FAIR Implementation Profiles for Social Science

Shuai Wang[1 [0000−0002−1261−9930]], Angelica Maineri[2 [0000−0002−6978−5278]], Navroop K. Singh[1 [0000−0001−9131−3528]], and Tobias Kuhn[1 [0000−0002−1267−0234]]

[1] Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands {shuai.wang |t.kuhn} @vu.nl, n.k2.singh@student.vu.nl
[2] ODISSEI, Erasmus School of Social and Behavioral Sciences, Erasmus University Rotterdam, 3000 DR Rotterdam, the Netherlands angelica@odissei-data.nl

Abstract. FAIR Implementation Profile (FIP) is a special kind of linked data that consists of questions and answers about communities' decisions about the use of resources regarding the FAIR principles. Some FIPs have been created to capture collective decisions by communities. However, FIPs have not been widely adopted by the communities in social science. In this paper, we explore how FIPs can capture the decisions of communities in social science by creating their FIPs and comparing them against existing attempts. The created FIPs could be used as a structured way of cataloging FAIR-related implementation, comparing the resources used across communities, as well as understanding the practice of FAIR principles. We perform our analysis by generating the FAIR Convergence Matrix and comparing the resources used and their use. Finally, we discuss the lessons learned and the limitations of this approach.

Keywords: FAIR Implementation Profile · Social Science · Nanopublication · Metadata · FAIR Convergence Matrix.

1 Introduction

The FAIR principles (Findable, Accessible, Interoperable, Reusable) [1] can be a useful framework for the sharing of data that maximum use and reuse for researchers and communities. A FAIR Implementation Community (FIC) is a self-identified organization sharing a common interest that aspires to the creation of FAIR data and services [2]. A FAIR Implementation Profile (FIP) captures the decision of a FIC on how to put the FAIR principles in practice [2, 1]. The FIP template consists of a set of 21 questions corresponding to the FAIR principles regarding datasets and their metadata. A FIP consists of the answers to the questions about how resources are used for the practice of FAIR principle. These resources are called FAIR Enabling Resources (FERs). Over the past few years, some FIPs have been created to capture collective decisions by the members of communities on the practice of FAIR principles. FIPs can be created using the FIP Wizard[3], a user-friendly interface, and can be published in the

[3] https://fip-wizard.ds-wizard.org/. Developed by the GO FAIR Foundation (https://ror.org/056j50v04).

format of nanopublication [3, 4]. The created FIPs can be used for comparing community decisions, serving as guides for practical FAIR data stewardship as well as understanding the trends in FAIR implementation [2, 4]. Well-established linked data techniques such as SPARQL queries can be handy for the analysis of published FIPs. More specifically, the FAIR Convergence Matrix has been developed to help compare the use of FERs in different communities [5].

FIPs have been used by several communities. In particular, the Environmental Research Infrastructures (ENVRI) community extensively used them to systematically assess the progress of FAIR implementation in their network[4]. Over the past years, despite the fact that many FIPs have been created [4], only a few of them refer to social science communities. In the transition towards data-intensive social science, the need for automated data linkages makes FAIR implementation a priority. FIPs can help social science communities document and establish FAIR standards, support seamless data linkage across communities, as well as help identify gaps and steer investments. As a new kind of resource, FIPs can also be used to facilitate easier adoption of community choices for Data Management Plans (DMP) (see an attempt to compare FIP and DMP in [6]).

In this paper, we create FIPs in social science and explore how FIPs can be used for the understanding and evaluation of decisions in FICs. **The research question of this paper is: what can be unveiled regarding communities' data management decisions by comparing FIPs and analyzing communities' use of FERs?** This paper makes the following two scientific contributions a) we provide three new FIPs in social science and update three existing FIPs; b) we analyze and evaluate FIP implementation across six social science communities by using the FAIR Convergence Matrix and more.[5]

The paper is organized as follows. The current status of FIPs and related work are included in Section 2. The description of new FIPs is provided in Section 3. Following that, Section 4 provides the detailed analysis including the FAIR Convergence Matrix. Finally, we present lessons learned and discuss the limitations of this approach in Section 5.

2 Related Work

We start our research by reviewing existing FIPs for FICs in social science. The GESIS Social Science Survey Research (GESIS SSSR) is a FIC that uses and generates the data for social science survey research. Its datasets are indexed in the GESIS data catalogue, a platform for disseminating quantitative data with a particular focus on survey research. Since the SSSR FIP was published using an outdated knowledge model (i.e. template of questions for FIP) we create an updated version that takes the changes into consideration. We consulted the authors of the FIP for ambiguous entries.

[4] ENVRI-FAIR is a subproject of the Environmental Research Infrastructure (ENVRI): `https://envri.eu/home-envri-fair/`.

[5] All the FIPs, FICs, and the extended tables for analysis can be found at `https://github.com/FAIR-Expertise-Hub/MTSR`.

More recently, FIPs have been adopted by the two-year project to advance implementation of the FAIR principles: *WorldFAIR: Global cooperation on FAIR data policy and practice*. Different case studies assisted by FIPs have been described in the report by WorldFAIR Project [7]. The report presents FIPs as a methodology for cataloging FAIR implementation decisions made by a specific community of practice, and all the WorldFAIR case studies [7] have developed at least one FIP towards the end of the project. In the framework of the WorldFAIR, two FIPs about social surveys have been described but not added in the FIP Wizard, thus not published as nanopublication. The corresponding communities represent the European Social Survey (ESS) and the Australian Social Survey International – European Social Survey (AUSSI-ESS). The two FICs and FIPs were thoroughly described in the Cross-national Social Science survey FAIR implementation case studies report [8]. ESS is a cross-national survey, measuring attitudes, beliefs, and behaviors across diverse populations in over 30 European countries. Similarly, AUSSI-ESS is a the implementation of the ESS survey conducted in an Australian context. The purpose of this community was to enable cross-continent comparisons to European countries to better understand the similarities and differences between the countries. For this study, two FIPs were created in the FIP Wizard for ESS and AUSSI-ESS respectively, based on the description in the report [8]. In summary, the above-mentioned FIPs are mostly about social survey research. However, the research in social science is much more diverse than that. Next, we create more FIPs for more analysis.

3 FIPs for Social Science

In addition to the FICs described in Section 2, we created three new FIPs working alongside three FICs. They are the SSHOC-NL Socio-Economic History (SEH), Media Content Analysis Lab scholars (MCAL), and LGBTQ+ Linked Open Vocabulary (LGBTQVoc) communities. As described in Section 1, the FIP Wizard can be used to create the new FIPs. The provided FIP template contains a questionnaire allowing community data stewards to specify the FERs as the answer to the questions. By using an interface with dropdown menus, created FIPs can be published as RDF files of nanopublication in the TriG format.

The SSHOC-NL Socio-Economic History community consists of researchers and data experts in the Netherlands who conduct research and publish data about social history and economic history research. Questions in the FIP were filled in two virtual meetings with members of the FIC. The FIP of MCAL-NL, a community of scholars in the field of communication sciences and media content in the Netherlands, was developed with the same approach.

The LGBTQ+ linked open data vocabulary community is a young community that develops, translates, maintains, and uses (multilingual) linked open vocabularies. Different from the previous two communities, its members are distributed across the world in various time zones, which makes it difficult to organize a series of virtual meetings. For this exceptional case, we drafted the first

version of FIP based on the documentation of Homosaurus,[6] the most representative project in the community. Following that we enriched the FIP in two ways: a) we added missing entries using the documentation of QLIT[7] b) we asked several active members of the community to enrich and check the answers to the questions in FIP according to their knowledge and experience.

4 Analysis

Next, we answer our research question by performing some comparative analysis of the FIPs. Due to the page limit, we include in Table 1 only that regarding each of the most representative FAIR principles with more full table in the supplementary material. They capture important decisions by communities about the repositories for (meta-)data publishing, protocol use, and licenses for reuse, respectively. Its first row below the header (F4 Data) deals with the different repositories used by FICs to store and publish their datasets. Each of the FIP, with the exception of LGBTQVoc, has designated repositories that are specific to the community. Only Dataverse and the DANS SSH data station are in the overlap of SEH and MCAL. The absence of a FER in LGBTQVoc is because the datasets are either not published or only available on their official websites. In contrast, the second row (A1.1 Metadata) shows that many FICs overlap regarding their metadata communication protocols. The common FERs include OAI-PMH, HTTPS, and REST API. The third row (I1 Data) shows that FICs use different formats for the representation of their data with little overlap despite that of SEH and LGBTQVoc. While the fourth (I2 Data) shows that FICs use different structured vocabularies, which could be a barrier to interoperability. The last row (R1.1 Data) shows that different licenses are used. Due to the sensitivity of data and its social impact, as for the LGBTQVoc community, Homosaurus is published with a very strict license. MCAL and AUSSI-ESS declare no implementation choice. SSSR claims the use of a license that is specific to their community. This reflects how much the license used differs across communities.

Due to the page limit, we selected rows from our FAIR Convergence Matrix [5] that best represent different roles of FER in different FIPs in Table 2. We modify the proposed matrix by specifying the corresponding FAIR Principle since we noticed that the same FER could play different roles when used as metadata and data, or used in multiple ways, or corresponds to different FAIR principles. For example, DDI-Codebook is more popular among social survey communities. The use of which has an impact on both F2 and R1.2 Data. As for F1 Metadata, DOI is less popular among communities using surveys intensively. We can also see that REST API is more popular than OAI-PMH as metadata protocol. Finally, we noticed that publishing data in RDF is not common in the selected social survey communities in comparison with the others.

[6] Homosaurus is a bilingual linked open vocabulary: `https://homosaurus.org/`.

[7] QLIT (`https://queerlit.dh.gu.se/`) a linked open vocabulary with a focus on literature with some terms translated from Homosaurus. It was selected due to its close relation to Homosaurus and the availability of its detailed documentation.

FAIR Principle	SEH	MCAL	LGBTQVoc	SSSR	ESS	AUSSI-ESS
F4 Data	Dataverse, DANS SSH data station (future use)	DANS SSH data station, Dataverse, OSF, Figshare, TriplyDB (future use)	-	GESIS Search	ESS Data Portal, EOSC Portal	ADA Dataverse FER
A1.1 Metadata	OAI-PMH, HTTPS, REST API	HTTPS, REST API	HTTPS, REST API	OAI-PMH	HTTPS, GraphQL API	HTTPS, REST API
I1 Data	OWL, RDF, RDFS, XMLS, HTML	-	JSON-LD, RDF, RDFS, N-Triples, Turtle, SKOS	-	Apache Parquet	SPSS, Stata, SAS, R, CSV
I2 Data	HISCO, AMCO, H-ICD, IDS	-	Homosaurus, QLIT, LCSH, LCDGT, RVM	-	ISO3166-1, ISO639-2, NACE Rev. 2, ISCO-08, NUTS 3, DDI Controlled vocabularies, CESSDA Vocabulary, ELSST	-
R1.1 Data	CC-BY-SA, CC-BY-NC	-	CC-BY-NC-ND 4.0	GESIS Usage Regulations 2018	CC-BY-NC-SA 4.0	-

Table 1. A comparison of decisions on the practice of the FAIR principle

FAIR Principle	FER	SEH	MCAL	LGBTQVoc	SSSR	ESS	AUSSSI-ESS
F1 Metadata	DOI	1	1	1	1	0	0
A1.1 Metadata	OAI-PMH	1	0	0	1	0	0
A1.1 Metadata	REST API	1	1	1	0	0	1
F2 & R1.2 Data	DDI-Codebook	1	0	0	1	1	1
I1 Metadata/Data	RDF	1	0	1	0	0	0

Table 2. Selected rows in the FAIR Convergence Matrix (1 indicates in use, 0 not)

FER	SEH	MCAL	LGBTQVoc	SSSR	ESS
MCAL	4,2,0,0 (6)	-	-	-	-
LGBTQVoc	2,2,4,0 (8)	1,2,0,0 (3)	-	-	-
SSSR	1,2,0,0 (3)	1,0,0,1 (2)	1,0,0,0 (1)	-	-
ESS	0,1,0,0 (1)	1,1,0,0 (2)	1,1,0,0 (2)	1,0,0,0 (1)	-
AUSSI-ESS	0,3,0,0 (3)	1,3,0,1 (5)	0,2,0,0 (2)	1,1,1,1 (4)	0,1,0,1 (2)

Table 3. Comparison of FERs across FICs (in each cell is the number of overlapping FERs between the FIPs for each aspect of F, A, I, R as well as their sum in parentheses; the color corresponds to the sum)

Finally, in Table 3, we summarize the FAIR Convergence Matrix with an accumulative analysis of these overlapping FERs. The entries in the table show that LGBTQVoc and SEH have the most common FERs while the decisions by community ESS differ most from the others.

5 Discussion and Conclusion

The creation of FIP can be time-consuming, especially when members of the communities are not very active or invested. The approach for the creation of

the LGBTQVoc FIP is limited due to its incomplete information and therefore can not be generalized. We added over 30 missing FERs to the FIP Wizard so there is no missing FER as an answer. As living documents, FIPs can be updated as projects and communities develop. With periodic updates, the emergence of different versions of FERs and FIPs can introduce inconsistency and redundancy that may confuse researchers and create barriers to the automated processing of FIPs. The results and insights of this paper can be used to steer investments and resources: for instance, after highlighting the low uptake of FERs tackling the I principle, some structured vocabularies were recommended to the MCAL community. Moreover, after realizing that the datasets of the LGBTQ+ linked open vocabulary community cannot be found in public repositories, some suggestions were given to the editorial board of Homosaurus.

Conclusion This paper provides some primitive study on using six FIPs for the analysis of decisions made by communities in social science. We studied the overlap and difference of the use of FERs and the impact on the practice of FAIR principles. In future work, we would like to study how FIPs in social science differ from other research fields and how multiple FIPs can be used by researchers when writing their DMPs to ensure that their choices align with those of the reference community.

References

[1] Mark D Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1 (2016), pp. 1–9.

[2] Erik Schultes et al. "Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence". In: *Advances in Conceptual Modeling*. Springer, 2020, pp. 138–147. ISBN: 978-3-030-65847-2.

[3] Tobias Kuhn et al. "Publishing Without Publishers: A Decentralized Approach to Dissemination, Retrieval, and Archiving of Data". In: *The Semantic Web - ISWC 2015*. Ed. by Marcelo Arenas et al. Springer International Publishing, 2015, pp. 656–672. ISBN: 978-3-319-25007-6.

[4] Barbara Magagna et al. "FIPs and Practice". In: *Research Ideas and Outcomes* 8 (2022). DOI: 10.3897/rio.8.e94451.

[5] Hana Pergl Sustkova et al. "FAIR Convergence Matrix: Optimizing the Reuse of Existing FAIR-Related Resources". In: *Data Intelligence* 2.1-2 (Jan. 2020), pp. 158–170. ISSN: 2641-435X.

[6] Kristina Hettne et al. "FIP2DMP: Linking data management plans with FAIR implementation profiles". In: *FAIR Connect* 1 (Jan. 2023), pp. 23–27. DOI: 10.3233/FC-221515.

[7] Arofan Gregory et al. *WorldFAIR Project (D2.1) 'FAIR Implementation Profiles (FIPs) in WorldFAIR: What Have We Learnt?'* Nov. 2022. DOI: 10.5281/zenodo.7378109.

[8] Steven McEachern et al. "WorldFAIR Project (D6.1) Cross-national Social Sciences survey FAIR implementation case studies". In: (2023). DOI: 10.5281/zenodo.7599652.