Towards Rigorous Data Upcycling using the FAIR Implementation Profile for the SSHOC-NL Socio-Economic History Community

Shuai Wang[1][0000−0002−1261−9930] and Angelica Maineri[2][0000−0002−6978−5278]

[1] Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
shuai.wang@vu.nl

[2] ODISSEI, Erasmus School of Social and Behavioral Sciences, Erasmus University Rotterdam, 3000 DR Rotterdam, the Netherlands
angelica@odissei-data.nl

Abstract. **Keywords:** FAIR Implementation Profile · socio-economic history · data upcycling · legacy datasets.

1 Introduction

Despite the wide endorsement of the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles [1] by research institutions and funders, the principles are implemented heterogeneously by researchers, communities, institutes, etc. To capture FAIR implementation by communities, the documentation of choices via FAIR Implementation Profiles (FIPs) was proposed as a solution to streamline the efforts of FAIR implementation [2]. A FIP consists of 21 questions asking which standards, resources and technologies are used to comply with each of the FAIR subprinciples (in some cases, split for data and metadata). A FIP can be compiled in user-friendly interface using the FIP Wizard[3] [3]. As a living resource, FIPs can be updated as projects develop.

The SSHOC-NL Socio-Economic History (SSHOC-NL-SEH) community is a sub-community of the Dutch Socio-Economic History community consisting of researchers and data experts from the International Institute of Social History (IISG), Utrecht University (UU), Radboud Universiteit Nijmegen, and the Vrije Universiteit (VU) Amsterdam) in the Netherlands. The community has strong links to the SSHOC-NL project[4]. They conduct research and publish data about social history and economic history research. The research output of members of the community includes the publication of historical and upcycled datasets, as well as their analytical results about social and economic history. The FIPs

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

[4] SSHOC-NL will officially start in January 2024 and it is a collaboration between ODISSEI and CLARIAH, the Dutch national infrastructures for, respectively, social sciences and humanities.

for the SSHOC-NL-SEH community along with five other social science communities have been published [4].[5] This paper explores how this FIP can be used as a community standard to guide data management within the community, especially for the upcycling of legacy data.

Some recent work indicates that these FIPs can be a helpful reference for data management, for example, as candidate answers to the questions in the Data Management Plans (DMPs) [5]. Inspired by this, we test if the FIP can suggest actions to take or improve data management decisions about the upcycling of the legacy datasets. More specifically, in this paper, we provide a proof-of-concept experiment using some legacy datasets from the SSHOC-NL-SEH community. We study the upcycling of legacy datasets and compare their current status with the FIP and provide suggestions and future steps to take for upcycling these datasets with consideration of aligning them with community standards that are captured by FIP. As a primitive study, we focus on the following aspects:

Findability: Persistent and resolvable identifier, search engine and repository.
Interoperability: metadata format, metadata knowledge representation language, the format for datasets, knowledge representation for datasets.
Reusability: the licence of datasets.

The study of accessibility can be more complicated and can be a topic in future projects.

The paper is organized as follows. Section 2 presents an analysis of the legacy datasets to be upcycled. Following that, Section 3 provides details of the creation of the SSHOC-NL-SEH FIP and demonstrates how it can provide suggestions for the upcycling of datasets. Finally, some discussion is included in Section 4.

2 Legacy Data

Legacy collections are often older materials that do not meet modern data curation standards and require considerable resources to be preserved for future research [6]. They were produced when there was little concern or knowledge about the sharing of data and reuse by future researchers [6]. There could be flaws and problems when examined against standards of modern best practices. Legacy data differs from dark data, which is not amenable to computer processing or not used. Such datasets are not maintained or used, which could be in a worse situation than legacy data. For this reason, in this study, we also take such datasets into consideration. The following five datasets were provided by members of the SSHOC-NL-SEH community. Table 2 provides an overview of these legacy datasets.

Volkstellingen[6] aims to provide access to all census tables published between 1795 and 1971. The project was completed in 1997. Since 2007, the dataset has

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

[6] `http://volkstellingen.nl/`

been published on EASY, a data archiving system by DANS (Data Archiving and Networked Services) [7]. Some datasets were further migrated to the newly DANS Data Station Social Sciences and Humanities (a.k.a. DANS Data Station - SSH) [8]. Due to its diversity in content (images, data, photos, etc.), the file formats include PDF, CSV, etc.

The Linked International Classification for Religions (LICR)[9] is a classification system that provides mappings to various other well-known systems such as IPUMS, NAPP, HL7. It is enriched with DBPedia descriptions. The dataset was reported to have been discontinued since 2017 (no update or maintainence since its first publication on Dataverse). Thus, it has the potential to become dark data. We therefore also include it in this study as a recent legacy dataset.

The History Of Work Information System (HISCO)[10] is a multilingual class scheme that provides historical occupational codes and classification of occupational activities. Since its publication in 2002, the dataset has been used by various projects and communities. Thus, some effort has been put into upcycling it. HISCO and related datasets have been published in the IISG Dataverse[11]. Despite all the updates, the original website[12] was maintained due to its detailed explanation of data curation, searching functionality, etc.

Gemeentegeschiedenis[13] (Municipal History) is a project by Hic Sunt Leones. It contains the boundaries of Dutch municipalities since 1812. Its datasets were published by Onno Boonstra and colleagues in 2007 and 2011[14]. 1542 datasets were published in the EASY repository by DANS but are being migrated to the DANS SSH Data Station[15]. The datasets were converted to GeoJSON and maintained in 2020 and published in IISG's Dataverse.[16]

Finally, the HDNG dataset (in English, the Historical Database of Dutch Municipalities) originated from the HED (Historisch-Ecologische Database) dataset. In its third version by Ruben Schalk, the data was converted to linked data and published on the Druid repository[17]. The latest version (v5.1, July 2023) was

[7] `https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:44159`

[8] `https://ssh.datastations.nl/`

[9] The official website (`www.licr.io`) is no longer accessible. Detailed description of the dataset is unavailable. The dataset is available on IISG's Dataverse: `https://hdl.handle.net/10622/MHJWRZ`.

[10] `https://historyofwork.iisg.nl/`

[11] `https://datasets.iisg.amsterdam/dataverse/historyofwork`

[12] It is worth mentioning that the errors on this website correspond to outdated data and can be misleading for users.

[13] `https://gemeentegeschiedenis.nl/`

[14] The DOI of the two datasets are `https://doi.org/10.17026/dans-xb9-t677` and `https://doi.org/10.17026/dans-xdr-cs36`.

[15] The data is not accessible in the SSH Data Station yet. Instead, in the ZIP file, the link is included: `https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:44426`.

[16] `https://hdl.handle.net/10622/URI802`

[17] https://druid.datalegend.net/dataLegend/HDNG

published on Dataverse. [18]. Table 1 shows the version history that indicates how the dataset received multiple contributions from various researchers/data experts.[19] HDNG is published with the licence CC0 1.0. However, HED, its predecessor, was initially published with its own licence, which specified several restrictions. The dataset is now open source.

Version	**Version History on Dataverse**	**Contributors**	**Date Published**
4	5.1	Ineke Kellij	Jul 2023
	5.0	Rick Mourits, Richard Zijdeman	Jul 2021
	4.0	Rick Mourits	Jul 2021
3	3.0	Richard Zijdeman	Feb 2021
	2.1	Richard Zijdeman, Ineke Kellij	Dec 2020
	2.0	Richard Zijdeman	Jan 2020
2	1.1	Richard Zijdeman	Feb 2019
	1.0	Richard Zijdeman	Mar 2016

Table 1. The version history of the HDNG dataset on the Dataverse of IISG

3 Rigorous Legacy Datasets Upcycling Using FIP

In the SSHOC-NL-SEH FIP, the identifier captured by the FIP are Handle and DOI (to be used in the future) for datasets. Dataverse and DANS SSH Data Station (to be used in the future) are the chosen search engines. XML, HTML, RDF are the knowledge representation languages captured in this FIP for its datasets. That for metadata includes RDF (OWL, RDFS), and SKOS. Finally, CC-BY-SA is being used by the community with CC-BY-NC to be used in the future. The last column in Table 2 provides a summary of these selected aspects of the FIP. Next, we show how this FIP can be used to provide suggestions on data management and future steps to take regarding the above-mentioned datasets.

The LICR has not been maintained since its publication. With the aspects assessed in Table 2, the dataset remains up to date with community standards captured by FIP. Its findability could increase if it were migrated to the DANS SSH Data Station, which can reduce the change to become legacy data or dark data.

The assessment of Volkstellingen shows that the current FIP misses URN as a persistent ID for datasets. Moreover, our examination shows that EASY[20]

[18] `https://hdl.handle.net/10622/RPBVK4` The corresponding code is on Github at `https://github.com/CLARIAH/HDNG`.

[19] Extracted from `https://datasets.iisg.amsterdam/dataset.xhtml?persistentId=hdl:10622/RPBVK4`.

[20] EASY was discontinued and is replaced by the DANS SSH Data Station.

Aspects	**LICR**	**Volkstellingen**	**HISCO**	**Gemeentegeschiedenis**	**HDNG**	**FIP**
Year created	2017	1997	2002	2007	1970	-
Year last maintained	2017	2007	2020	2020	2021	-
Dataset PID	Handle	DOI, URN	Handle	DOI, Handle	Handle	DOI(to use), Handle
Repository	IISG Dataverse	EASY, DANS SSH Data Station	IISG Dataverse	DANS SSH Data Station, EASY, IISG Dataverse	Druid, IISG Dataverse	Dataverse, DANS SSH data station (future use)
Metadata format	Dublin Core, DDI, DataCite, DDI Codebook, JSON, OAI_ORE, OpenAIRE, Schema.or JSON-LD	XML, CSV	Dublin Core, DDI, DataCite, DDI Codebook, JSON, OAI_ORE, OpenAIRE, Schema.or JSON-LD	Dublin Core, DDI, DataCite, DDI Codebook, JSON, OAI_ORE, OpenAIRE, Schema.or JSON-LD	Dublin Core, DDI, DataCite, DDI Codebook, JSON, OAI_ORE, OpenAIRE, Schema.or JSON-LD	-
Knowledge representation language for metadata	XML, Dublin Core (DCMI Metadata Terms), DataCite, HTML, JSON-LD	Dublin Core (DCMI Metadata Terms), EASY Metadata Terms	XML, Dublin Core (DCMI Metadata Terms), DataCite, HTML, JSON-LD	XML, Dublin Core (DCMI Metadata Terms), DataCite, HTML, JSON-LD	XML, Dublin Core (DCMI Metadata Terms), DataCite, HTML, JSON-LD	RDF (OWL, RDFS), SKOS
Data format	NQ, TAB	Various	CSV, TAB, PDF, TXT	Shapefile (.dbf, .shp, .shx), GeoJSON	RDF (Trig), CSV	-
Knowledge representation language for datasets	SKOS, RDF, RDFS	Not found by manual examination	None	None	RDF, SKOS	XML, HTML, RDF (RDFS, OWL)
Licence	CC0 1.0	Not found	CC0 1.0	CC BY-SA 4.0	CC0 1.0	CC-BY-SA, CC-BY-NC (new to the community)

Table 2. Dataset features compared against FIP

could be not captured in the FIP as a previously used search engine. There are many datasets in various formats in the Volkstellingen (corresponding to different years), and each consists of many files, which makes it difficult to perform a manual assessment of all the knowledge representation languages. Converting data into their corresponding RDF format would make them better align with community standards. Given that no licence is specified, we can suggest the use of CC-BY-SA as documented in the FIP.

Our examination shows that there are several upcycling attempts in which there is no redirection from the outdated to the more recent ones. Taking HISCO for example, Table 3 shows how the landing pages and repository pages changed over time. Our examination shows that this can lead to confusion because of the lack of references between them. Worse of all, the latest version cannot be found using the first five links. This shows the importance of a unified repository where the datasets are published. Recall that the FIP suggests the use of IISG's Dataverse and will also be using DANS SSH data station. The findability of datasets related to HISCO could be improved by providing links to location of future migration. Finally, the use of DOI and Handle are suggested by FIP but only the last page can be found by dereferencing the corresponding Handle PID. Moreover, we noticed that the licence used is CC0 1.0. However, the FIP suggests the use of CC-BY-SA for datasets.

Index	Page	URL	Comment
1	HISCO page	`https://historyofwork.iisg.nl/`	No link to the datasets, nor any other page in the table.
2	IISG HISCO page	`https://iisg.amsterdam/en/data/data-websites/history-of-work`	Link to 1, 3, 4, 5
3	History Of Work Information System page on IISG Dataverse	`https://datasets.iisg.amsterdam/dataverse/historyofwork`	Link to 4
4	HISCO page on IISG Dataverse	`https://datasets.iisg.amsterdam/dataverse/HISCO`	
5	The HISCO collaboration website	`https://datasets.iisg.amsterdam/`	Link to 3.
6	The HSNDB Occupations page on IISG Dataverse	`https://datasets.iisg.amsterdam/dataset.xhtml?persistentId=hdl:10622/88ZXD8`	The latest version of HISCO with the Persistent ID `https://hdl.handle.net/10622/88ZXD8`, which lands on IISG's page on Resources for HSNDB's Historical Sample of the Netherlands.

Table 3. Attempts for the upcycling of HISCO

As for Gementegeschiedenis, our examination shows that this is the only dataset whose licence is aligned with the community standard captured by the current FIP. Its data are shapefiles and GeoJSON files, which could be better as linked data, in RDF for example.

Finally, for HDNG, Druid was not mentioned in the FIP. The licence used is CC0 1.0. However, the FIP suggests CC-BY-SA, which is worth some discussion. SKOS is used in the data but missing in the FIP.

4 Discussion and Conclusion

Conclusion In this paper, we explored the benefit of FIP in data upcycling in the SSHOC-NL Socio-Economic History Community. The FIP, an instrument capturing past data management decisions, present choices, and future plans of a community, can guide the upcycling of legacy data by offering suggestions as to which standards and tecnologies should be adopted. For instance, it can give explicit indications on the repositories to use to store new versions of upcycled data, and also suggest what PID system should be used. In a fast-evolving data management landscape, this makes the upcycling efforts less dispersive. Moreover, this ensures the interoperability of the data over time within a community.

Discussion It should be noted that upcycling is a continuous process, and the FIP is a living document. As such, they are meant to inform each other. For instance, our examination highlighted how some resources in use by the community were missing from the FIP: this is especially the case for licensing, whereby license CC0 1.0 detected in legacy data will be added to the FIP as it is in use, cannot be revoked, and it is not an outdated standard. In this sense, future investigations should focus on the bidirectionality of the relationship between FIP and data upcycling, enabled by a critical examination of the standards used to process and document legacy data. In future versions, since the FIP also captures past decisions about data management, some FERs could be added as resources used in the past.

This paper also provides insights into the data management practice of the community. Datasets in the community are maintained by different researchers at different stages of the data lifecycle. Sometimes due to external reasons, datasets are moved to different platforms (e.g. following the decision of DANS to move from EASY to the Data Station SSH, which is based on Dataverse). In addition, our examination shows that tracking the entire version history for some datasets can be difficult. If more information is available, users should be notified about the changes made and where the errors have been fixed in the legacy datasets.

In this study, we manually assessed aspects regarding three principles out of four in the FAIR principles. The study of accessibility can be more complicated and thus left for future work. We also noticed that in some special cases, manually assessing each file is difficult, which calls for the design of an automated tool. Some more datasets under the process of upcycling could be assessed with a future version of the FIP.

Acknowledgement

The authors received kind help from Rick Mourits during the project. Richard Zijdeman, Sytze van Herck, and colleagues from the Vrije Universiteit Amsterdam have also contributed to the forming of ideas and reviewing of the paper.

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

[4] Shuai Wang et al. "FAIR Implementation Profiles for Social Science". In: *Preceeding of 17th International Conference on Metadata and Semantics Research*. Springer, 2023.

[5] 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.

[6] Werner Scheltjens. "Upcycling historical data collections. A paradigm for digital history?" In: *Journal of Documentation* (2023).