participation among their staff and institutions, and make efforts to keep their work aligned. Notably, the various "interoperability frameworks" are not duplicative, but to a large degree address different aspects of interoperability. All of the initiatives mentioned are still ongoing, so it is difficult to say with certainty where they will eventually sit relative to one another, but they are not being conducted in isolation, nor are they competitors.

GO FAIR Foundation "FAIR Implementation Profiles" (FIPs) are a key tool for evaluating an infrastructure, domain, or large organization in terms of FAIRness. The GO FAIR Foundation has been a major force in the promotion of the FAIR data principles, and they have developed several tools that implementers may find useful. The FIPs are perhaps the most popular of these – you can learn more at the GO FAIR site.[34] There is also a "FIP Wizard" under development which is one of the tools used by other FIR projects to support the creation of profiles (free, but requires registration).[35]

IUSSP/CODATA Working Group on FAIR Vocabularies[36] is a now-finished project to make recommendations regarding the dissemination of controlled vocabularies relevant to demographic studies. IUSSP is the International Union for the Scientific Study of Population, and they have partnered with experts from the academic social sciences, the UN system, SDMX, and elsewhere to produce a set of recommendations for better defining and providing controlled vocabularies of all types to researchers in both the scientific and official statistical communities.

[31] European Social Survey Labs [https://www.europeansocialsurvey.org/esslabs] [Accessed 7 February 2024].

[32] EOSB Interoperability Framework [https://eosc-portal.eu/eosc-interoperability-framework] [Accessed 7 February 2024].

[33] EOSB Core Services [https://faircore4eosc.eu/eosc-core-components] [Accessed 7 February 2024].

[34] FAIR Implementation Profiles [https://www.go-fair.org/how-to-go-fair/fair-implementation-profile/] [Accessed 7 February 2024].

[35] FIP Wizard [https://fip-wizard.ds-wizard.org/wizard/] [Accessed 7 February 2024].

[36] IUSSP-CODATA FAIR Vocabularies Working Group [https://iussp.org/en/iussp-codata-fair-vocabularies-working-group] [Accessed 7 February 2024].