Project Title	Global cooperation on FAIR data policy and practice
Project Acronym	WorldFAIR
Grant Agreement No	101058393
Instrument	HORIZON-WIDERA-2021-ERA-01
Topic, type of action	HORIZON-WIDERA-2021-ERA-01-41 HORIZON Coordination and Support Actions
Start Date of Project	2022-06-01
Duration of Project	24 months
Project Website	http://worldfair-project.eu

D6.1 Cross-national Social Sciences survey FAIR implementation case studies

Work Package	WP06 – Social Surveys
Lead Author (Org)	Steven McEachern (Australian Data Archive, Australian National University)
Contributing Author(s) (Org)	Hilde Orten (Sikt.no), Hanna Thome Petersen (Sikt.no), Ryan Perry (Australian Data Archive, Australian National University)
Due Date	31.01.2023

Date	30.01.2023
Version	1.0 DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION
DOI	https://doi.org/10.5281/zenodo.7584438

Dissemination Level

- [X] PU: Public
- [ ] PP: Restricted to other programme participants (including the Commission)
- [ ] RE: Restricted to a group specified by the consortium (including the Commission)
- [ ] CO: Confidential, only for members of the consortium (including the Commission)

Versioning and contribution history

Version	Date	Authors	Notes
0.1	16.01.2023	Steven McEachern, Hilde Orten, Hanna Thome Petersen, Ryan Perry	Draft for internal review
1.0	30.01.2023	Steven McEachern	Content ready

Disclaimer

WorldFAIR has received funding from the European Commission's WIDERA coordination and support programme under the Grant Agreement no. 101058393. The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

Abbreviations and Acronyms

ADA	Australian Data Archive
API	Application Programming Interface
AUSSI-ESS	Australian Social Survey International – European Social Survey
CORDIS	Community Research and Development Information Service
DDI	Data Documentation Initiative

DOI	Digital object identifier
EOSC	European Open Science Cloud
ESS	European Social Survey
EVS	European Values Survey
FAIR	Findable, Accessible, Interoperable, Reusable
FER	FAIR-Enabling Resource
FIP	FAIR Implementation Profile
HEIs	Higher Education Institutions
ISSP	International Social Survey Program
NCI	National Computational Infrastructure (NCI)
WVS	World Values Survey

Executive Summary

This report provides an overview of the data harmonisation practices of comparative (cross-national) social surveys, through case studies of: (1) the European Social Survey (ESS) and (2) a satellite study, the Australian Social Survey International – European Social Survey (AUSSI-ESS). To do this, we compare and contrast the practices between the Australian Data Archive and Sikt.no, the organisations responsible for the data management of ESS and AUSSI-ESS.

The case studies consider the current data management and harmonisation practices of study partners in the ESS, including an analysis of the current practices with FAIR data standards, particularly leveraging FAIR Information Profiles (FIPs) and FAIR Enabling Resources (FERs).

The comparative analysis of the two case studies considers key similarities and differences in the management of the two data collections. Core differences in the use of standards and accessible, persistent registry services are highlighted, as these impact on the potential for shared, integrated reuse of services and content between the two partner organisations.

The report concludes with a set of recommended practices for improved management and automation of ESS data going forward—setting the stage for Phase 2 of WorldFAIR Work Package 6—and outlines the proposed means for implementing this management in the two partner organisations. These recommendations focus on three areas of shared interest:

1. Aligning standards
2. Establishing common tools
3. Establishing and using registries

in order to advance implementation of the FAIR principles, and to improve interoperability and reusability of digital data in social sciences research.

Table of contents

1. Introduction

This report, the first deliverable of WorldFAIR WP06 Social Surveys, provides an overview of the data harmonisation practices of comparative (cross-national) social surveys, through case studies of: (1) the European Social Survey (ESS) and (2) a satellite study, the Australian Social Survey International – European Social Survey (AUSSI-ESS). The report outlines the reasons for comparative social surveys such as the ESS and related initiatives, and the foundations of the European Social Survey and its satellite studies. It then continues to consider the current data management and harmonisation practices of study partners in the ESS. To achieve this, we provide an overview of the current practices with FAIR data standards, particularly leveraging the FAIR Information Profiles (FIPs) and FAIR Enabling Resources (FERs) to compare and contrast the practices between the Australian Data Archive and Sikt.no, the two partners leading Work Package 6 and the organisations responsible for the data management of ESS and AUSSI-ESS. The report concludes with a set of recommended practices for improved management and automation of ESS data going forward—setting the stage for Phase 2 of Work Package 6—and to outline proposed means for implementing this management in the two partner organisations. These recommended practices will also provide a foundation for establishing FAIR practices in similar studies in the social sciences and related domains into the future.

1.1 Project partners

WorldFAIR Work Package 6 is being lead by two organisations with extensive track records in the management of social science research data: the Australian Data Archive[1] at the Australian National University in Canberra, Australia; and Sikt, the Norwegian Agency for Shared Services in Education and Research[2] in Bergen, Norway.

The Australian Data Archive (ADA) was established at the Australian National University (ANU) in 1981 to provide a national service for the collection and preservation of digital research data. ADA disseminates this data for secondary analysis by academic researchers and other users, in Australia and around the world. It has a long history in the development of data archiving standards and practices, and was an early adopter of the Data Documentation Initiative (DDI), the dominant standard for the management of social science research data since its release in 2001.

ADA is based in the ANU Centre for Social Research and Methods (CSRM). The CSRM is a dedicated research and teaching centre within the ANU, providing four key activities:

- The development of social research methods
- Analysis of social issues and policy
- Training in social science methods
- Providing access to social scientific data

Sikt, the Norwegian Agency for Shared Services in Education and Research, develops, acquires and delivers services for education and research in Norway. It was established in 2022, from the merger of the former NSD (Norwegian Centre for Research Data AS), Uninett

[1] https://ada.edu.au
[2] https://sikt.no/

AS and Unit, the Directorate for ICT and Joint Services in Higher Education & Research. The data archiving services of Sikt come from NSD, which had a long history in research data archiving, established in 1971 by the Norwegian Research Council. In collaboration with users, Sikt offers a common infrastructure for education and research. The aim is to free capacity for customers, and to meet overarching goals of digitalisation, data sharing and open research.

There are several similarities in the organisational structures of both these data archives and their broader organisations. Both archives:

- are involved in collaborative international social survey projects, including the European Social Survey (ESS-ERIC) and the International Social Survey Programme (ISSP)
- provide data archiving services to a national and international research community
- conduct social survey research projects as part of their daily research and business activities

This co-location with a social survey unit, for ADA within the CSRM and for Sikt.no as two organisational departments, provides unique opportunities for studying the management and dissemination of social science data, as both are involved in the creation and collection of data, as well as its management and dissemination.

In addition, both organisations have had long history of participation in comparative social surveys: Australia was one of the founding members of the International Social Survey Programme (ISSP) in 1984, with Norway joining in 1989, and both countries have been members of the European and World Values Surveys (EVS/WVS) since the 1980s. This long-term commitment to social surveys as both a research methodology and a data sharing activity provides a sound foundation for studying and understanding data integration practices in this domain.

There is also a fourth area of common practice between ADA and Sikt. Both organisations are members of the DDI Alliance[3], the organisation responsible for the dissemination of the DDI standard, and have been involved in the development of the DDI-CDI standard. The foundations of DDI-CDI align closely with the requirements of the Cross-Domain Interoperability Framework (CDIF), a key activity of the WorldFAIR project, and the study of the practices within the two organisations should therefore provide insight into the implementation of DDI-CDI through the WorldFAIR project.

2. Comparative social surveys and the need for data integration

Cross-national social surveys (also known as comparative or cross-cultural surveys) have a long history in the social sciences. In a review of the long-term development of cross-national surveys, Smith and Fu (2015) identified three broad phases of cross-national survey development:

[3] https://ddialliance.org/

- Initial establishment (1930s to early 1970s), where public opinion polls were progressively established across countries, and then between small numbers of similar countries, particularly in Europe;
- Expansion (1973 to 2002), where studies expanded in both scope and breadth of countries, and were increasingly coordinated and sustained. Studies such as the ISSP and the World Values Survey were established in this period;
- Surveys as infrastructure (2002 to present), where "survey research became part of the social-science infrastructure … [and] the degree of central coordination and control notably increased" (Smith and Fu 2015, p.7).

This increased central coordination and control has led to the establishment of a broad set of recommended practices for the conduct and management of cross-country social surveys. These practice guidelines are often the result of a collaboration of international survey data practitioners, such as the Cross-Cultural Survey Guidelines[4] established by the Comparative Survey Design and Implementation group[5].

There has, however, been less work on the documentation and data integration requirements of cross-national surveys, particularly in the processes of conducting the data harmonisation itself. Many of the recommended practices for data management and integration of such studies were documented by Peter Granda at the Inter-University Consortium for Social and Political Research (Granda and Blasczyk 2010; Vardigan, Granda and Hoelter 2016). More recent work has however sought to address this gap, particularly Slomczynski, Dubrow, Tomescu-Dubrow and colleagues at the Ohio State University (OSU) and Polish Academy of Sciences (PAS) (Dubrow and Tomescu-Dubrow 2015). Early on, the OSU group identified a core problem in understanding and improving practices:

> "... it is clear that there is now a large methodological literature on SDH in the social sciences but there is no coherence across projects. The problem is that this literature exists in various documents scattered over place and time, and little of it has been synthesised into a manageable and accessible format. In short, there is a need to pull this literature together, to create a handbook of SDH in the social sciences based on the many existing efforts …"
>
> (Dubrow and Tomescu-Dubrow 2015, p. 16).

The OSU-PAS team has subsequently spent significant time in compiling current secondary documentation to improve this situation consistent with their recommendations. This has seen the establishment of the Harmonization Project and follow-on Survey Data Recycling project, and the subsequent publication of the Survey Data Recycling database[6], which provides pooled methodological information from 23 different cross-country survey projects. In a published report on this work, Tofangsazi and Lavryk (2018), summarised the variation in practices in three dimensions:

- File format variety
- Document location variety

[4] https://ccsg.isr.umich.edu/
[5] https://csdiworkshop.org/
[6] https://wp.asc.ohio-state.edu/dataharmonization/data/

- ‘Difficult cases’ associated with inconsistencies in methodological details such as documentation of survey response rates and target populations

Alongside this work, the OSU team has also sought to establish standards for cross-national research data quality and documentation[7], leveraging the DDI standard as the basis for this work, although the outputs of this work have not yet been published. The work of the OSU team is however suggestive as to where improvements in the data management process may be made. In particular, to address the inconsistency issues across projects that they identified in their study, Tofangsazi and Lavryk made three recommendations that would assist in improving transparency and reuse:

1. All documents should be available in PDF and HTML, so that they can be read across different computer operating systems and software. HTML might also enable automation of some of the more boring tasks that should not, in a perfect world, require humans to do.
2. Project websites should provide a clear description of all the survey documentation that they provide, and what languages the documents are available in. Ideally, project web pages would have a stable address such as one finds in GESIS and ICPSR.
3. Considering that international survey projects vary in terms of quality of documentation, a wider adoption of DDI standards might improve the situation, including the production of structured, machine-processable metadata

(Tofangsazi and Lavryk 2018, p.30).

While these recommendations are relatively simple in current terms, they do point to opportunities for machine-readable and actionable content. Notably, the suggestion for stable addresses, use of DDI standards, and the production of structured, machine-processable metadata provide clear opportunities for using FAIR data principles for driving some of the improvements required. It is on this basis that the case studies for the ESS and AUSSI-ESS have leveraged the FAIR Implementation Profiles (FIPs), in order to study the use of FAIR practices across the two partner archives, and to consider where these could be more closely aligned.

3. The European Social Survey and AUSSI-ESS

A brief description of the ESS and the conduct of satellite studies such as AUSSI-ESS follows here. The case studies for this report are based on the conduct of ESS Wave 9, which was the first wave to be conducted in Australia as a satellite study.

3.1 The European Social Survey

The ESS was established in 2001 to study the attitudes, beliefs and behaviours of the populations of European countries. In 2013, it was established as a European Research Infrastructure Consortium under EU legislation, and is one of the three ERICs in the social sciences. At time of writing, the ESS has undertaken 10 waves of data collection, with Wave

[7] https://consirt.osu.edu/research/standards-for-cross-national-research/

11 in the field and planning for a further wave currently underway, and now includes over 30 countries in the core programme.

The methodological requirements and project logistics for conducting the ESS in participating countries are outlined by the Core Scientific Team of the ESS, based at City University London. All documentation for each wave of the survey is available from the ESS website[8]. The core description of methods is detailed in the Survey Specification for each wave of the ESS, providing an overview of each stage of the process for country teams collecting ESS data. A summary of the main sections of the methodology is included in ***Figure 1.***

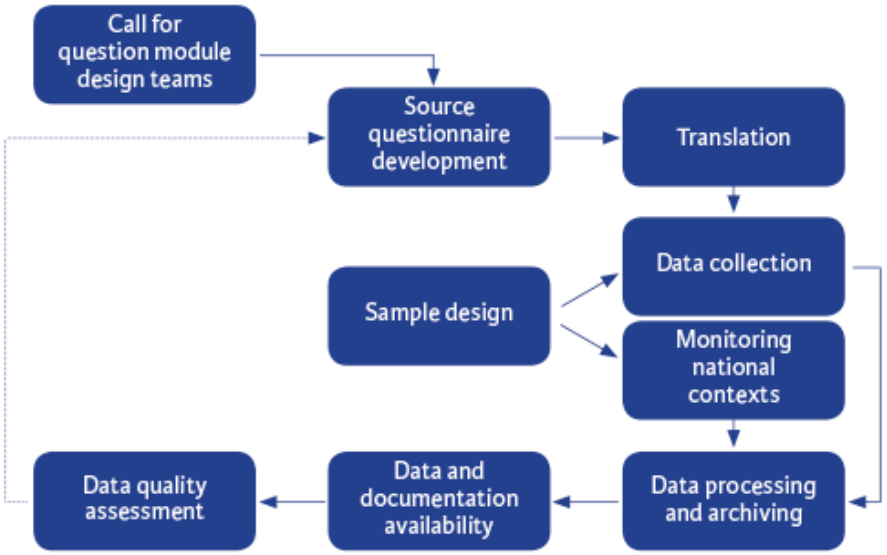

Figure 1. ESS methodology flowchart (Source: ESS website)

Each of the sections then has more detailed documentation for the national and central team, also available from the ESS website. For example, the Data collection section for Wave 9 of the ESS includes the following:

- ESS9 Interviewer briefings: NC manual
- ESS9 Interviewer briefings: Interviewer manual
- ESS9 Guidelines for enhancing response rates and minimising nonresponse bias
- ESS9 Guidelines on fieldwork progress reporting
- ESS9 FMS data upload portal - User manual
- ESS9 Fieldwork questionnaire (blank)

3.2 Extending the ESS to Australia – AUSSI-ESS

The Australian National University (ANU) conducted the ESS in an Australian context, under the title of the Australian Social Survey International – ESS (AUSSI-ESS). The purpose of the Australian study was to conduct a national study that enabled cross-continent comparisons

[8] https://www.europeansocialsurvey.org/methodology/

to European countries that participate, to better understand the similarities and differences between countries.

The project was led by Ben Edwards from the ANU, with data collection undertaken by the Social Research Centre using their probability-based panel, Life in Australia™. The fieldwork methodology for AUSSI-ESS was completed using self-completion of surveys via the web. This was different to that traditionally used in the ESS, which had used face-to-face survey methods almost exclusively through Wave 9. (Notably, however, the advent of COVID-19 did necessitate the switch in data collection methods in many ESS core countries in Wave 10, conducted in 2020.) Fieldwork for the AUSSI-ESS was undertaken from 17 February to 2 March 2020.

4. Data management for the ESS

Data management for the ESS is managed by Sikt, as a Work Package of the European Social Survey ERIC. Sikt and its predecessor NSD have been the data custodians for the ESS since its inception in 2001.

4.1 Technical infrastructure

As part of a technical infrastructure refresh over the past three years, the management of ESS has been progressively moving from largely manual processes conducted on internal systems to a cloud-based infrastructure, based on the Parquet data format developed by Apache[9] and using Microsoft Azure blob storage. A combination of databases are used in the system from both open (PostgreSQL) and proprietary (Datomic) providers.

Internal systems for the management and processing of data and metadata are workflow-oriented, using Python and Jupyter notebooks as the core software for the processing and management of ESS data, and infrastructure provided through Microsoft Azure. Metadata is managed through an internal instance of Colectica Repository, a commercial metadata repository platform, with access to the repository enabled through an API and GraphQL front-end query language[10]. External systems for dissemination of data are custom-built, based on JSON and Parquet data format. A representation of the data processing pipeline and relevant infrastructure components is included in ***Figure 2***.

[9] https://parquet.apache.org/
[10] https://api.nsd.no/graphiql

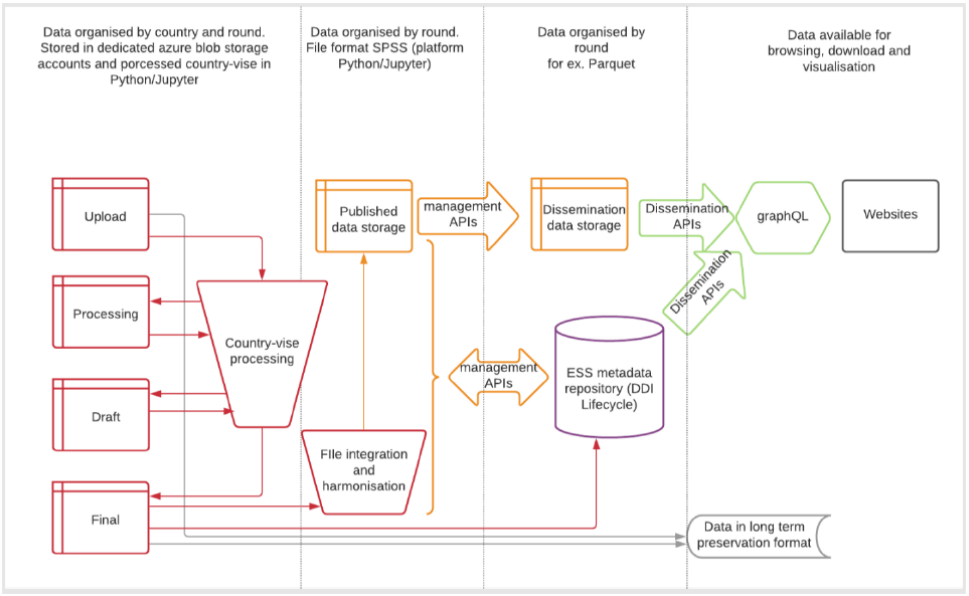

Figure 2. System design of Sikt data management system for ESS (Source: Agasøster et al., 2022, Fig. 1)

4.2 Data management and processing

The ESS Data management and processing platform is a collection of tools, files and programmes intended to help the staff of the ESS Team streamlining data processing while avoiding repeated manual work.

The main components of this system are a collection of Python-scripts and Jupyter Notebooks developed by Sikt's data scientists in cooperation with members of the ESS team. These files are maintained in a GitLab repository and regularly synchronised with a storage location in the Microsoft Azure Machine Learning Studio. In that way, it is guaranteed that the files are shared safely across the data team and get updated as soon as the programs change in the original location. The Python scripts in the Jupyter Notebooks access the data files that are stored in the Azure blob storages and its related metadata that is stored in Colectica repository via GraphQL API, ensuring that data processing is done in alignment with relevant metadata (Bidargaddi et al 2022).

The data processing workflow consists of multiple stages, from data ingest to data cleaning, data harmonisation and beyond. Each stage in the workflow is supported by multiple Jupyter Notebooks that the team members execute during the processing and edit if needed. Even if there is no absolute way to prescribe the exact steps in the data cleaning process because the data quality can vary from country to country, the standardised programs and templates ensure that the data cleaning is done the right way every time.

The aim of this workflow is to improve the overall data quality and to harmonise and standardise the data to as high a quality as possible (Kolsrud 2010).

Figure 3. ESS data management process

The main steps of data processing, outlined in Figure 3, are as follows:

1. Data ingest:

   For the ESS, each national team undertakes the initial data harmonisation for the data collected in their country. This is supported by template processing scripts, along with coding rules for processing the data, and additional guidelines as specified in the ESS Data protocol. Each country submits their data files in an agreed format, along with relevant context information and documentation (such as the country-specific questionnaire, sampling methodology, etc.) (as described by Kolsrud 2010).

   Once the data is submitted through the ESS Survey platform MyESS, the data files are automatically synchronised with the respective Microsoft Azure blob storage and thus available for the team members through the Python scripts that are executed in the Jupyter Notebooks. The files are finally imported into Pandas dataframes and can be processed together with related metadata from the Colectica Repository that serves as the basis for further processing.

2. Processing and cleaning:

   In this step, the data team tries to detect and fix incorrect, incomplete or illogical data within a dataset. There is a wide range of checks developed for the different types of data that are included in eight different Notebooks.

   The most important checks are listed below (as described by Kolsrud 2010):

   - Identification of duplicate or missing identification numbers and consistency of identification numbers across files
   - Content checks – aligning with the specifications in metadata, Data protocol and source questionnaires
   - Identification of duplicate data across rows
   - Wild codes – checking for invalid, out of range, extreme and missing values
   - Logical and consistency checks – checking for consistency of responses between related questions (e.g. between ages of parents and children)
   - Routing checks – checking for consistent application of routing through conditional sections of the survey

- Consistency over time for selected background variables – testing the distributions of background variables (such as religion and education distributions) to compare with previous waves.

3. Processing report:

   Once the checks in the Notebooks are executed, an HTML report for the National Team is automatically generated by a Notebook in addition to an Excel file with actions and notes that are required to address identified inconsistencies.

   The HTML report and the Excel document are then uploaded to MyESS and shared with the National team. After reviewing the report, the National team should correct the data and submit a new version of the data if necessary. This review process continues iteratively between the data team at Sikt and each National Team until the inconsistencies have been resolved or documented sufficiently.

4. Harmonisation and validation:

   In this step the data team inspects the final frequencies for each variable in the file. Additionally, the team compares the distribution of repeat variables, usually with the data from the previous round. Variables that have changed considerably over time will be flagged in the auto-generated HTML output and sent to the National team for verification.

   Further steps of this stage are:
   - Variable metadata preparation in Colectica Designer - Applying the DDI-Variable Cascade structure to country-specific variables that is needed for cross round-comparison as described in chapter 4.4.
   - Data editing – Assignment of missing values
   - Approval of the final draft file – the final draft file and the output from the variable comparison is shared with the National team on MyESS for a final check and approval

5. Anonymisation and integration:

   In the last step of processing, the national datafiles are integrated into one cross-national file. The integration program in the Notebook combines the data from the participating countries and adds final metadata from the Colectica repository such as variable labels and formats to the file. Before storing the final integrated dataset as an SPSS file, an anonymisation script is executed to ensure anonymity in the ESS data that will be published.

   Other steps of this processing stage:
   - ID scrambling – Identification numbers for respondent and interviewer will be replaced with random numbers to make the original value unrecognisable
   - Computation of Post-coded variables - Variables such as final interview, age of respondent and highest level of education are computed after the integration of the national files

4.3 Data access and dissemination

After data management and processing described in the previous section, the final version of the data is then made available through the ESS data platform. This platform is a custom-built site that generates real-time processing of ESS data from the ESS backend systems to produce custom data sets for users. The platform uses the Parquet data format for storage, and JSON for representation of the data on presented web pages for users.

Access to the data is available in three ways:

1. Users can download the data in common statistical package formats including Stata, SAS and SPSS.
2. Through an Application Programming Interface (API), leveraging the GraphQL interface to Sikt systems[11].
3. Accessing the data is through the data wizard. This tool enables users to customise their own dataset. In the first step, you can choose which ESS rounds and which countries you want to download data from. Step two makes it possible to choose which variables you want to include. It is possible to download entire variable groups, but also specific variables. In addition, there is a code comparison feature that makes it possible to compare codes over time. This feature is based on the DDI-Variable Cascade structure. Step three enables the download of this data in different formats (SAV, DTA, CSV).

4.4 Metadata management

Metadata for the ESS is managed using the Colectica Repository, providing a single point of truth for all metadata associated with the ESS. This work, undertaken through three rounds of EU funding (DASISH, SERISS and SSHOC), has enabled the creation of the ESS Question and Variable Database (QVDB)[12]. Using the DDI Lifecycle variable cascade[13], ESS variables are now documented at three levels: Conceptual Variable, Represented Variable and Instance Variable. This use of the variable cascade enables coordination of the longitudinal characteristics of the ESS data, providing comparability across time.

The use of the variable cascade and the Colectica repository has enabled additional options for Sikt for metadata reuse, particularly in combination with the use of Python and Jupyter notebooks. Data archivists at Sikt interact directly with the ESS metadata through the GraphQL API, enabling reuse of the metadata for both processing of data and access for dissemination purposes. This external API access is also an option for streamlining reuse across countries, an issue which will be explored further in the recommendations below.

[11] https://api.nsd.no/graphiql

[12] See Agasøster, B. et al. (2022) for details of the integration of Sikt data and metadata systems in the management of ESS

[13] https://doi.org/10.5281/zenodo.5180568

4.5 Standards and identifiers

A core principle of the ESS data preservation and management since inception has been the use of standards for the management of all data and metadata. All metadata for the ESS was managed initially following the DDI Codebook (Version 2.1) standard, and subsequently migrated to the DDI Lifecycle (Version 3.3) standard following the migration of Sikt systems to the Colectica and Azure systems in recent years.

Sikt has also worked to provide persistent identifiers for ESS and other data in its archive. This includes the use of DOIs for all ESS waves at the study level, with DOIs soon to be made available at the data file level (Agasøster et al., 2022).

The use of the DDI Lifecycle standard and its implementation in the Colectica Repository also have enabled Sikt to establish an internal identifier system for each metadata item in the repository - this includes all variables, questions, codes, code lists and categories - making each of these items potentially reusable artefacts for future reuse. These are not persistently identified at this point through an external registry, but this may be possible in future through proposed updates to the DDI Alliance's agency registry[14]. The registry is accessible through the Sikt API, making each of the metadata elements in the system reusable for external parties such as ESS national teams in each country.

5. Data management for AUSSI-ESS

Data management for the AUSSI-ESS is managed by the Australian Data Archive. The AUSSI-ESS has only been run once in Australia, with plans to repeat the survey in 2023, replicating much of the ESS Wave 11 survey.

5.1 Technical infrastructure

ADA uses the OAIS Reference model[15] as the basis for the design of technical and administrative infrastructure and services. ADA infrastructure is hosted and managed on secure virtual machines and storage operated by the National Computational Infrastructure (NCI), one of Australia's two Tier-1 high performance computing services supported under the National Collaborative Research Infrastructure Strategy[16] (NCRIS), the Federal strategy developed by the Department of Education, the Australian national education ministry.

ADA has been undertaking a major technical infrastructure refresh, similar to Sikt, following a decision to transition away from the Nesstar publishing platform in 2017. The first phase of this refresh focussed on the core public-facing platform for the external publishing of data from 2017-2020, and current updates are focussed on internal business systems and tools.

Internal data processing for ESS is conducted through file sharing on the NCI server, and accessed via Windows Remote Desktop clients using a file mount to the data store. Data

[14] https://registry.ddialliance.org/
[15] https://public.ccsds.org/Pubs/650x0m2.pdf
[16] https://www.education.gov.au/ncris

processing is completed using both SPSS and R and RMarkdown (through the RStudio desktop client), with processing content (syntax and documentation) stored on the NCI file mount. ADA is also currently implementing the Colectica repository and GraphQL API and query language for supporting AUSSI-ESS and other major ADA series.

For external data publishing, ADA has now implemented the the Dataverse repository platform[17], a purpose-built data publishing software developed by Harvard University and a community of over 50 repository service providers. Dataverse is Java-based software with a postgres database backend running on secure virtual machines hosted through NCI. This platform supports the DDI Codebook standard, along with API access to core data and metadata, and both system-internal and OIDC-based access and authentication services through AAF (the Australian national authentication provider and EduGAIN partner) and ORCID[18].

5.2 Data management and processing

Data deposits are completed by the data owner and deposited with the archive through ADA's online self-deposit system. ADA data archivists process the AUSSI-ESS data using a set of standardised data processing rules established and documented in an ADA internal wiki. Each of the steps in these procedures is supported by a template processing script in either SPSS or R/RMarkdown—for AUSSI-ESS this was completed using R. Scripts, data and processed materials are then managed and stored in an archival file store in the ADA storage on NCI.

The ADA data processing procedure, detailed in ***Table 1***, involves three phases:

1. Data preparation: review of files and related materials prior to ingest, and initial data privacy and data quality (e.g. spelling and appropriate labelling) checks.
2. Data cleaning: checking of variable and metadata characteristics, and editing of original data toward producing versions for publication.
3. Problem resolution: documentation of any data and metadata issues, and communications with the data owner to address issues.

Table 1 ADA data processing steps for AUSSI-ESS

Step No.	Activity
1	***Data Preparation***
1.1	Data File Format
1.2	Converting string to numeric variables
1.3	Privacy Act checks
2	***Quantitative Data Cleaning***
2.1	Variable name checking
2.2	Variable label checking
2.3	Value label checking
2.4	Value range checking

[17] https://dataverse.org/
[18] https://orcid.org/

2.5	Logic checking
2.6	Checking for direct identifiers
2.7	Checking for indirect identifiers
2.8	Checking for uncommon variable types
2.9	Recoding
2.10	Creating map variable
2.11	Anonymisation of data collected from online panel data
2.12	Generating data dictionary
3	***Problem Resolution***
3.1	Before Contact
3.2	Contact

As AUSSI-ESS also required harmonisation of the data files, additional harmonisation processing was undertaken with the data in Step 2.9 to harmonise the content with the ESS variable specifications. This processing, completed in R, included harmonisation of variable names to align with ESS mnemonics (also used in the QVDB). Sample R code for the variable naming is included in ***Table 2***.

Table 2 Sample variable name editing in R

```
names(aussi)[names(aussi) == 'A1'] <- 'nwspol'
names(aussi)[names(aussi) == 'A2'] <- 'netusoft'
names(aussi)[names(aussi) == 'A3'] <- 'netustm'
names(aussi)[names(aussi) == 'A4'] <- 'ppltrst'
names(aussi)[names(aussi) == 'A5'] <- 'pplfair'
names(aussi)[names(aussi) == 'A6'] <- 'pplhlp'
```

Other characteristics of the variables, such as codes and category labels, were already harmonised as the ESS source questionnaire had been used for the coding of the questionnaire by the Social Research Centre (the data collection agency). This use of standardised metadata in the data collection process is an additional means for ensuring harmonisation of content, referred to as 'input harmonisation' in cross-national survey design guidelines. Such input means that collected data and metadata can be tested and checked against expected inputs, a process which can be automated throughout the data lifecycle.

5.3 Data access and dissemination

Dissemination of the AUSSI-ESS data is published through ADA's Dataverse portal. Consistent with Dataverse's design principles, the AUSSI-ESS is assigned a DOI and a landing page. A representation of the AUSSI-ESS landing page is included in ***Figure 5***.

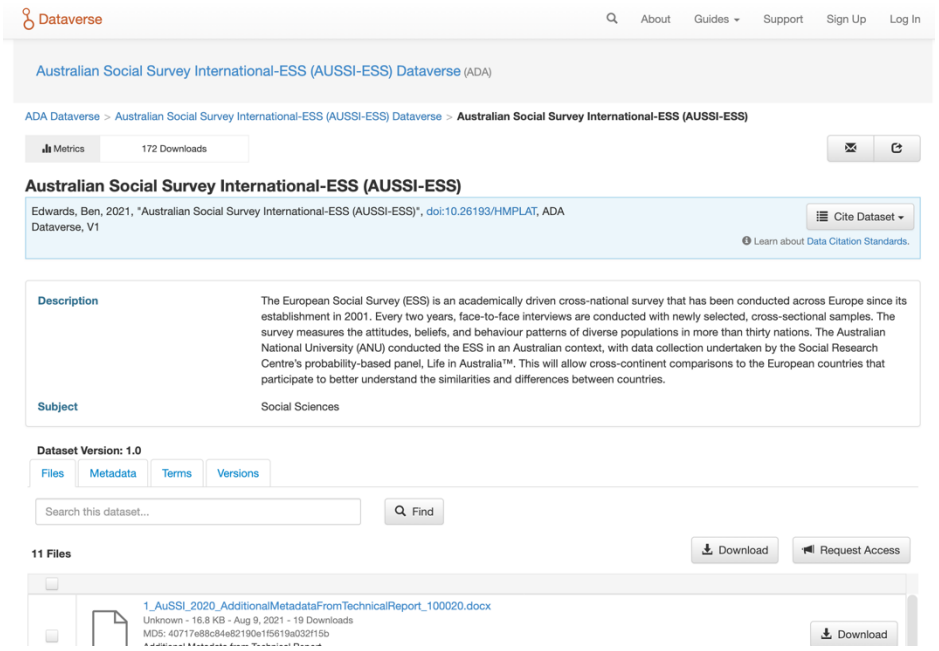

Figure 5. The AUSSI-ESS landing page (Source: http://dx.doi.org/10.26193/HMPLAT)

Metadata is documented and accessible in DDI, Dublin Core and JSON formats through both 'point and click' and API access mechanisms. Data for AUSSI-ESS is released in ADA's standard formats—SPSS, Stata, SAS and CSV—along with documentation such as the fielded questionnaire, data dictionary and other materials. Access to the data is restricted but is available on request using the 'data request' button on the AUSSI-ESS landing page.

5.4 Metadata management

Metadata for the AUSSI-ESS occurs at three broad levels - study, file and variable levels - consistent with the DDI Codebook standard used at ADA. Variable metadata for AUSSI-ESS is managed using the ADA Dataverse system, generated from the SPSS statistical file format produced in ADA data processing. The authoritative record for ADA variable metadata is therefore the processed SPSS file and data processing scripts and thus not externally accessible and reusable. Processing scripts are preserved in the ADA archive. Study level metadata is then manually added to the Dataverse catalogue using a manual entry form in the Dataverse repository[19]. Basic file metadata is generated automatically with the ingest of files into the repository and then supplemented with additional, manually created metadata.

5.5 Standards and identifiers

As noted earlier, ADA makes extensive use of the DDI Codebook standard in the management of AUSSI-ESS and other datasets in its collection. This has migrated from DDI Codebook version 1.1.2 in the previously-used Nesstar platform, to version 2.1 as implemented in the Dataverse repository platform. In addition, ADA publishes (study-level) metadata to the Datacite registry and mints DOIs for all datasets, and will be making DOIs available at the data file level in 2023.

[19] https://dataverse.ada.edu.au/

ADA also incorporates the use of DDI controlled vocabularies into the AUSSI-ESS and other metadata records. These vocabularies, however, require manual management through archival practices, rather than being controlled directly through technical controls, due to the design of the Dataverse repository interface for this metadata, which allows only text entry in the implementation in use at ADA[20].

6. Comparative analysis of data management practices

The previous sections of this report provide case studies of the core management practices of the central data archives for the ESS and the satellite AUSSI-ESS. In this section, we consider now a comparison of the two sets of practices and their potential for closer coordination with a particular emphasis on the potential for machine-to-machine (M2M) integration in the reuse of data and metadata.

6.1 FAIR Implementation Profile - ESS

For the WorldFAIR project, a FAIR implementation profile for the ESS has been completed using the FIP Wizard tool Excel template, and is currently being added to the online FIP Wizard[21]. A summary of the profile is included in Appendix One (Section 9.1) of this report.

Sikt recently undertook a FAIR assessment of its current technical infrastructure as part of the SSHOC project funded under the EU Horizon 2020 program[22]. The details of this assessment can be found in Agasøster et al. (2022).

6.2 FAIR Implementation Profile - AUSSI-ESS

A FAIR implementation profile for the AUSSI-ESS has been completed using the FIP Wizard tool Excel template and the online FIP Wizard. A summary of the profile is included in Appendix Two (Section 9.2) of this report.

An overview of the implementation of the FAIR principles in the Dataverse platform was undertaken in 2019 by Mercè Crosas[23], a co-author on the original FAIR principles and the project lead for Dataverse at that time. Details of that assessment can be found in Crosas (2019). In her assessment, Crosas noted the means through which the Dataverse platform leverages both dataset and variable information in the Dataverse system to enable findability of content (through dataset and variable searches) and interoperability and reusability (for conducting online analyses of data through Dataverse extensions such as the Dataverse Explorer tool developed by Scholars Portal in Canada).

---

[20] Recent releases of the Dataverse platform have included support for controlled vocabularies using Javascript to interact with external vocabulary services such as the CESSDA vocabulary service. ADA is currently looking to implement this in the next platform upgrade in 2023.

[21] The Wizard is available at https://fip-wizard.ds-wizard.org/.

[22] Social Sciences & Humanities Open Cloud (SSHOC), Project Number: 823782, INFRAEOSC-04-2018.

[23] See presentation:
https://scholar.harvard.edu/mercecrosas/presentations/fair-guiding-principles-implementation-dataverse

Variable metadata in Dataverse is, however, not fully interoperable or reusable. ADA is therefore in the process of implementing the Colectica repository to provide a registry service for use in archivalist data processing. This provides both internal reuse and interoperability for data archivists, and also potential reuse for findability and reusability in dissemination.

6.3 Comparable practices and content

In comparing the processing methods and current standards in use for ESS and AUSSI-ESS, it can be seen that there are a number of similarities in the approaches used, albeit with differing levels of automation and different versions of standards and software in use. ***Table 3*** outlines key points of similarity between the two services.

Table 3 Comparison of ESS and AUSSI-ESS infrastructure for FAIR data integration

Item	ESS	AUSSI-ESS
Source content	ESS survey specifications	ESS survey specifications
Technical infrastructure	Microsoft Azure	
Machine interoperability	API (GraphQL)	API (Dataverse, GraphQL*)
Access and authentication	Metadata: not required Data: EduGAIN/OIDC APIs: API keys	Metadata: not required Data: Internal, EduGAIN/OIDC APIs: API keys
Processing software	Python, Jupyter notebooks	SPSS, R, RMarkdown
Data standards	DDI Lifecycle version 3.1	DDI Codebook version 2.1
Data repository	Custom built	Dataverse
Metadata repository	Colectica Repository	Dataverse, Colectica Repository*
Persistent identifiers	ESS survey specifications	ESS survey specifications
Identified content	Studies, waves, data files*	Studies, waves, data files*
Controlled vocabularies	DDI and CESSDA vocabularies	DDI vocabularies**

** denotes currently in testing and/or implementation*
*** denotes human controlled (through text input) rather than machine controlled (through software)*

As this comparison demonstrates, there is a large degree of similarity in the current approaches to data management between the two services from Sikt and ADA. While there are both similarities and differences in the practices of the two services, both groups use common standards (DDI), technologies (Colectica, GraphQL) and resources (shared source content for the surveys, common metadata models and semi-automated processing scripts). The core processing workflows of the two archives, as documented in Sections 4.2 and 5.2, are also closely in parallel, with consistent use of core processing checks, editing processes and data release protocols.

6.4 Differences in implementation

The differences between the two groups, in terms of both practices and technical implementations, can create limits on interoperability. This impacts particularly on the

potential for machine-to-machine interoperability. In the two case studies, three notable examples of such differences were identified:

1. ***Different processing software:*** Sikt has adopted Python and Jupyter notebooks as its core archive processing tools, while ADA has adopted SPSS and progressively R and RMarkdown.
2. ***Different versions of standards:*** the differences in which version of the DDI standard is used - DDI Lifecycle for ESS and DDI Codebook for AUSSI-ESS - limits the current possible reuse of the ESS Question and Variable Database (QVDB), as DDI Codebook Version 2.1 does not have the capacity for identifiers or reusability of fine-grained metadata such as variables.
3. ***Use of variable metadata registries:*** Sikt has adopted the Colectica registry and GraphQL for the management of metadata, providing full access to the DDI variable cascade. By comparison, ADA has only recently commenced this activity, and makes little reuse of metadata in its day to day operations.

The implications of these three differences are important for considering the extent to which interoperability can occur at a machine-to-machine level. The impact of differences in processing software are small - users (human and machine) can work between technologies effectively by the use of interchange data formats and shared libraries (such as the Python "rpython" library[24] and R "reticulate" package[25]). The inconsistencies in standards and registries are more problematic. Moving between standards or even versions of standards requires cross-mapping of the content of the standards, which may result in conversion challenges or even incompatibilities. The lack of a registry for ADA means that access to the variable content is limited or absent. While humans can work around these issues through use of additional documentation, machines cannot - they depend on these services to execute code and provide interoperable services.

6.5 Potential for coordination and reuse

Given that many of the processing steps in the two archives can likely be mapped together, there is significant potential for reusability of resources between the two country datasets, and for reuse between archives more generally. For example: for harmonised content, the case study analysis of the processing of AUSSI-ESS illustrated that the use of harmonised inputs - in the form of a common source questionnaire used by data collection agencies in each country - reduced the need for manual data harmonisation in the management and archiving of data (Section 5.2). These consistencies in processes, standards and content suggest significant potential for enabling both human and, progressively, machine interoperability for archives processing common content, particularly where metadata support can be introduced early in the data lifecycle.

Some potential areas for exploring such integration are suggested in the case studies above. As part of the ESS FAIR assessment process (Agasøster et al., 2022), the ESS archiving team outlined a set of recommendations for implementation of a FAIR-compliant integrated repository. Of particular note were the following:

[24] https://pypi.org/project/rpython/
[25] https://rstudio.github.io/reticulate/

> *(1) Controlled vocabularies (CVs)/standards should be used where possible. At least, use of domain-specific CVs should be applied but, if possible, use cross-domain CVs/standards, as CVs help increase machine actionability.*
>
> *(2) Standardised licences for data and metadata should be set up. The use of standardised licences such as Creative Commons gives data access and is often used as a provenance indicator in the FAIR measurement matrixes thus reassuring researchers in their use and reuse of data.*
>
> *(3) Persistent identifiers should be allocated to their own resources to ensure persistence and trackability of changes in data and metadata.*
>
> *(5) Thorough assessment of the data and metadata organisation in repository (-ies) should be made. This should include the implementation of a stable domain-specific metadata standard. ESS ERIC chose the DDI LifeCycle standard, which ensures interoperability and reusability of metadata and data.*
>
> *(6) "Off-shelf" opportunities should be explored i.e. what can be acquired elsewhere - for example Colectica. For NSD, use of Colectica ensured a time- and resource-efficient development process.*
>
> *(7) The use of APIs should be implemented as it is the contemporary way of accessing various data.*
>
> *(9) Authentication, access control and user statistics are key elements in data dissemination. Use of single sign-on gives a better user experience and reduces double registration, thus facilitating better user management.*
>
> (Agasøster et al. 2022, p. 27)

These recommendations outline some core technical and process requirements which will be beneficial for managing interoperable and FAIR-compliant services. It is for this reason that, as part of the case study investigation, ADA and Sikt have explored some initial candidate resources that may be able to be shared and reused in Phase 2 of this work package. Initial testing of access to the Question and Variable Database (QVDB) through the GraphQL API indicates that ADA staff will be able to access the API directly, using Australian national EduGAIN provider credentials. Sikt staff have been using these APIs in conjunction with their Python/Jupyter environments to interact with ESS metadata in processing.

This initial testing, along with the case study comparisons of the core features of both the ADA and Sikt practices, suggest that a number of these recommendations are already likely to be achievable within the lifetime of the WorldFAIR project; in particular, greater use of controlled vocabularies (Rec. 1), common standards (Rec. 5), API access (Rec. 7) and standardised authentication and access control (Rec. 9). The two partners have therefore proposed the sharing of API access and processing libraries developed by each group along with key outputs such as Jupyter notebooks and RMarkdown reports, as a first stage of activity for Phase 2 of WP6.

Other items on the FAIR recommendations list may require more detailed analysis and resource support in order to be achieved. Notably, the use of persistent identifiers for registry content (Rec. 3) requires some implementation changes with the Sikt (and ADA) Colectica registries, while the reuse of such content depends also depends on licences associated with the metadata that are sufficiently permissive to allow extensive reusability (Rec. 2). However, these actions can also be explored in the second phase of the Work Package.

7. Recommendations and next steps

The previous section articulates the key points of similarity and difference in the ESS and AUSSI-ESS, and key challenges for M2M interoperability. Identifying, surfacing and comparing these differences provides the opportunity to revise and update practices to improve such integration. This is the focus of the recommendations for this report, as Work Package 6 moves into the next stage of activity focussing on development of shared tools and resources. These recommendations fall into three categories - 1. aligning standards; 2. establishing common tools; and 3. use of registry services.

7.1 Aligning standards

Recommendation 1: The ADA should move to the use of the DDI Lifecycle Version 3.3

Recommendation 2: The ADA should adopt the use of the DDI Variable Cascade for the management of their time series and longitudinal content

In order to harmonise and interoperate using common content, the two archives need to be able to ensure that their content can be effectively exchanged by machine-based services in both facilities. While technical workarounds are possible in some circumstances, this process is more readily enabled by the use of a mutually agreed standard. As both organisations are already users of DDI standards, and the use of DDI is prevalent in the social science community, alignment on DDI is appropriate. In addition, the use of DDI Lifecycle Version 3.3 will enable the use of identifiable, maintainable and reusable content.

7.2 Establishing common tools

Recommendation 3: The ADA and Sikt should undertake a pilot to test the use of common libraries and scripts in the processing of ESS and AUSSI-ESS content

Recommendation 4: A public repository of template scripts should be made available for reuse in processing other cross-national datasets

The case studies identified consistency in the broad processes and approaches both organisations use for the processing of ESS and other data. There is also progress in both groups in the use of scripting tools for increasing automation of processing and data harmonisation. The establishment of a shared script repository - first internally, and then externally where suitable, given the confidentiality of data sources - provides a means of

further harmonising practices and increasing reuse of both processes and metadata content, particularly where used in conjunction with variable and other registry services (see below).

7.3 Establishing and using registries

Recommendation 5: ADA and Sikt should establish formal registries of variables and other reusable metadata and content, and expose current internal content from these registries for reuse through API services.

Recommendation 6: Where possible, common content such as harmonised variables and code mappings should be persistently identified and made available through such registries to enable standardised and reusable harmonisation practices

Analysis of ADA and Sikt systems and infrastructures identified that both organisations are making progress towards the establishment of internal registry services for their metadata and data content. The use of the Colectica environment at Sikt has made it possible to expose that content through a GraphQL API interface, while ADA is also moving towards a shared registry for its content. These registries however, while of benefit, could be further optimised through the use of persistent identifiers attached to the registry content. This would achieve two outcomes:

1. Ensuring a consistent and persistent means for machine access to content, using structured, standardised web services models;
2. Enabling the standardised process (described in 7.2 above) to leverage these registry services to reuse the content of the registries in current and future processes. By aligning the templated content of the script library with the persistent content of the variable registries, ADA and Sikt - and progressively other data management organisations - should be able to develop a common core of complex, machine-led data management services that rely on well-established and persistent metadata.

8. References

Agasøster, B., Havåg Bergseth, G., Beuster, B., Bidargaddi, A., Risnes, Ø., Kalgraff Skjåk, K., Stavestrand, E.. (2022). *D5.13 Recommendations for a FAIR compliant integrated data and metadata repository (ESS as a service) (v1.0).* Zenodo. https://doi.org/10.5281/zenodo.6564151

Bidargaddi, A., Agasøster, B., Kalgraff Skjåk, K., Risnes, Ø. (2022). D5.14 Report on Preparing the ESS for Services in the EOSC (ESS as a Service) (v1.0). Zenodo. https://doi.org/10.5281/zenodo.6779526

DDI Training Group. (2021, August 11). Variables and the Variable Cascade. Zenodo. https://doi.org/10.5281/zenodo.5180568

Dubrow, J.K., Tomescu-Dubrow, I. (2015) The rise of cross-national survey data harmonization in the social sciences: emergence of an interdisciplinary methodological field. Quality and Quantity, 50, 1449–1467. https://doi.org/10.1007/s11135-015-0215-z

ESS-ERIC (n.d.). Methodology Overview. European Social Survey. https://www.europeansocialsurvey.org/

Granda, P., Blasczyk, E. (2010) Data harmonization. In: *Cross-cultural survey guidelines*. http://ccsg.isr.umich.edu/

Kolsrud, K., Midtsæter, H., Orten, H., Kalgraff Skjåk, K., Øvrebø, O. (2010). Processing, Archiving and Dissemination of ESS data. The Work of the Norwegian Social Science Data Services. *Ask: Research and Methods*, 19(1), 51-92.

Smith, T., Fu, Y. (2016) The Globalization of Surveys. In Wolf, C., Joye, D., Smith, T., Fu, Y. *The SAGE Handbook of Survey Methodology*, Chapter 41, DOI: https://dx.doi.org/10.4135/9781473957893

Tofangsazi, B., Lavryk, D. (2018). We coded the documentation of 1748 surveys across 10 international survey projects: This is what data users and providers should know. *Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences*, *4*(2), 27-31.

Vardigan, M., Granda, P., Hoelter, L. (2017) Documenting Survey Data Across the Life Cycle. In Wolf, C. et al (2017) The SAGE Handbook of Survey Methodology. http://dx.doi.org/10.4135/9781473957893

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.

9. Appendices

NB: In both the following tables, colour is used to visually group the sub-principles of each FAIR principle; i.e. all 'F' sub-principles are pink, 'A' sub-principles are yellow, 'I' are green, and 'R', blue.

9.1 Appendix One: FAIR Implementation Profile for ESS (managed by Sikt)

This table summarises the FAIR Implementation Profile for the Sikt services supporting the European Social Survey through the ESS Data Portal and related services.

Table 5 FAIR Implementation Profile (FER) for European Social Survey - Sikt

FAIR Principle name	Referring to MetaData or Data	FIP question	FER Enabling Resource used in ESS (Sikt)
F1	MD	What globally unique, persistent, resolvable identifier service do you use for metadata records?	DDi URN - used internally only
F1	D	What globally unique, persistent, resolvable identifier service do you use for datasets?	DOI
	D	What globally unique, persistent, resolvable identifier service do you use for datasets?	DDI URN - Not externally resolvable. Structure exists in Colectica, but the final persistent link is missing.
F2	MD	What metadata schemas do you use for findability?	DDI-Lifecycle 3.3 - Study Unit
F3	D	What is the schema that links the persistent identifiers of your data to the metadata description?	DDI-Lifecycle 3.3
F4	MD	Which service do you use to publish your metadata records?	GraphQL API (is there a specific and exposed NSD/SIKT API endpoint)
F4	MD	Which service do you use to publish your metadata records?	Colectica web services.
F4	D	Which service do you use to publish your datasets?	ESS Website landing page, API
F4	D	Which service do you use to publish your datasets?	EOSC Portal
A1.1	MD	Which standardized communication protocol do you use for metadata records?	HTTPS
A1.1	D	Which standardized communication protocol do you use for datasets?	HTTPS
A1.2	MD	Which authentication & authorisation service do you use for metadata records?	No authorisation. eduGAIN/OIDC in GraphQL API
A1.2	D	Which authentication & authorisation service do you use for datasets?	eduGAIN/OIDC, transport: GraphQL + data file formats.

A1.2	D	Which authentication & authorisation service do you use for datasets?	Azure Active Directory, transport: Azure APIs + data file formats.
A2	MD	What metadata preservation policy do you use?	ESS Policy?
I1	MD	What knowledge representation language (allowing machine interoperation) do you use for metadata records?	JSON in GraphQL
I1	D	What knowledge representation language (allowing machine interoperation) do you use for datasets?	Parquet
I2	MD	What structured vocabulary do you use to annotate your metadata records?	DDI-Lifecycle 3.3 structured codelists
I2	D	What structured vocabulary do you use to encode your datasets?	ISO3166-1 for country and ISO639-2 for language, NACE Rev 2 for Industry, ISCO08 for occupation, NUTS for regions.
I2	D	What structured vocabulary do you use to encode your datasets?	DDI Controlled vocabularies, CESSDA vocabularies, ELLST
I3	MD	What semantic model do you use for your metadata records?	DDI-Lifecycle
			DDI-CDI
I3	D	What semantic model do you use for your datasets?	DDI-Lifecycle
R1.1	MD	Which usage license do you use for your metadata records?	CC BY-SA 4.0
R1.1	D	Which usage license do you use for your datasets?	CC BY-NC-SA 4.0
R1.2	MD	What metadata schema do you use for describing the provenance of your metadata records?	DDI-Lifecycle 3.3
			DDI-CDI
R1.2	D	What metadata schema do you use for describing the provenance of your datasets?	DDI-Lifecycle, DDI-Codebook, DDI-CDI (future)

9.2 Appendix Two: FAIR Implementation Profile for AUSSI-ESS (managed by ADA)

This table summarises the FAIR Implementation Profile for the ADA services supporting the AUSSI-ESS through the ADA Dataverse and related services.

Table 6 FAIR Implementation Profile (FER) for AUSSI-ESS – Australian Data Archive

FAIR Principle name	Referring to MetaData/Data	FIP question	FER Enabling Resource used in WP06 Social Surveys
F1	MD	What globally unique, persistent, resolvable identifier service do you use for metadata records?	DataCite DOI resolution service
F1	D	What globally unique, persistent, resolvable identifier service do you use for datasets?	DataCite DOI resolution service
F2	MD	What metadata schemas do you use for findability?	DDI Codebook Version 2.1
F2	MD	What metadata schemas do you use for findability?	DataCite metadata schema version 3.1
F3	D	What is the schema that links the persistent identifiers of your data to the metadata description?	No implementation choice has been made by this community
F4	MD	Which service do you use to publish your metadata records?	ADA Dataverse
F4	D	Which service do you use to publish your datasets?	ADA Dataverse
A1.1	MD	Which standardized communication protocol do you use for metadata records?	HTTPS|Hypertext Transfer Protocol Secure
A1.1	MD	Which standardized communication protocol do you use for metadata records?	REST|Representational state transfer
A1.1	D	Which standardized communication protocol do you use for datasets?	HTTPS|Hypertext Transfer Protocol Secure
A1.1	D	Which standardized communication protocol do you use for datasets?	REST|Representational state transfer
A1.2	MD	Which authentication & authorisation service do you use for metadata records?	None for open records; SAML2|Security Assertion Markup Language 2.0
A1.2	D	Which authentication & authorisation service do you use for datasets?	SAML2|Security Assertion Markup Language 2.0
A2	MD	What metadata preservation policy do you use?	RDA Core Trust Seal Certification
I1	MD	What knowledge representation language (allowing machine interoperation) do you use for metadata records?	JSON|JavaScript Object Notation
I1	MD	What knowledge representation language (allowing machine	XMLS|eXtensible Markup Language Schema

		interoperation) do you use for metadata records?	
I1	D	What knowledge representation language (allowing machine interoperation) do you use for datasets?	SPSS, Stata, SAS, R, CSV
I2	MD	What structured vocabulary do you use to annotate your metadata records?	DDI Vocabularies, CESSDA Vocabularies (note that these are not currently controlled - text fields in Dataverse)
I2	D	What structured vocabulary do you use to encode your datasets?	None
I3	MD	What semantic model do you use for your metadata records?	DDI Codebook Version 2.1
I3	D	What semantic model do you use for your datasets?	DDI Codebook Version 2.1
I3	D	What semantic model do you use for your datasets?	SPSS, Stata, SAS, R, CSV
R1.1	MD	Which usage license do you use for your metadata records?	CC-0
R1.1	D	Which usage license do you use for your datasets?	Custom licenses
R1.2	MD	What metadata schema do you use for describing the provenance of your metadata records?	DDI Codebook Version 2.1
R1.2	D	What metadata schema do you use for describing the provenance of your datasets?	DDI Codebook Version 2.1