

Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782

Start Date of Project: 01/01/2019

Duration: 40 months

Deliverable 5.14 Report on Preparing the ESS for Services in the EOSC (ESS as a Service)

Dissemination Level	PU
Due Date of Deliverable	30/04/2022, M40
Actual Submission Date	29/04/2022
Work Package	WP5 - Innovations in Data Access
Task	Task 5.5 ESS as a service: a pilot making cross-national survey data FAIR
Type	Report
Approval Status	Approved by EC - 04 May 2022
Version	V1.0
Number of Pages	p.1 – p.27

Abstract:

This report documents the process of preparing the ESS for services in the EOSC (ESS as a service) in Task 5.5 of the SSHOC project. It encompasses a set of workflows, architectures, protocols of the ESS services set up as part of the pilot and offers recommendations for other data archives based on the lessons learned.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability. This deliverable is licensed under a Creative Commons Attribution 4.0 International License.



History

Version	Date	Reason	Revised by
0.0	22/03/2022	Compilation of technical elements	Archana Bidargaddi (AB), Ørnulf Risnes (ØR) Bodil Agasøster (BA)
0.1	24/03/2022	Team review	Knut Kalgraff Skjåk
0.2	25-31/03/2022	Peer review	Matti Heinonen (FSD)
0.3	04-18/04/2022	Address peer review comments	AB, BA
0.4	20/04/2022	Final restructured version to CESSDA	BA, AB

Author List

Organisation	Name	Contact Information
Sikt ¹ (CESSDA/NSD)	Archana Bidargaddi	archana.bidargaddi@sikt.no
Sikt (CESSDA/NSD)	Bodil Agasøster	bodil.agasoster@sikt.no
Sikt (CESSDA/NSD)	Knut Kalgraff Skjåk	knut.skjak@sikt.no
Sikt (CESSDA/NSD)	Ørnulf Risnes	ornulf.risnes@sikt.no

¹ Sikt - Norwegian Agency for Shared Services in Education and Research is the result of a merger between NSD (Norwegian Centre for Research Data), Uninett AS and Unit into the Norwegian Directorate for ICT and Joint Services in Higher Education and Research. See: <https://sikt.no/about-sikt> (Accessed April 2022)

Executive Summary

Task 5.5. ESS as a service: a pilot making cross-national survey data FAIR aims to increase the FAIRness of ESS data, with particular focus on increasing the interoperability of data holdings, and to make ESS data and metadata available from the European Open Science Cloud.

As part of the task, a new repository platform, APIs, landing pages and solutions for authentication and authorization were developed and deployed.

Deliverable D5.14, Report on preparing the ESS for services in the EOSC (ESS as a service), documents the implementation of the ESS Pilot project and presents recommendations for similar data archives based on ESS as a service's workflow, architecture, and protocols.

The main general recommendations of the ESS data archive for similar development work are to:

- 1) use cross-functional teams and an agile approach,
- 2) build API-led systems,
- 3) use relevant protocols in a systematic way,
- 4) develop metadata-driven systems,
- 5) keep a close connection between data and metadata, and
- 6) establish solutions for data download and analysis.

Abbreviations and Acronyms

AIP	Archival Information Package
API	Application Programming Interface
Azure / Microsoft Azure	Cloud computing service operated by Microsoft for application management via Microsoft-managed data centres
CESSDA (ERIC)	Consortium of European Social Science Data Archives European Research Infrastructure Consortium
REST	Representational state transfer (a software architectural style)
REST API	API that follows REST architectural style and allows for interaction with RESTful web services
CSV	Comma-Separated Values file
DataCite	Leading global provider of DOIs for research data
DDI	Data Documentation Initiative
DDI LifeCycle	The Data Documentation Initiative LifeCycle (model)
DIP	Dissemination Information Package
DOI	Digital Object Identifier
eduGAIN	Interfederation service connecting identity federations around the world, simplifying access to content, services and resources for the global research and education community
ELSST	European Language Social Science Thesaurus
EOSC	European Open Science Cloud
ESS (ERIC)	The European Social Survey (European Research Infrastructure Consortium)
FAIR	Findable, Accessible, Interoperable, Reusable
Feide	The national solution for secure login and data sharing in the educational and research sector in Norway
FSD	The Finnish Data Archive
GA	Grant Agreement
GraphQL	Query language for APIs and a runtime for fulfilling queries with existing data

HTTP	HyperText Transfer Protocol
IaC	Infrastructure as Code
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
Nesstar	NSD-developed software system for data publishing and online analysis
NSD	Norwegian Centre for Research Data (since 2022 part of Sikt)
NumPy	Package for scientific computing with Python
OAIS	Open Archival Information System
OAuth	Open Authorization
OIDC	OpenID Connect
Pandas	Fast, powerful, flexible and easy to use open-source data analysis and manipulation tool built on top of Python
Parquet (Apache)	File format for data storage
PDF/A	Archival format of Portable Document Format
Python	Programming language
QVDB	The Question Variable Database
R	Programming language
SAS	Statistical Analysis Software
.SAV	File format created by SPSS
SciPy	Fundamental algorithms for scientific computing in Python
SERISS	The Synergies for Europe's Research Infrastructures in the Social Science project
Sikt	Norwegian Agency for Shared Services in Education and Research
SIP	Submission Information Package
SPSS	Statistical Package for the Social Sciences
SSHOC	The Social Sciences & Humanities Open Cloud project
Stata	Software for Statistical and Data Science

Statsmodels	A Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration
URL	Uniform Resource Locator
XML	Extensible Markup Language

Table of content

1.	Introduction.....	9
1.1	The Achievements of SSHOC Task 5.5	9
1.2	The structure of the Document	10
2.	The ES Workflow	11
3.	System Architecture.....	12
3.1	Ingest.....	13
3.2	Data Management.....	13
3.3	Preservation.....	14
3.4	Dissemination Backend:.....	14
3.5	ESS Data Portal:	14
4.	Protocols.....	18
4.1	DDI-Lifecycle	18
4.2	GraphQL API	18
4.3	OIDC/OAuth (for login).....	19
4.4	Terraform.....	19
4.5	Apache Parquet	19
4.6	JSON.....	20
4.7	HTTPS	20
5.	Recommendations.....	20
5.1	Agile Approach, Cross Functional Teams.....	20
5.2	API First	21
5.3	Enabling Metadata Driven Services.....	22
5.3.1	Controlled Vocabularies	22
5.3.2	Versioning of Data and Documents	23
5.3.3	Versioning of Metadata	23
5.4	Build Data Solutions for Download and Analysis.....	23
5.5	Connect Data and Metadata.....	24

6.	Conclusions	25
7.	References.....	26
	List of Figures	27
	List of Tables	27

1. Introduction

The SSHOC project aims to build a Social Sciences and Humanities Open Cloud (SSHOC) as part of the European Open Science Cloud (EOSC) with FAIR² data, tools, and training in a cloud-based infrastructure. Better access to data and innovative tools and services has been provided for users, and reuse of data has been facilitated through Open Science and FAIR principles and practices.

The European Social Survey (ESS) has from 2001 in collaboration with national teams collected data on attitudes, beliefs, values, and behaviour from interviews with respondents who are representative of each country's population in 40 countries. Since 2023, the ESS has offered open and free online access to the cross-sectional research data to all non-commercial users.

By 2018, the infrastructure for curation and dissemination of data and metadata needed major upgrading to satisfy the needs of modern users and to comply with ambitions and requirements of Open Science, the FAIR principles and the EOSC. Participation in the SSHOC project has permitted ESS to develop and realize the needed upgrading of the infrastructure of the survey. The successful implementation presents new ways of using modern technologies and industry protocols to increase FAIRness of data archives.

1.1 The Achievements of SSHOC Task 5.5

Based on the ESS, Task 5.5, ESS as a service: a pilot making cross-national survey data FAIR, pilots how cross-national, longitudinal survey data and metadata can be prepared and provide services for the EOSC. D5.13 Recommendations for a FAIR compliant integrated data and metadata repository aimed to explain choices taken as part of the development process of the new infrastructure with data and metadata repositories and APIs, and to explain lessons learned and opportunities for replication (Agasøster et al., 2022). In this report, the focus is on documenting the work undertaken.

Preparing ESS as a service for EOSC required modification and modernization of the complete workflow from data ingest to data dissemination. The solution has taken into consideration integration with (external) other ESS services such as the project management system with inbuilt data deposit service - MyESS³ during the design and development phase, providing a well synchronized holistic system and

² Refer to the European Commission, 2018.

³ My ESS is an open-source collaborative environment for documenting the life cycle of major international survey projects, which has been adopted by the European Social Survey, please refer to: *UPF becomes the new 'hub' of a management platform to improve the European Social Survey* | myScience / news / science wire available here:

better workflow. The aim has been to generate value through the various phases of the workflow and increase the overall efficiency and FAIRness of ESS services.

The report undertakes to document the various steps of the work and to offer recommendations to other data archives based on this.

1.2 The structure of the Document

Chapter 2 of the report documents the new workflow of the ESS survey, followed by Chapter 3, which outlines the system architecture. The chapter explains the various components of the service briefly.

In Chapter 4, the protocols used in the new system are presented and discussed. The section explains the considerations behind the selection of these protocols and their benefits.

Organizational, domain specific, technical and development project recommendations for other data archives endeavoring to do similar upgrades have been added in Chapter 5.

Section 6 concludes the report with an overview of achievements made and how challenges were met and lists some next steps in the FAIRification work of the ESS data archive at Sikt that will add further value to ESS as a service.

Like for deliverable report D5.13, the main target audiences of this report are other data archives and interested data and survey managers as well as data scientists. Chapter 6, Recommendations, is however of value to a wider audience than the target audience of the deliverable.

Of the other SSHOC work packages, the document is mainly relevant to WP3 and WP7.

https://www.myscience.org/en/news/wire/upf_becomes_the_new_hub_of_a_management_platform_to_improve_the_european_social_survey-2022-upf ; My ESS includes the external deposit service for data from national teams to the ESS data archive. (Accessed April 2022)

2. The ES Workflow

The data and metadata deposited by the ESS national teams through the MyESS deposit service are collected in the Azure storage. The data in the cloud is processed using Jupyter Notebooks. The metadata is imported into the Colectica repository and curated. Upon completion of data and metadata processing and quality check, they are published to the respective repositories. Upon publishing, they are assigned DOIs and are available for search and download.

Figure 1 shows how the data and documentation is organized at various stages in the process. The blue, red and gray elements illustrate components that lie outside of Task 5.5. However, as they are important in understanding the structure of the data storage and API solutions developed within the task, they have been included.

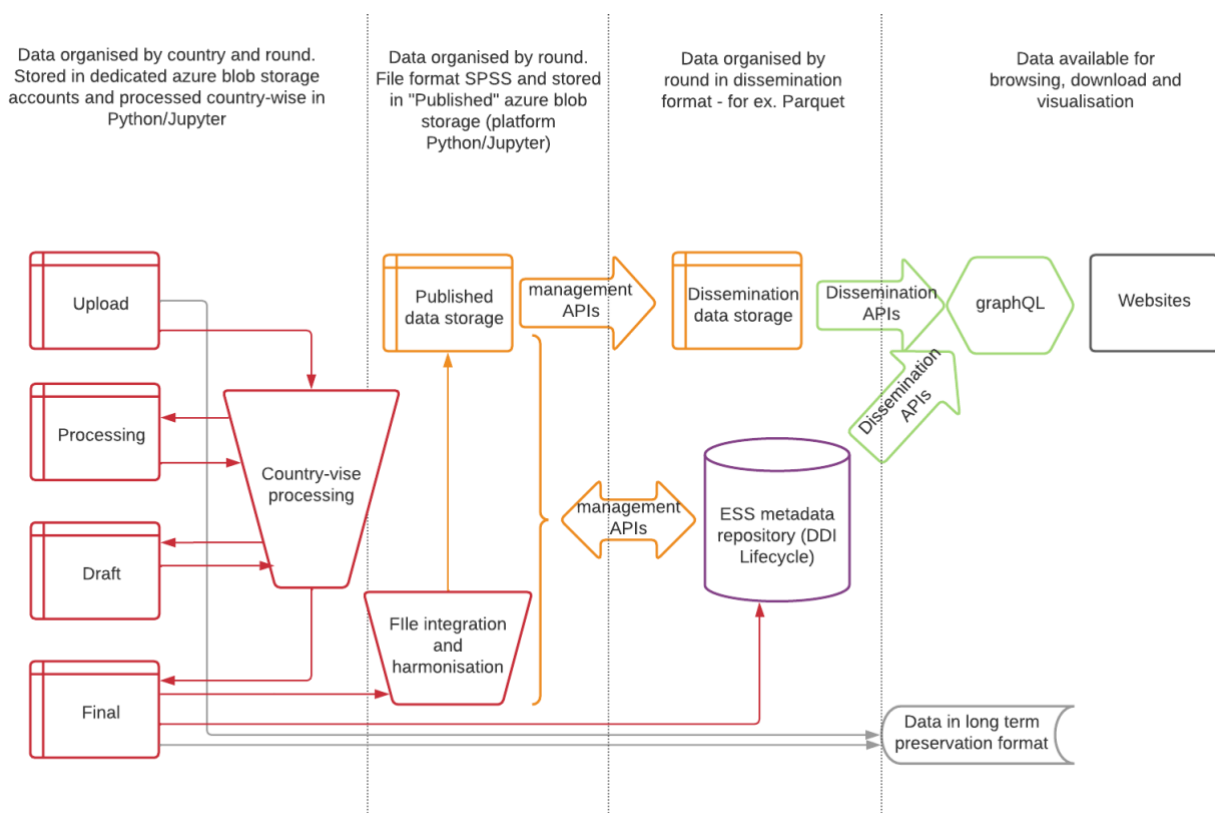


FIGURE 1: ESS DATA AND DOCUMENTATION WORKFLOW FROM DATA INGEST TO DISSEMINATION

3. System Architecture

The ESS as a service implements the Open Archival Information System (OAIS) model⁴ in the cloud – and covers Ingest, Data Management, Preservation and Dissemination processes of digital asset preservation and dissemination.

Figure 2 gives an overview of the system architecture of the new ESS service with main components - metadata repository, data storage in cloud, dissemination service, ESS API and clients metadata curation application, data processing platform in cloud and the ESS data portal with metadata search and landing pages, and data analysis and download, functionalities with single sign on integration.

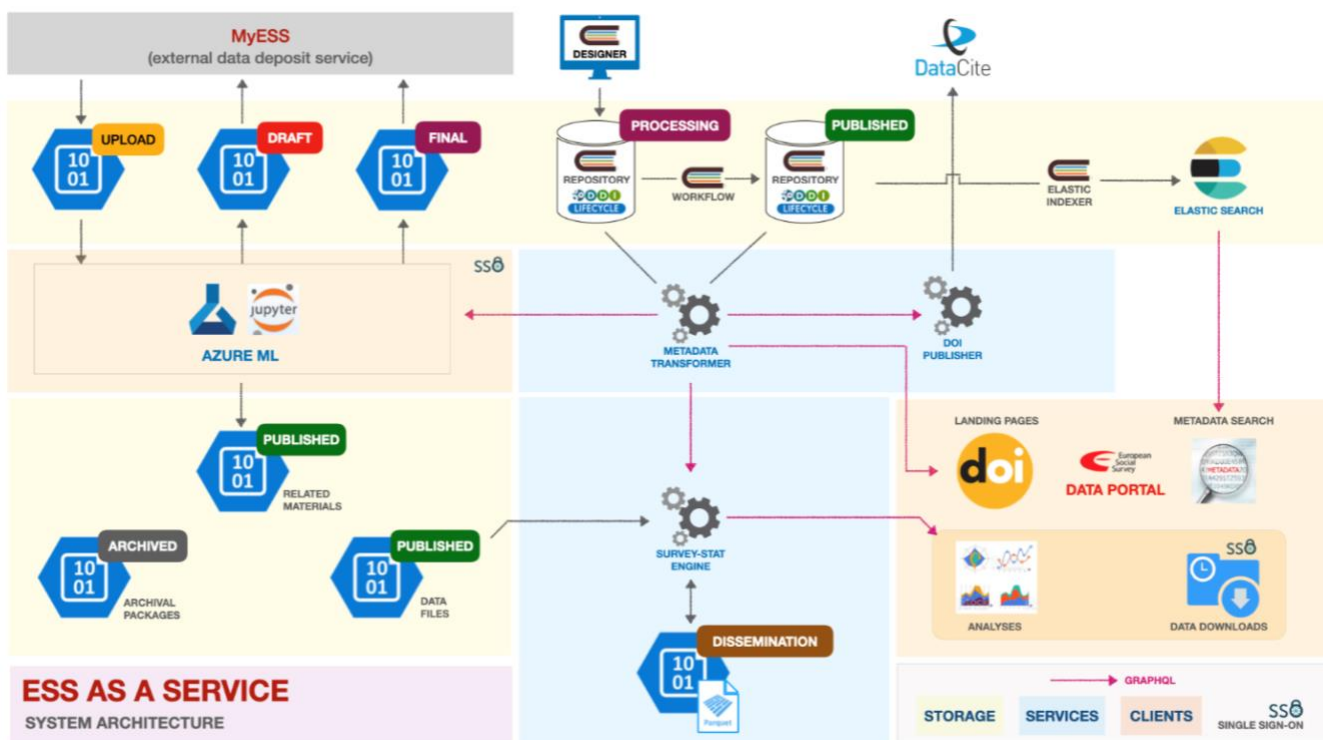


FIGURE 2: ESS AS A SERVICE – SYSTEM ARCHITECTURE

⁴ OAIS: <http://www.oais.info/> (Accessed April 2022)

3.1 Ingest

The external data deposit service of the MyESS project management system interfaces with the data storage in Azure to upload SIPs to “Upload” blob storage and sync back the processed data and related materials from the “Draft” and “Final” blob storages for review purposes. This supports complete transparency of data processing to the data depositors.

3.2 Data Management

Data Processing Platform: Azure Machine Learning is used as collaborative workspace to run Jupyter Notebooks⁵. Standardized data processing Jupyter notebooks written in Python⁶ are available across the data processing team, streamlining data processing workflow while eliminating repeated manual work. The Python scripts access data stored in the Azure blob storages⁷ and metadata stored in Colectica repository via GraphQL API, ensuring that data processing is done in alignment with relevant metadata.

Once the final version of data and related materials have been approved (via external services) by the depositors, they are published to their corresponding “Published” blob storages. The related materials are published as PDFs while the country wise datafiles are integrated into the round datafile and published as .SAV files with metadata-based machine-actionable names wherein the naming convention followed is *<agencyID>_<physicalInstanceID>_v<physicalInstanceVersion>.sav* .

Metadata Curation: The metadata deposited by the country wise data deposit team are imported into Colectica Designer⁸, a metadata curation desktop application based on DDI-lifecycle standard, and stored in the “Processing” instance of the Colectica Repository⁹. Once the curation is completed, metadata is published to the “Published” instance of the Colectica Repository via Colectica Workflow¹⁰.

DOI Publisher, a batched microservice, mints DOI for the newly published metadata from DataCite¹¹.

⁵ <https://docs.microsoft.com/en-us/azure/machine-learning/how-to-run-jupyter-notebooks> (Accessed April 2022)

⁶ <https://www.python.org/> (Accessed April 2022)

⁷ <https://docs.microsoft.com/en-us/azure/machine-learning/how-to-access-data> (Accessed April 2022)

⁸ <https://www.colectica.com/software/designer/> (Accessed April 2022)

⁹ <https://www.colectica.com/software/repository/> (Accessed April 2022)

¹⁰ <https://www.colectica.com/software/workflow/> (Accessed April 2022)

¹¹ <https://support.datacite.org/reference/mint-doi> (Accessed April 2022)

3.3 Preservation

An archival package containing data, metadata, and related materials in respective open long-term archival formats — CSV, XML and PDF/A — is created and stored in the long-term preservation storage in Azure once the data and metadata are published.

3.4 Dissemination Backend:

Elastic Search: Published metadata is indexed by the Elastic Indexer¹² and the elastic search index¹³ is stored in the Elastic Cloud¹⁴ to enable a powerful search capability. The Elastic Index API¹⁵ is wrapped in GraphQL API for consumption by clients.

Metadata Transformer: Metadata stored in the Colectica Repository is accessible via a Colectica REST API¹⁶. However, to support flexible and efficient metadata consumption and metadata driven services, a microservice, Metadata Transformer, was built. Metadata Transformer reads metadata as JSON SET and transforms it as required by consuming services and clients and exposes metadata via the GraphQL API.

SurveyStatEngine: Data analysis, custom datafile creation and datafile download functionalities are performed by SurveyStatEngine. The statistical actions are performed on datafiles stored in Apache Parquet¹⁷ format in the “Dissemination” blob storage while the required metadata is retrieved via GraphQL API. If a datafile is not available in parquet format, SurveyStatEngine reads from “Published” blob storage and creates the required datafile in the parquet format. It also converts data into various formats, such as CSV, SAS and SPSS, for download.

3.5 ESS Data Portal:

Search: A simple, user-friendly search user interface, powered by an elastic search engine in the backend, gives ESS users an easy way of searching and finding exactly what they are looking for in the 10 ESS rounds, 60 data files, 18,139 variables and 213 question texts. Figure 3 shows the variables search results for the search term “Income”.

¹² <https://docs.colectica.com/repository/technical/elastic-indexer/> (Accessed April 2022)

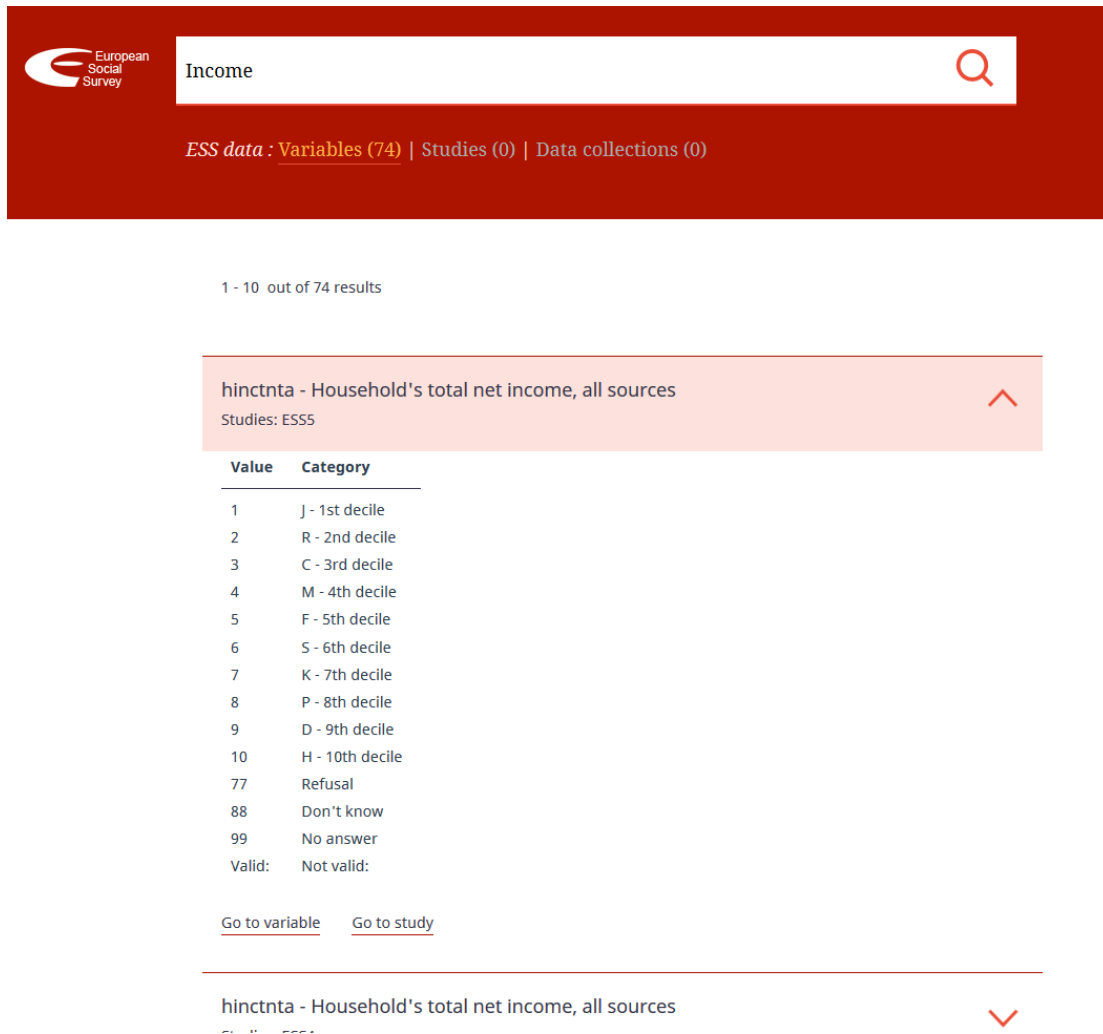
¹³ <https://docs.colectica.com/repository/technical/elasticsearch/> (Accessed April 2022)

¹⁴ <https://www.elastic.co/cloud/> (Accessed April 2022)

¹⁵ <https://www.elastic.co/guide/en/elasticsearch/reference/current/docs-index.html> (Accessed April 2022)

¹⁶ <https://docs.colectica.com/repository/functionality/rest-api/web-services/> (Accessed April 2022)

¹⁷ <https://parquet.apache.org/> (Accessed April 2022)



The screenshot shows a search interface with a red header. On the left is the European Social Survey logo. The search bar contains the text 'Income' and a magnifying glass icon. Below the search bar, it displays 'ESS data : Variables (74) | Studies (0) | Data collections (0)'. Underneath, it says '1 - 10 out of 74 results'. A search result is shown in a light pink box, titled 'hinctnta - Household's total net income, all sources' with 'Studies: ESS5' and an upward arrow icon. Below this is a table with two columns: 'Value' and 'Category'. The table lists values from 1 to 10, 77, 88, and 99, corresponding to deciles and other categories. At the bottom of the result box are links for 'Go to variable' and 'Go to study'. Below the result box, the search result is repeated with a downward arrow icon.

Income

ESS data : [Variables \(74\)](#) | [Studies \(0\)](#) | [Data collections \(0\)](#)

1 - 10 out of 74 results

hinctnta - Household's total net income, all sources
Studies: ESS5

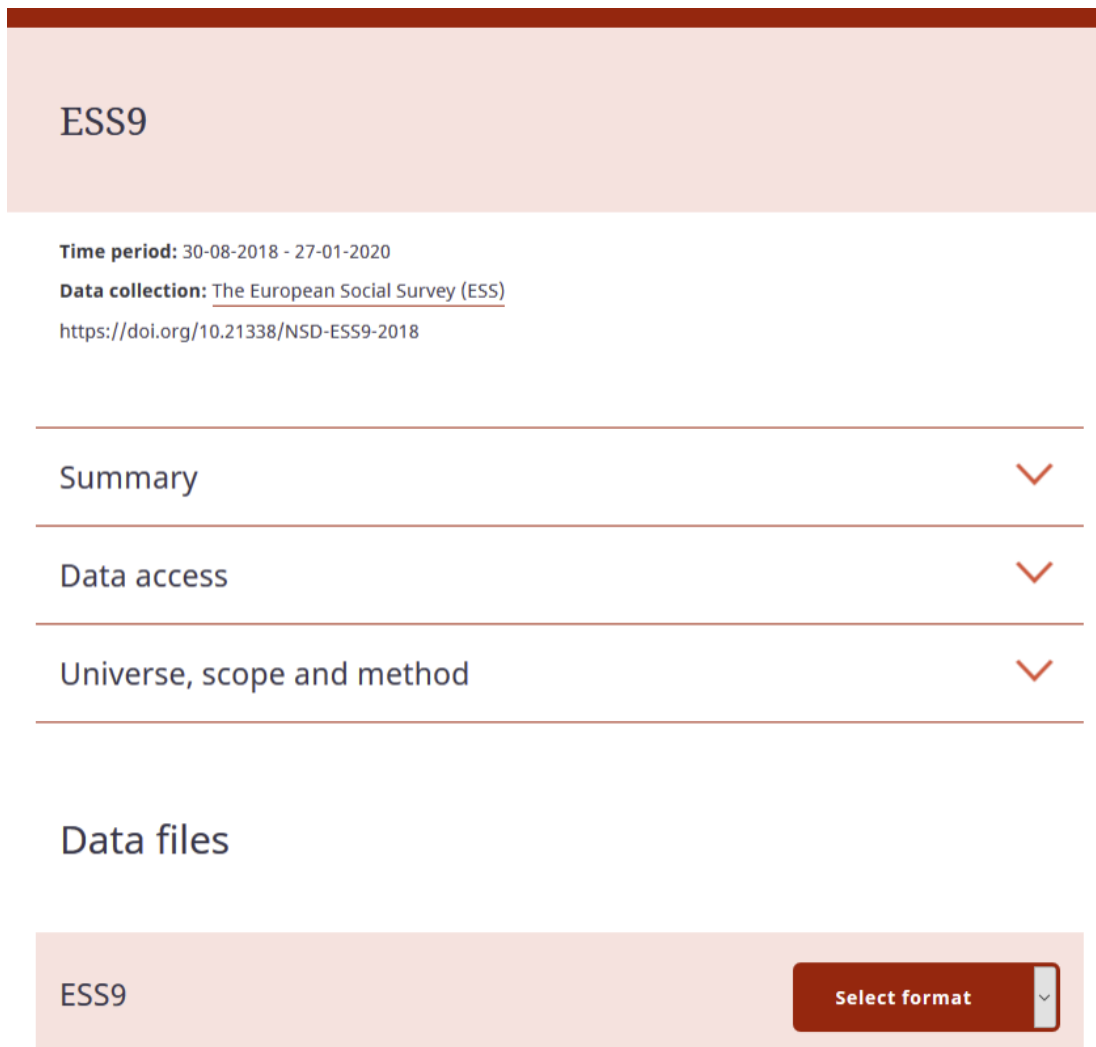
Value	Category
1	J - 1st decile
2	R - 2nd decile
3	C - 3rd decile
4	M - 4th decile
5	F - 5th decile
6	S - 6th decile
7	K - 7th decile
8	P - 8th decile
9	D - 9th decile
10	H - 10th decile
77	Refusal
88	Don't know
99	No answer
Valid:	Not valid:

[Go to variable](#) [Go to study](#)

hinctnta - Household's total net income, all sources

FIGURE 3: SCREENSHOT OF VARIABLES SEARCH RESULTS

DOI landing pages: The DOIs resolve to the landing pages that display metadata stored in the cloud (Colectica Repository) via the GraphQL API. As the coverage of the metadata elements is expanded, the landing pages are updated to display these elements in a comprehensible way. Figure 4 shows the DOI landing page for a round (ESS9).



ESS9

Time period: 30-08-2018 - 27-01-2020

Data collection: The European Social Survey (ESS)

<https://doi.org/10.21338/NSD-ESS9-2018>

Summary

Data access

Universe, scope and method

Data files

ESS9

Select format

FIGURE 4: SCREENSHOT OF DOI LANDING PAGES FOR ESS ROUNDS (HERE ESS9)

Data Analysis: The landing pages are continuously developed to support rich visualization of variable level statistics such as frequencies, timelines, and geographic distribution; as well as data analysis such as cross-tabulation of maximum three variables and bar charts.

Data Download: Data download requires users to login to the data portal in case they are not already logged in. The various types of data made available for download are contact form data, country specific data, parents' occupation string data, parents' occupation numeric data, sample data, interviewer data, experiment data, weights, interview time data and main data files. All data files can be downloaded in supported formats as single files.

A cumulative data download wizard for the main integrated data files, with the same functionalities as the one currently existing in the ESS website¹⁸, will allow users to create custom download files from selected rounds, countries and variables.

User Management and Login: The integration of the user management and authentication solutions with the landing pages gives new and existing ESS users broad and easy access to the new services via their Google, Edugain or Feide (the Norwegian version of Edugain) accounts – without prior registration as ESS users, as illustrated by Figure 5. Users can also register as ESS users by creating an ESS-specific account, should they do not have/wish to use Google/Edugain or Feide login. The new features ensure that ESS services are more accessible and reusable than before.

Create one here.'. At the bottom right, there is a blue question mark icon followed by the text 'Help'." data-bbox="134 349 864 711"/>

Log in
to continue to ess-search...

Log in with Feide/eduGAIN

Log in with Google

Log in with username/password

[Personvern og brukerbetingelser](#)

Don't have an account on any of these providers? [Create one here.](#)

[? Help](#)

FIGURE 5: LOGIN VIA FEIDE/EDUGAIN, GOOGLE OR ESS SPECIFIC USERNAME/PASSWORD

¹⁸ <https://www.europeansocialsurvey.org/downloadwizard/> (Accessed April 2022)

4. Protocols

ESS as a Service has adopted relevant most modern yet mature industry standard protocols, wherever applicable, to ensure robustness, performance, sustainability, and interoperability. The protocols are of various types – metadata standard, file formats, API, cloud infrastructure procurement, communication and so on.

4.1 DDI-Lifecycle

DDI-Lifecycle is an international stable metadata standard¹⁹ that supports data documentation and management across the entire lifecycle, from conceptualization to dissemination. DDI encourages comprehensive description of data for discovery and analysis and supports effective data sharing. Because DDI is a structured standard, it facilitates machine-actionability and interoperability and it can be used to drive systems. Another feature of DDI is its focus on metadata reuse; “enter once, use often” means you can reuse metadata over the course of the data lifecycle to avoid costly duplication of effort.²⁰

4.2 GraphQL API

The APIs uses a modern API language/interface, GraphQL, to express search and other elements. GraphQL is designed to support strong types and offers flexibility for various clients consuming GraphQL APIs. Moreover, GraphQL APIs encourage decoupling of the externally available API and its types and relations from internal implementation details. Decoupling ensures that API consumers are shielded from internal APIs and systems, offering a stable, flexible and domain-centric API likely to outlive many of its internal technological implementations at any given time.

The dissemination API supports searches against the DDI Lifecycle-based metadata repository. Running a search for a set of terms returns a result-set in the form of a list of studies, modules, or variables, depending on the search query. From the result set, clients may retrieve complete metadata records for all studies or variables in the result-set, including relevant relations expressed in the underlying DDI-Lifecycle repository.

¹⁹ <https://ddi-lifecycle-documentation.readthedocs.io/en/latest/User%20Guide/Introduction.html> (Accessed April 2022)

²⁰ <https://ddialliance.org/learn/why-use-ddi> (Accessed April 2022)

4.3 OIDC/OAuth (for login)

The ESS user login is built on OpenID Connect²¹ (OIDC) and OAuth 2.0²² (Open Authorization) protocol. OAuth 2.0 is an industry-standard protocol for authorization by which users can grant websites access to their information held by other services. OIDC is a thin identity layer built on OAuth 2.0 which allows users to login into multiple services using single user identity such as google or EduGain/Feide. OIDC is API-friendly and allows clients from various platforms to request and receive information about authenticated sessions and end users. OAuth 2.0 capabilities are integrated in the OIDC protocol.

4.4 Terraform

The cloud infrastructure of ESS as a service is provisioned and managed as Infrastructure as Code (IaC)²³ via Terraform²⁴. Terraform is an open-source IaC software tool where one can create, read, manage, and destroy infrastructure using declarative configuration files. Terraform is human-readable machine-actionable code to automate cloud resource management. Being an Open-Source language and with support for various cloud solutions mitigates the risk of vendor lock-in. Provides a Uniform Syntax for Infrastructure as a Code increasing sustainability. The language is highly expandable, making it support newer cloud services as they are offered by the service providers. It can generate dependency graphs thus helping in infrastructure maintenance work.

4.5 Apache Parquet

The ESS dissemination service uses Parquet file format for data analysis and download. Apache Parquet is an open-source columnar data storage file format designed to support fast data access and analysis. In Parquet, compression is performed column by column and supports flexible compression options per data type. Parquet is optimized for performance and supports data schema evolution.²⁵

²¹ <https://openid.net/connect/> (Accessed April 2022)

²² <https://oauth.net/2/> (Accessed April 2022)

²³ <https://docs.microsoft.com/en-us/devops/deliver/what-is-infrastructure-as-code> (Accessed April 2022)

²⁴ <https://www.terraform.io/> (Accessed April 2022)

²⁵ <https://www.upsolver.com/blog/apache-parquet-why-use> (Accessed April 2022)

4.6 JSON

The ESS dissemination service uses JSON format for metadata retrieval. JSON (JavaScript Object Notation)²⁶ is an open text-based data-interchange format which is easy for humans to read and write, while machines can easily parse and generate it.

4.7 HTTPS

The ESS Data Portal uses https:// communication protocol. HTTPS (Hypertext Transfer Protocol Secure)²⁷ is a secure version of the HTTP protocol that uses the SSL/TLS protocol for encryption and authentication. The main benefits of using HTTPS protocol are that it adds security and trust by protecting from security attacks launched from compromised networks. Data is protected during transit. Protects the website from phishing and other attacks.

5. Recommendations

This section encompasses recommendations to similar data archives based on ESS as a service's workflow, architecture, protocols, should they want to implement similar infrastructure.

5.1 Agile Approach, Cross Functional Teams

Building a modern, well-functioning and robust digital infrastructure for ESS and similar survey data collections requires skills from several domains. It is recommended organizing this type of work under a largely autonomous cross functional team with competence/expertise on:

- Data management and data curation,
- Research needs (methodological and other),
- System architecture and API-design,
- Software development,
- Data science tools,
- Dev/ops,
- Metadadata standards (typically DDI Lifecycle),
- Procurement, and
- Cloud infrastructure.

²⁶ <https://www.json.org/json-en.html> (Accessed April 2022)

²⁷ <https://en.wikipedia.org/wiki/HTTPS> (Accessed April 2022)

Furthermore, the Task 5.5 team recommends that the team is equipped with sufficient mandate and resources to be able to iterate quickly through different proof-of-concept-approaches, quickly analyse results/feedback (preferably with important stakeholders such as researchers) and be prepared and equipped to make mistakes along the way.

An agile approach should not be limited to the core team, but also apply to stakeholders within and outside of the organization; building trust by aligning on goals but being flexible on the path towards the goals we see as crucial.

The current solution is more than anything a result of structured experimentation, communication, trial and error, and it ended up both more flexible and more powerful and robust than first envisioned. Early on, the opportunity to use the project as a learning experience was taken, where collaborative experiments and structured analysis/re-planning resulted in surprising results and substantial learning across the team and organization beyond.

Daily digital stand-up-meetings (15-30 mins) between core team members were held. Key experts and/or leaders from outside the team were frequently invited to speed up information gathering and decision making.

Multiple workshops and presentations were organised regularly to cross-pollinate technical and domain knowledge. Noteworthy results include increased competence on research needs and methodology among technical team members and increased technical creativity/knowledge among data curators. This also enhanced communication and team cohesiveness.

5.2 API First

APIs were front and centre of the work of the implementing team at Sikt. This included both Colectica APIs and APIs that were developed throughout the project. Ultimately, APIs need to support all user-needs, and coordinate behaviour and content from the backend to fulfil those needs.

During the daily stand-up meetings, APIs were typically the focus of attention, enabling (after a while) a common language for communication within the core team.

Our choice of GraphQL as the API language supported and reinforced API communication for several reasons:

- Non-technical people can create API-requests themselves, and experiment with results,
- The language is quite expressive, albeit simple enough to be understood by most, and
- GraphQL is particularly useful because it hides implementation details from consumers, resulting in high flexibility for the backend-service developers to refactor and restructure their services without disturbing/coordinating with API users (data curators and frontend developers).

The API is the *only* channel between frontend and backend services, and the API used by in-team developers is the exact same API that is available for other teams and external consumers. We believe

so-called “dogfooding” (i.e., using our own API for everything) is crucial when developing APIs for external use.

5.3 Enabling Metadata Driven Services

5.3.1 Controlled Vocabularies

One of the main purposes to achieve increased FAIRness of ESS data and metadata was to enable building metadata driven systems. Increased interoperability was achieved by implementation of controlled vocabularies and statistical classifications. Several Controlled Vocabularies have been implemented in the metadata repository (refer to list in Table 1). This, together with the use of controlled vocabularies in the ESS data (values and categories), ensures a high degree of interoperability with other data within the social science domain. The DDI Alliance and CESSDA controlled vocabularies are registered in Recommendation 8 of Ten simple rules for making a vocabulary FAIR (S.J.D. Cox et al. 2021).

Table 1. Controlled vocabularies and statistical classifications used in the repositories

CV Name	Publisher	Context used	URL
Keywords	ELSST	Study documentation	https://elsst.CESSDA.eu/
Topic Classification	CESSDA	Study documentation	https://vocabularies.CESSDA.eu/vocabulary/TopicClassification?lang=en
Contributor role	DDI-Alliance	Study documentation	https://vocabularies.CESSDA.eu/vocabulary/ContributorRole?lang=en
Analysis Unit	DDI-Alliance	Study documentation	https://vocabularies.CESSDA.eu/vocabulary/AnalysisUnit?lang=en
Type of instrument	DDI-Alliance	Fieldwork documentation	https://vocabularies.CESSDA.eu/vocabulary/TypeOfInstrument?lang=en
Data source type	DDI-Alliance	Fieldwork documentation	https://vocabularies.CESSDA.eu/vocabulary/DataSourceType?lang=en
Time method	DDI-Alliance	Fieldwork documentation	https://vocabularies.CESSDA.eu/vocabulary/TimeMethod?lang=en
Sampling procedure	DDI-Alliance	Fieldwork documentation	https://vocabularies.CESSDA.eu/vocabulary/SamplingProcedure?lang=en
Date type	DDI-Alliance	Variable documentation	https://vocabularies.CESSDA.eu/vocabulary/DateType?lang=en
ISO 3166-1 country codes	ISO	Study / Variable documentation	https://www.iso.org/
ISO 639-2 language codes	ISO	Variable documentation	https://www.iso.org/
ISCO-08 Occupation codes	ILO	Variable documentation	https://www.ilo.org/public/english/bureau/stat/isco/isco08/

NACE Rev.2 Industry	EUROSTAT	Variable documentation	https://ec.europa.eu/eurostat/web/nace-rev2/overview
NUTS Region codes	EUROSTAT	Variable documentation	https://ec.europa.eu/eurostat/documents/3859598/6948381/KS-GQ-14-006-EN-N.pdf
Ancestry codes	ESS	Variable documentation	http://www.europeansocialsurvey.org/
Education codes	ESS	Variable documentation	http://www.europeansocialsurvey.org/
Religion codes	ESS	Variable documentation	http://www.europeansocialsurvey.org/
Marital status	ESS	Variable documentation	http://www.europeansocialsurvey.org/

5.3.2 Versioning of Data and Documents

The ESS versioning for data files and documents will be updated from 2 level versioning to a type of semantic versioning. The three levels thus being MAJOR (inclusion of new country data), MINOR (Changes in data or metadata that will influence the use of data or the results of data analysis) and PATCH (insignificant changes such as spelling errors). A new DOI will be minted for changes in MAJOR and MINOR levels. As changes at PATCH level does not influence the results of data analysis or how these are interpreted, a PATCH update will not lead to a new DOI.

5.3.3 Versioning of Metadata

ESS metadata elements are versioned at one level only. Every version of an item committed to the repository is saved, allowing clients to retrieve a full version history of any item in the repository. Each version includes additional information such as, for instance, information about the user who committed the version, the date and time the version was committed and an optional message describing the reason for the change (Beuster et al., 2019).

It is recommended to adopt and implement sector standard controlled vocabularies and statistical classifications along with a good versioning regime for data, documents and metadata.

5.4 Build Data Solutions for Download and Analysis

NSD (now: Sikt) has developed in-house solutions for data analysis and download services (such as Nesstar) for almost 40 years. In the beginning, this was demanding work, and analytical functionality had to be implemented more or less from scratch. Today, the barrier to build software solutions for analysis and data download is dramatically reduced with the proliferation of Python and other open-source data science ecosystems.

With Python (which in contrast to R, Stata and SPSS is a complete programming language and software engineering environment), it is now fairly trivial to build services for simple and more advanced data analysis in house. It is now a matter of packaging relevant open-source libraries (Pandas, Statsmodels, SciPy, NumPy) in simple and small applications, and various analyses and data assurance techniques that need to run directly on data can be implemented with ease.

In retrospect, this project played a significant role in adopting novel and open-source solutions where it was earlier relied on in-house-developed and other proprietary and less flexible software products.

The *recommendation* here is thus to take a structured approach to adopting Python and associated packages not only at the data curators' desktops and in data processing pipelines, but also in backend services.

5.5 Connect Data and Metadata

Traditionally (with the notable exception of Nesstar), data and metadata live in separate systems and technology stacks. Keeping the two in sync and calibrated has traditionally been a manual, time-consuming and error-prone undertaking. The adoption of Colectica (and migration from Nesstar to Colectica) was a step back in this sense, since Colectica itself does not support continuous data-metadata integration the same way Nesstar did.

However, Colectica does support DDI Lifecycle, and thereby versioned metadata elements.

It was decided to utilize this versioning feature, along with an *immutable, flat-file approach* to data to ease and automate automated and semi-automated data/metadata validation.

Since data sets are now kept as flat files, it is trivial to keep/retain earlier versions of data files. For replication and reproducibility purposes, keeping older data file versions around and available can be crucial. The Colectica versioning scheme is represented/exposed through the APIs, enabling API consumers to reason about version history, and consume content and analyses from all data versions not just the last one.

Again, GraphQL can be recommended as API language, since GraphQL queries *from* the data processing environments in Jupyter Notebooks are utilised to talk directly to the metadata services (GraphQL libraries exist for Python). This way, data curators can calibrate data with metadata real-time, without leaving their data processing environment.

A set of standardized quality-assurance Jupyter Notebooks with associated libraries has been developed with built-in data-metadata calibration/validation support, so data curators do not need to write GraphQL API calls directly.

6. Conclusions

The development of the components described and documented in the report has given ESS a state-of-the-art infrastructure for data management and dissemination purposes. Major outcomes of the Task 5.5 are:

- Increased FAIRness of ESS data and services,
- Enhanced user services (improved landing pages & search, login options),
- New modern cloud-based data management infrastructure,
- Transparency with the data depositors,
- Streamlined data processing workflow, and
- Updated metadata and data processing skills among ESS data curation team.

Challenges that were met underway included lack of time and resources, new technology (Azure cloud infrastructure) and complex problem-space. These were overcome by running the development project as a part of Sikt's larger Data and Metadata Platform infrastructure project, which was consolidation of multiple similar projects. Some of the measures undertaken by the project team to overcome challenges were:

- Building a cohesive team with shared vision, motivation to understand problem-space and ownership of solutions developed,
- Investment in increasing/cross-pollinating domain and technical competence,
- Implementation of Design Thinking in praxis wherein, a high-functioning cross-functional team with relevant domain and technical expertise was able to discuss, debate and implement solutions quickly,
- Successful lean and agile project management,
- Effective use of available resources, and
- Finetuned prioritization of deliverables based on deadlines and availability of resources.

The hope is that the report will give other data archives sufficient information to examine, replicate parts of solution and test the technical achievements documented.

Some of the next steps that will add value to ESS as a service are:

- Expanding the coverage of DDI-Lifecycle metadata elements used for data documentation,
- Revising applied Controlled Vocabularies and applying for other relevant metadata elements,
- Adoption of ELSST,
- Increasing quality of metadata to support machine actionability and enriched data visualization,
- Continuous development of the service to keep it relevant, modern and support changing user needs, and

- Exploring DDI-CDI metadata standard for newer possibilities to increase FAIRness of ESS data, refer to Gregory et al. (2021) for a use case on possible benefits from applying the model to the complex ESS data.

Some of the above steps will be taken on as part of the general development work of the data archive. Other elements will depend on access to further funding. An example of this would be application of the DDI-CDI metadata model, which for a big survey like the ESS could reduce the workload with minor and larger data upgrades and to give cost savings.

7. References

Agasøster, Bodil, Bergset, Gyrid Havåg, Beuster, Benjamin, Bidargaddi, Archana, Risnes, Ørnulf, Skjåk, Knut Kalgraff and Stavestrand, Eirik (2022). SSHOC D3.13 Recommendations for a FAIR compliant integrated data and metadata repository (ESS as a service). Zenodo (*in publication*)

Beuster, B. & Agasøster et al. (2019), Report on populating a Question Variable Data base (QVDB) with ESS questions and variables, as well as having other SERISS survey partners /third parties test the feasibility of such a tool. Deliverable 4.2 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables.

CESSDA (2008): CESSDA user guide for digital preservation. [Microsoft Word - D10.4 Data Formats.doc \(cessda.eu\)](https://www.cessda.eu/Microsoft%20Word%20-%20D10.4%20Data%20Formats.doc).

The DDI Alliance (2020), DDI 3.3 documentation, available: <https://ddi-lifecycle-documentation.readthedocs.io/en/latest/User%20Guide/Introduction.html>.

Cox, S.J.D.; Gonzalez-Beltran, A. N.; Magagna, B & Marienscu, M.-C. (2021), Ten simple rules for making a vocabulary FAIR, <https://doi.org/10.1371/journal.pcbi.1009041> ,

European Commission, Directorate-General for Research and Innovation, *Turning FAIR into reality: final report and action plan from the European Commission expert group on FAIR data*, Publications Office, (2018), <https://data.europa.eu/doi/10.2777/54599>.

Gregory, Arofan, Hodson, Simon, & Wackerow, Joachim. (2021). The Role of DDI-CDI in EOSC: Possible Uses and Applications. Zenodo. <https://doi.org/10.5281/zenodo.4707263>

List of Figures

[Figure 1: ESS data and Documentation workflow from data ingest to dissemination](#)

[Figure 2: ESS as a service – System architecture](#)

[Figure 3: Screenshot of variables search results](#)

[Figure 4: Screenshot of landing pages for ESS rounds \(here ESS9\)](#)

[Figure 5: Login via Feide/Edugain, Google or ESS specific username/password](#)

List of Tables

[Table 1. Controlled vocabularies and statistical classifications used in the repositories](#)