Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782 Start Date of Project: 01/01/2019 Duration: 40 months

Deliverable 1.6 Data Management Plan - DMP

Dissemination Level	PU
Due Date of Deliverable	30/06/2020 (M18)
Actual Submission Date	30/06/2020
Work Package	WP 1- Project Management and Administration
Task	T1.1 Administrative Project Management
Type	ORDP: Open Research Data Pilot
Approval Status	Waiting EC approval
Version	V1.0
Number of Pages	p.1 – p.54

Abstract:
Data Management Plan (DMP) is one of the main SSHOC project management documents. It gives an overview of data handling during and after the SSHOC project implementation and provides information on creation, management, sharing, curation and preservation of the project data and data usage for implementation of project activities.

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020); H2020-INFRAEOSC-04-2018, under the agreement No. 823782

History

Version	Date	Reason	Revised by
0.1	20/07/2019	Structure draft and sections options	Martina Drascic
0.2	19/05/2020	Redrafting to include EC template sections	Vanja Komljenovic, Ivana Ilijasic Versic, Irena Vipavc Brvar
0.3	05/06/2020	New structure and guidelines prepared	Vanja Komljenovic
0.4	11/06/2020	Review by all PMB members	Vanja Komljenovic
0.5	22/06/2020	PMB comments addressed Draft input per WP/Task collected	Vanja Komljenovic
0.6	23/06/2020	Marketplace input added	Vanja Komljenovic
0.7	26/06/2020	Data Communities input and SHARE tools input added	Vanja Komljenovic
0.8	29/06/2020	ESS and Wage Indicator input added, comments addressed	Vanja Komljenovic
0.9	29/06/2020	Revision of text in all sections	Vanja Komljenovic, Ivana Ilijasic Versic
1.0	30/06/2020	Final editing; final version	Martina Drascic, Vanja Komljenovic

Author List

Organisation	Name	Contact Information
CESSDA ERIC	Ivana Ilijasic Versic	ivana.versic@cessda.eu
CESSDA ERIC	Vanja Komljenovic	vanja.komljenovic@cessda.eu
CESSDA ERIC	Martina Drascic	martina.drascic@cessda.eu
CESSDA / ADP-UL	Irena Vipavc Brvar	irena.Vipavc@fdv.uni-lj.si
CESSDA /GESIS	Elizabeth Lea Bishop	elizabethLea.Bishop@gesis.org
CLARIN ERIC	Daan Broeder	daan.broeder@di.huc.knaw.nl
CLARIN ERIC	Maria Eskevich	maria@clarin.eu
CLARIN ERIC	Kea Tijdens	k.g.tijdens@uva.nl
CLARIN ERIC	Dieter Van Uytvanck	dieter@clarin.eu

CLARIN / Athena	Maria Gavriilidou	maria@athenarc.gr
CNR	Emiliano Degl'Innocenti	emiliano.deglinnocenti@cnr.it
CNR	Monica Monachini	monica.monachini@ilc.cnr.it
CNRS	Nicolas Larrousse	nicolas.larrousse@huma-num.fr
DAI	Wolfgang Schmidle	wolfgang.schmidle@dainst.de
DARIAH ERIC	Laure Barbot	laure.barbot@dariah.eu
DARIAH ERIC	Erzsébet Toth-Czifra	erzsebet.toth-czifra@dariah.eu
DARIAH / OEAW	Klaus Illmayer	klaus.illmayer@oeaw.ac.at
DARIAH / OEAW	Matej Ďurčo	matej.durco@oeaw.ac.at
DARIAH / UGOE	Stefan Buddenbohm	buddenbohm@sub.uni-goettingen.de
DARIAH / UGOE	Nanette Rissler Pipka	rissler-pipka@sub.uni-goettingen.de
EEP PSE	Lana Yoo	lana.yoo@psemail.eu
ESS / UPF	Diana Zavala Rojas	diana.zavala@upf.edu
KNAW	Tom Emery	emery@nidi.nl
KNAW	Marion Wittenberg	marion.wittenberg@dans.knaw.nl
LIBER	Vasso Kalaitzi	vasso.kalaitzi@kb.nl
NG	Joseph Padfield	joseph.padfield@ng-london.org.uk
SHARE ERIC	Johanna Bristle	bristle@mea.mpisoc.mpg.de
SHARE ERIC	Fabio Franzese	franzese@mea.mpisoc.mpg.de
SHARE ERIC	Stefan Gruber	gruber@mea.mpisoc.mpg.de
SHARE ERIC	Annette Scherpenzeel	scherpenzeel@mea.mpisoc.mpg.de
SHARE ERIC	Stephanie Stuck	stuck@mea.mpisoc.mpg.de
SHARE ERIC	Luzia Weiss	l.weiss@mea.mpisoc.mpg.de
TRUST-IT	Marieke Willems	m.willems@trust-itservices.com
UoN	Cees van der Eijk	cees.van_der_eijk@nottingham.ac.uk
UoY-ADS	Holly Wright	holly.wright@york.ac.uk

Executive Summary

The Deliverable 1.6 Data Management Plan (DMP) provides structured description of data collected, created, and used for and by the SSHOC project. It also describes the procedures and policies as well as data protection and preservation mechanisms used during the implementation of the SSHOC project and planned in the post implementation period.

This document follows the structure of Horizon 2020 DMP template[1] and outlines the key points regarding:

- purpose of the data collection in relation to SSHOC project objectives and activities,
- types and formats of SSHOC project data,
- reuse of existing data,
- origin of data and data usefulness,
- alignment with FAIR principles,
- resources needed and responsibilities within the project,
- data security,
- ethical and intellectual property aspects related to data.

The SSHOC DMP describes the overall methodology, standards and technical aspects to enable sound and tenable data management system for the SSHOC project. It presents the SSHOC project datasets, the basic rules of conduct in handling project data, and serves as a guide for members of the SSHOC project consortium.

SSHOC recognizes the following data handled throughout the project duration:

- Survey data
- Case studies / pilots data
- Tools and Service data
- SSHOC Marketplace data
- SSHOC user communities data
- Other data

Descriptions of the recognised data follow the same structure throughout the document giving the overview of the key points for each type of data.

The SSHOC Data Management Plan is a living document and shall be updated continuously throughout the project in line with the new information gathered via conducting the project activities.

[1] H2020 templates: Data management plan v2.0 – 15.02.2018 https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm#A1-template [24.06.2020]

Abbreviations and Acronyms

API	Application Programming Interface
AUSSDA	The Austrian Social Science Data Archive
CAT	Computer-Aided Translation
CAWI	Computer-assisted web interviewing
CDI	Customer data integration
CIDOC-CRM	Comité International pour la Documentation – Conceptual Reference Model
CTS	Core Trust Seal
DANS	Data Archiving and Network Services
DBSS	Dried Blood Spot Samples
DDI	Data Documentation Initiative
DMP	Data Management Plan
DOI	Digital Object Identifier
DPO	Data protection officer
EMMs	Ethnic and Migrant minorities
EOSC	European Open Science Cloud
ERIC	European Research Infrastructure Consortium
E-RIHS	European Research Infrastructure for Heritage Science
ESS	European Social Survey
ETHMIGSURVEYDATA	The International Ethnic and Immigrant Minorities' Survey Data Network
EU	European Union
EURHISFIRM	Historical high-quality company-level data for Europe
EVS	European Values Study
FAIR	Findable, Accessible, Interoperable, Reusable
GA	Grant Agreement
GDPR	General Data Protection Regulation
GESIS	Leibniz Institute for the Social Sciences
GGP	Generations and Gender Programme
GUI	Graphical User Interface
HQ	Head Quarters
IIIF	International Image Interoperability Framework
IPERION HS	Integrated Platform for the European Research Infrastructure
IPR	Intellectual Property Rights

ISCO	International Standard Classification of Occupations
KNAW	Royal Dutch Academy of Science
LIBER	Ligue des Bibliothèques Européennes de Recherche
LOD	Linked Open Data
MCSQ	Multilingual Corpus of Survey Questionnaires
MT	Machine Translation
NIDI	The Netherlands Interdisciplinary Demographic Institute
NLP	Natural language processing
NSD	Norwegian Center for Research Data
OEAW	Austrian Academy of Sciences
OECD	Organisation for Economic Co-operation and Development
OLA	Operation Level Agreement
PID	Persistent identifiers
PDF	Portable Document Format
RI	Research infrastructure
SAS	Statistical Analysis Software
SHARE	Survey of Health, Ageing and Retirement in Europe
SLA	Service Level Agreement
SPSS	Statistical Package for the Social Sciences
SSH	Social Sciences and Humanities
TEI	Test Encoding Initiative
TGIR	Très grande infrastructure de recherche
TMX	Translation Memory Exchange
SSHOC	Social Science and Humanities Open Cloud
UK	United Kingdom
UPF	Pompeu Fabra University
XML	Extensible Markup Language
WP	Work Package

Table of Contents

1. Introduction

SSHOC - "Social Sciences and Humanities Open Cloud" project (GA No.823782) is one of the 5 ongoing EOSC thematic cluster projects. SSHOC aims to provide a full-fledged Social Sciences and Humanities Open Cloud where data, tools, and training are available and accessible for users of SSH data. It will build the SSH Cloud, maximise the re-use through Open Science and FAIR principles, interconnect existing and new infrastructures, and establish governance for SSH-EOSC. Development, realisation and maintenance of user-friendly tools & services will allow to align, analyse, present and encompass vast heterogeneous collections of SSH data available within data repositories and other institutions in Europe.

SSHOC project objectives and activities are set to enable data collecting, processing, usage, re-usage and accession in line with high standards and generally acknowledged data management regulations and procedures. Project participates in the Pilot on Open Research Data in Horizon 2020, in line with the Commission's Open Access to research data policy for facilitating access, reuse and preservation of its research data. It is also aligned with the European Commission's approach towards research data which is "as open as possible, as closed as needed" and with provisions of the GA.

The SSHOC DMP provides a structured description of data collected, created and used for and by the SSHOC project. It follows the Guidelines on FAIR Data Management in Horizon 2020[2]. Processing of personal data is handled with utmost responsibility, professionality and care. It is in line with the article 39 of the GA and it fully respects the General Data Protection Regulation[3] (GDPR) provisions.

The SSHOC project consortium has the expertise to cover the whole data cycle: from data creation and curation to optimal re-use of data, as well as addressing training and advocacy to increase actual re-use of data.

The research underlying the project publications will be made available under the conditions of the Creative Commons open license.

[2] H2020 Programme Guidelines on FAIR Data Management in Horizon 2020: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf [30.06.2020]

[3]Regulation of the European Parliament- General Data Protection Regulation: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN [30.06.2020]

2. Survey data

Several cross-national surveys have been part of the SSHOC consortia, the biggest ones being SHARE[4] (Survey of Health, Ageing and Retirement in Europe) and ESS[5] (European Social Survey). Data produced during regular data collections (rounds) is used in the SSHOC project for various purposes (from testing tools, creating pilots/case studies, enhancing mechanisms for data protection, security and access, to designing user services).

To avoid repetitions and enable clarity for the purpose of this document, SHARE data has been explained in more detail below as a representative of survey data in the SSHOC project while other surveys' data information was added where appropriate bearing additional value.

2.1 Data Summary

The purpose of the SHARE data is to observe demographic change in a harmonized, cross-nationally comparable way across Europe. Within SSHOC, the processing and dissemination of biomedical data (e.g, blood data and accelerometer data) will be implemented in accordance with the FAIR principles, which closely relates to one of the main objectives of SSHOC, namely making social science data FAIR.

In addition, the occupation ontology developed in Task 3.2 "Essential SSH ontologies and vocabularies" will be incorporated in the SHARE survey and data will be collected in all SHARE countries, by means of the job coder technology.

The ESS is an academically driven cross-national survey that has been conducted across Europe since its establishment in 2001. Every two years, face-to-face interviews are conducted with newly selected, cross-sectional samples. The survey measures the attitudes, beliefs and behaviour patterns of diverse populations in more than thirty nations. In SSHOC, ESS leads and coordinates work in several tasks in WP4 such as T4.1 "A sample management system for cross-national web surveys" or T4.7 "Modelling the SSHOC data life cycle", as well as actively participates in a number of other WPs.

The Generations & Gender Programme Research Infrastructure (GGP) provides scientists and policy makers with high quality and timely data about families and life course trajectories of individuals to enable researchers to contribute insights and answers to current societal and public policy challenges. Users have access to an open-access data source of cross-nationally comparative surveys and contextual data. GGP has several components out of which GGP Survey data will be used in the SSHOC project. Main contribution to the SSHOC project is in tasks T4.4 "Voice recorded interviews and audio analysis" and T4.5 "Social policy APIs for social surveys".

[4] Survey of Health, Ageing and Retirement in Europe (SHARE) website: :http://www.share-project.org/organisation/share-eric.html [30.06.2020]

[5] European Social Survey European Research Infrastructure – ESS ERIC website: https://www.europeansocialsurvey.org/about/ [30.06.2020]

SHARE data

SHARE data is made available to the research community for two of the most common statistical software packages Stata and SPSS. Jobcoder data and generated variable modules for accelerometer and biomarker data will strictly follow the same data management procedures that are applied to the already existing regular SHARE data and thus will be released in the same formats as the regular SHARE data. In addition, potential types and formats for releasing accelerometer data will be further investigated within WP5 Task 1 of SSHOC in order to allow data access following the FAIR guidelines. The job coder data is released together with the regular SHARE data modules. The size depends on the number of interviews conducted.

SHARE Dried Blood Spot Samples (DBSS) data has been collected in SHARE wave 6 (2014/2015) and will be used for SSHOC deliverables. Accelerometer data has been collected during SHARE wave 8 in 2020 until its early termination due to the Covid-19 pandemic. Both will be re-used for SSHOC and thereafter. Jobcoder data has been collected since wave 6. All data can be merged to previously collected SHARE panel data.

ESS data

In SSHOC, ESS planned to commission data collection with human subjects under Task 4.1 "A sample management system for cross-national web surveys". In order to test the operation of the sample management system being developed under Task 4.1 a small sample of adults aged 18+ (c. 50) in up to three EU countries is invited to join a two wave panel and complete up to two short 10 minute surveys on general, non-sensitive, themes. In SSHOC, ESS planned to commission data collection with human subjects under Task 4.1 "A sample management system for cross-national web surveys". In order to test the operation of the sample management system being developed under Task 4.1, a small sample of adults aged 18+ (c. 50) in up to three EU countries is invited to join a two wave panel and complete up to two short 10 minute surveys on general, non-sensitive, themes. Recruitment is planned off the back of the ESS Round 10 pilot and is entirely voluntary.

Informed consent is sought at the time of the invitation. The information sheet is compliant with the requirements of GDPR. Panel members are able to easily withdraw their consent and fully exercise their data access rights. Data protection is a primary consideration and is intended to provide efficient coordination of cross-national panel management between national and central teams.

2.2 FAIR Data

2.2.1 Findable Data

SHARE uses Digital Object Identifiers (DOI) to make datasets permanently identifiable and locatable. The repository of Da|ra links every DOI to a set of metadata, a collection of bibliographical and content information, referring to the registered dataset (title, author, publication date, copyright etc.). The metadata of the SHARE survey data (like questionnaires or show cards) is provided to the users on the SHARE website in a generic English version as well as in each language of the participating countries. Moreover, the DDI based SHARE Data & Documentation Tool is provided to users as a fast, customizable, easy-to-use web interface for browsing and searching the SHARE (meta)data.

2.2.2 Openly accessible data

For both ESS and SHARE and other data infrastructures in the consortium, the research community can access the data openly. Ever since the beginning of the ESS, the principle of free and immediate access to the data for all has been an integral part of the design. The ESS data are available without restrictions, for not-for-profit purposes, pending only a very simple registration procedure[6] on ESS website. For SHARE, data access is given via the following procedure (all details are available on website):

1) The user first requests access to the data by email, fax, or mail with credentials of being a scientist from a known scientific institution (university, research institute, research department of a public policy institute) or detailed information about the scientific project for which the data is intended to be used. Non-EU users also need to sign the user statement binding them to EU data protection standards.

2) Upon acceptance of the credentials, within a few working days, access will be given to the secure SHARE Research Data Centre website via a personal user ID and a password.

3) The data can then be downloaded by the individual user after successful registration. SHARE will support users by a website with all public information and by a combination of central and national support points that answer questions and respond to user requests.

For both infrastructures, data is deposited in public data archives.

In the SHARE case, the only preconditions for data access are scientific affiliation and the submission of a user statement containing identity details of the user (name, email, scientific institution) and a signature for confirming the EU data protection regulations. In case special requirements for biomedical data is needed, these will be evaluated throughout the SSHOC project within Task 5.1 Data access protocols for biomedical data will be published with Deliverable D5.2 for DBSS data (due in M22) and with D5.3 for accelerometer data (due in M32).

Moreover, regarding confidentiality, ESS data use is in accordance with data protection regulations in participating countries, so only anonymous data are available to users. Before depositing data to NSD (the Norwegian Centre of Research Data), each national team is responsible for checking their data with confidentiality in mind and to undertake the necessary measures to ensure anonymity of the data files and to foresee that anonymity is also maintained after merging of data files.

In SSHOC, Task 5.4 "Remote Secure Access to Data'' is developing infrastructure to enable remote access to sensitive data, that is, data with higher than average disclosure risk, e.g. detailed employment data. Due to disclosure risks, these data cannot be made open. However, the purpose of developing this infrastructure is to fulfil the FAIR goal: open as possible, closed when necessary. Although these data are under controlled access, without this type of infrastructure, the data could not be shared at all, so this work supports the goal ''open as possible''.

[6] Registration procedure on ESS website: https://www.europeansocialsurvey.org/user/new [30.06.2020]

Moreover, the major output of this task is the Specification of an infrastructure that SSHOC could build and sustain to extend this capability for any relevant SSHOC data and data from other EU projects.

2.2.3 Interoperable data

SHARE data is made available to the research community for two of the most common statistical software packages Stata and SPSS. However, users of other statistical software (e.g., R, SAS or Python) can access the data as well because conversion of datasets between statistical software packages is a standard procedure by now. For example, the open software application R supports packages to read-in Stata and SPSS files, which allow SHARE data to be analysed with open software as well. Within all these statistical software applications, SHARE data can be combined with other data sources, e.g. regional data, in any format the software supports.

The standards and vocabularies used allow interdisciplinary interoperability. The SHARE (meta)data is comprehensively documented, based on DDI standards, to enable researchers from all disciplines to use it for their research and combine it with other datasets. Metadata for biomedical data in SHARE will be documented along the regular SHARE data and available in a similar manner in the SHARE Data and Documentation Tool, which strictly follows DDI standards.

The ESS Survey Specification describes the tasks and responsibilities with regard to conducting the European Social Survey. The specification is adapted before every ESS Round. The Specification outlines the tasks and responsibilities of the National Coordinators and the Survey Agency, list all activities, present a timetable and refer to a set of documents that provide guidelines in specific areas (sampling, translation, interviewer briefing, data delivery).

The aim of the Specification is to ensure that high quality data are collected in a comparable way in each participating country. These data should represent both the populations of the participating countries and the concepts under study.

2.2.4 Re-usable data

Re-usability of SHARE data and the SHARE biomedical data, as well as ESS data in general, is not restricted in re-use, except for commercial use. Archiving the data together with a detailed documentation of the (meta)data permits re-use of the data without any time restrictions.

The data is made available to the scientific research community as soon as possible, e.g. if the data is cleaned and checked for data protection/GDPR related issues. Both types of biomedical data need additional efforts in comparison to traditional survey data. Biomedical data must often be extensively processed, validated and calibrated before they can be made accessible and subsequently analysed with the statistical techniques and models that are common in the Social Sciences and Humanities (SSH). While survey data usually is available directly after the data collection (after the necessary subsequent data cleaning and quality checks are carried out), the inclusion of biological samples into the database, e.g., requires an additional step: the analyses of the samples (cf. Schmidutz 2019: 8). Only by means of the analyses the samples are converted into variables, so-called biomarkers (i.e., objective health data), which can be used in statistical analyses for research on e.g.,

health and ageing. An overview of the derived biomarkers is documented in MS21 "Protocol of laboratory processing of DBSS data.

2.3 Allocation of resources

SHARE follows the FAIR principles from its beginning. Additional costs for making biomedical data FAIR are included in Task 5.1 and relate to data processing, data cleaning, releasing, archiving and providing documentation. Database Management team is responsible for SHARE database. It is supported by country team operators who are taking care of national data, also when it comes to the knowledge of the national languages to e.g. code open answers from the job coder software further developed within SSHOC Task 4.3.

2.4 Data Security

Both SHARE and ESS place particular importance on the compliance with European and national data protection law as well as on the safeguarding of sensitive personal data and confidential information.

In accordance with the Statutes of SHARE-ERIC and the Consortium Agreement between SHARE-ERIC and its Scientific Partner Institutions, SHARE-ERIC and its Scientific Partner Institutions act as "joint controllers" pursuant to Article 26 of the European General Data Protection Regulation (EU-GDPR). The SHARE data are distributed to registered users through the SHARE Research Data Centre (FDZ-SHARE) that complies with the Criteria of the German Council for Social and Economic Data (Rat für Sozial- und Wirtschaftsdaten, RatSWD) for providing access to microdata. The released data are provided in anonymised form only. All names and other personal information that could be used to identify individuals or households are removed from these datasets. The same format and protection apply to the accelerometer data, which also do not contain any GPS or other location data. Additionally, for the biomarker data, the physical DBSS are stored in a biobank located at the Institute of Public Health at the University of Southern Denmark in Odense. Neither the biobank nor the processing laboratories have access or connection to the SHARE interview database and do not keep any names or other personal information in their database[7].

ESS ERIC subscribes to the Declaration on Ethics of the International Statistical Institute. All organisations which collect data on behalf of ESS ERIC are similarly asked to do so in accordance with the Declaration as well as any coexisting national provisions. All data collection commissioned directly by ESS ERIC are subject to approval by the ESS ERIC Research Ethics Board. ESS ERIC activities are implemented in line with applicable international, EU and national law. From May 25th 2018, this includes compliance with the General Data Protection Regulation (GDPR). ESS ERIC has registered as a data controller with the UK Information Commissioner. All organisations involved in processing personal data on behalf of ESS ERIC will be required to sign a data processor agreement with ESS ERIC confirming that all personal data will be handled in compliance with the GDPR. Sub-processors can only be engaged with the prior written consent of the Data Controller and subject to a written contract.

[7] SHARE Methodology Volumes and Compliance Profiles can be accessed via the SHARE homepage: http://www.share-project.org/data-documentation/methodology-volumes.html [30.06.2020]

ESS ERIC's data processor agreement is currently being updated to ensure compliance with GDPR available from May 2018.

2.5 Ethical aspects

Three beneficiaries in the SSHOC project, ESS, SHARE and GGP, have marked some of the issues in the ethics checklist as relevant in this project. They are related to section 2. Humans, since the three infrastructures in their surveys collect data from human participants (volunteers). Under section 3. Human cells/tissues, SHARE ERIC's data is related to secondary use of data from human cells or tissues, and some are obtained/stored in biobank. The three infrastructures in their work use personal data (section 4. Personal data) collection and/or processing, including processing of sensitive personal data, and further processing of previously collected personal data (secondary use). As a general rule, templates of the informed consent/assent forms and information sheets (in language and terms intelligible to the participants) will be kept on file, along with the copies of opinions/approvals by ethics committees and/or competent authorities for the research with humans must be kept on file. Under section 6. Third countries, SHARE ERIC, plans to import deidentified survey data from Israel and Switzerland. ESS ERIC will export deidentified survey data collected in EU countries to the ESS data archive in Norway (NSD). Detailed explanations for the three beneficiaries are provided below.

ESS ERIC Ethics information for WP4

ESS ERIC will commission data collection with human subjects under T4.1 "A sample management system for cross-national web surveys". In order to test the operation of the sample management system being developed under T4.1 a small sample of adults aged 18+ (c. 50) in Austria and the United Kingdom will be invited to join a two-wave panel and complete up to two short 10 minute surveys on general, non-sensitive, themes. Recruitment is done off the back of the ESS Round 10 pilot (which involves a representative sample of the general adult population aged 18+) with pilot respondents asked at the end of the face to face interview if they would be willing to join the panel. Participation will be entirely voluntary. According to the Post-Grant Requirements, it will be verified if a declaration on compliance and/or authorisation is required under national law for collecting and processing personal data as described in the proposal. If yes, the declaration on compliance and/or authorisation will be kept on file. If no declaration on compliance or authorisation is required under the applicable national law, a statement from the designated Data Protection Officer that all personal data collection and processing will be carried out according to EU and national legislation will be kept on file. Informed consent will be sought with respondents given a participant information sheet at the time of the invitation. The information sheet will contain the information required by compliant with the requirements of GDPR. Panel members will also be able to easily withdraw their consent and fully exercise their data access rights. Detailed information on the informed consent procedures in regard to the collection, storage, and protection of personal data will be kept on file. Templates of the informed consent forms and information sheets (in language and terms intelligible to the participants) will be kept on file. In case of further processing of previously collected personal data, relevant authorisations will also be kept on file. The test of the sample management system will involve the processing of personal data, including panel members contact details (name and email address for survey invites and mobile telephone number for SMS reminders).

Data protection is a primary consideration behind the design of the sample management system which is intended to provide a means of co-ordinating cross-national panel management between national and central

teams efficiently and securely. The system will be designed from the very start with data minimization in mind, as well as access control to personal data on a strict distinction between the various roles and institutions taking part in the project. Industry-standard safeguards for data storage or transfer including encryption will be employed.

Finally, detailed information on the procedures for data collection, storage, protection, retention, and destruction, and confirmation that they comply with national and EU legislation will be kept on file. Survey data collected during the test of the sample management system will be transferred outside of the EU (passed on to the ESS Data Archive at NSD - the Norwegian Centre of Research Data, Norway) at the end of the project. According to the Post-Grant Requirements, in case personal data are transferred from/to a non-EU country or international organisation, confirmation that this complies with national and EU legislation, together with the necessary authorisations, will be kept on file. Only de-identified data will be transferred and respondent contact details and any key linking respondent ids to personal information will be destroyed at the end of the project. Norway will implement the GDPR in 2018. The data will be treated in the same way as in all EU Member States. ESS ERIC subscribes to the Declaration on Ethics of the International Statistical Institute. All organisations which collect data on behalf of ESS ERIC are similarly asked to do so in accordance with the Declaration as well as any coexisting national provisions. All data collection commissioned directly by ESS ERIC are subject to approval by the ESS ERIC Research Ethics Board. ESS ERIC activities are implemented in line with applicable international, EU and national law. From May 25th, 2018, this will include compliance with the General Data Protection Regulation (GDPR). ESS ERIC has registered as a data controller with the UK Information Commissioner. All organisations involved in processing personal data on behalf of ESS ERIC will be required to sign a data processor agreement with ESS ERIC confirming that all personal data will be handled in compliance with the GDPR. Sub-processors can only be engaged with the prior written consent of the Data Controller and subject to a written contract. ESS ERIC's data processor agreement is currently being updated to ensure compliance with GDPR but will be available from May 2018.

For Task 4.3 "Applying Computer Assisted Translation tools in Social Surveys", participants of the machine translation pilot study will be informed on the data use and asked for their consent prior to taking part in the study.

SHARE ERIC Ethics information for WP3 and WP5

SHARE ERIC will commission data collection with human subjects under T3.2 "Essential SSH ontologies and vocabularies". The occupation ontology developed in this task will be incorporated in the SHARE survey and data will be collected in all SHARE countries, by means of the job coder technology. This SHARE data collection requires collection and processing of personal data. As previously stated, SSHOC project will comply and keep on file detailed information on the informed consent procedures in regard to the collection, storage, and protection of personal data. Templates of the informed consent forms and information sheets (in language and terms intelligible to the participants) will also be kept on file. In addition, SHARE ERIC will process biomarker data (secondary use) obtained from dried blood spot samples (DBSS) which were collected in SHARE wave 6 in 2015 and used under T5.1 "Legal, ethical and technical aspects of access to biomedical data". The data collected in both T3.2 and T5.1 will be linked with all previous waves stored in the SHARE database. As this is the case of further processing of previously collected personal data, relevant authorisations will also be kept on file. The following subsection describes SHARE's general compliance with data protection laws and ethics

approval of the project, and the subsections thereafter the specific details of the project's efforts to meet all national and international legal requirements and address all ethical issues in an appropriate manner.

SHARE's compliance with data protection laws and approval by ethics review boards

All work will be carried out in compliance with European Union and national data protection laws. The collection and processing of data of previous waves of SHARE (waves 1 to 7) have been carried out in compliance with the directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data and the national implementations of this non-self-executing legal instrument in the EU Member States and all countries in which the SHARE data are collected. All data collection and processing as of 25 May 2018 is a subject to and carried out in compliance with Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (which has entered into force on 24 May 2016 and is directly applied from 25 May 2018 in all EU Member States). The SHARE project and its data collection procedures are reviewed on a regular basis prior to each new wave of data collection. The SHARE data collection procedures have been reviewed and approved by the Ethics Council of the Max Planck Society (MPG) in June 2012, March 2014 and February 2016. SHARE-ERIC's activities related to human subjects' research are guided by international research ethics principles such as the Respect Code of Practice for Socio-Economic Research and the Declaration of Helsinki (as prepared by the Council for International Organizations of Medical Sciences in collaboration with the World Health Organization, last revised at the 64th WMA Meeting held in Fortalezza/Brazil in October 2013). Moreover, the Max-Planck-Institute for Social Law and Social Policy assured by means of its Federal wide Assurance (FWA) for the Protection of Human Subjects to the Office for Human Research Protections (OHRP) of the U.S. Department of Health and Human Services HHS that all of its activities related to human subjects research will be guided by the Declaration of Helsinki' and the International Ethical Guidelines for Biomedical Research involving Human Subjects'. In addition, the country implementations of SHARE – in particular with regard to the international collection of Dried Blood Spots Samples (DBSS) in SHARE Wave 6 – have been reviewed between February and May 2014 and approved by the respective ethics committees or institutional review boards (IRBs). The implementation of the collection of accelerometer data has been approved along the regular SHARE approval by the ethics review board for wave 8.

SHARE´s sampling and informed consent procedures

All respondents in SHARE are volunteers and the entire data collection is based on informed consent. Before starting the interview of each wave, each respondent's consent is obtained with regard to his/her participation. Additionally, during the interview answers to all questions are voluntary. Furthermore, respondents consent to the storage and the tracking of their addresses as long as they are participants of the SHARE panel and may revoke their consent at any time. The samples are drawn randomly from national, regional, or telephone-generated registers in order to create a representative probability sample. All registered persons in private households have the same ex-ante probabilities to be drawn. The procedures to ensure the protection of personal data in SHARE have been approved by the Data protection Officer of the Max-Planck Society who is a personal union Data protection officer of SHARE ERIC.

SHARE´s secondary use of human blood data

The collection of dried blood spot samples (DBSS) in SHARE wave 6 took place in 2015 and has been set up in strict compliance with appropriate European and national data protection and ethical standards (see section "SHARE's compliance with data protection laws and approval by ethics review boards". and section "Sampling and informed consent procedures"). SSHOC project includes only secondary processing of previously collected DBSS, no (new) dried blood spot samples are collected in this project. Copies of relevant documents for using, producing or collecting human cells or tissues (e.g. ethics approval, import licence, accreditation/designation/authorisation/licensing) are kept on file.

GGP Ethics information for WP4

The GGP[8] (Generations & Gender Programme) component, the Generations and Gender Survey (GGS) will be used for testing tools developed in Task 4.4 and Task 4.5. Participants in the survey consent to the processing of their data for research purposes and are informed about the data processing, storage and archiving procedures. The data provided is depersonalized and access to the data is provided only to researchers who have signed a data agreement on the condition that it is used for research purposes only. For the audio data collection in Task 4.4, respondents will be required to consent to the recording and the consent procedures of the Generations and Gender Survey will be adapted accordingly. Both Task 4.4 and 4.5 will be subject to an ethics review in the first year of the project which will be conducted by the Sociology Department at the University of Groningen with whom the GGP is affiliated. Both tasks will also be evaluated by the data protection officer of the Royal Dutch Academy of Science (KNAW) and data protection officers within any country in which the fieldwork is conducted. For the purposes of this work, data will only be collected in European Member States and will be wholly processed and held within the European Union. All requirements regarding Ethics stated in Post-Grant evaluation will be taken into account and complied with.

[8] Generations & Gender Programme website: https://www.ggp-i.org/about/ [30.06.2020]

3. Case studies / pilots data

3.1 Data Summary

Implementation of the SSHOC project anticipates several pilots and case-studies with the purpose of implementing enhancements and upgrades in the existing methodologies and technologies.

3.1.1 Case-study for piloting NLP technologies for translation

The task 3.1. in SSHOC project will collect multilingual metadata, multilingual vocabularies, multilingual terminologies, (e.g. occupation), currently used and relevant for the research infrastructures involved in SSHOC and managed by the SSHOC main stakeholders. These will constitute the background resources to support the development of services accessible and usable in SSH in order to:

- to facilitate knowledge discovery and make content searchable across different languages
- to permit metadata based discovery using different languages
- to improve discovery by non-native speakers and maximise the accessibility of such archives.

After the collection process, the collected vocabularies will be enriched by extracting terminologies with NLP extraction techniques from relevant documents and will be translated using MT techniques.

Depending on the type of data, different data formats will be collected by the project (harmonization and interoperability of vocabularies will be dealt with in other tasks). The format of the multilingual vocabularies will depend on the SSHOC recommended vocabulary publication platform decided in the project.

As far as the vocabularies are concerned, metadata catalogues of the main infrastructures involved in the project will be reused:

- The CESSDA Data Catalogue
- ESS Data Nestar
- The SHARE Survey of Health Aging and Retirement in Europe
- The CMDI CLARIN Component Metadata
- The DARIAH collection registry
- The ISIDORE platform of SSH digital data
- The Parthenos Joint Resource Registry

In the short term, the first data providers are the research infrastructures involved in SSHOC. In the long term, the origin of the data could become any relevant SSH repository in the European research area.

The collected multilingual metadata and taxonomies will be useful for the SSHOC users to maximize the accessibility and to improve discovery by non-native speakers. The data will be useful firstly to the project consortium. In the future the data will be used by researchers interested in SSH studies.

3.1.2 Case-studies for piloting NLP technologies for text analysis

The main objective of Task 3.3. is to bring to practice and to adopt where needed the developed natural language processing solutions into the context of social sciences and humanities.

Task 3.3. will do a number of case-studies piloting NLP technologies for text analysis. Currently the datasets have not yet been determined in detail, but the originators of those datasets are assumed to adhere to all requirements wrt GDPR. Since the datasets are already existing and are meant for testing the NLP technologies, no new responsibilities with respect to their data management is necessary. In order to provide demonstration of the algorithms and pipelines a potential test set has been created which consists of drama corpora from Baroque era in English (Shakespeare), French (Moliere et aalii; complete catalogue of authors and pieces is available, if wished), and Spanish works from Calderon, Cervantes, Tirso de Molina and many more, again, if wished, a list of writers exists.

By using these dramas for respective NLP tools, a substantial contribution shall be made answering following, subject-specific research question: In European drama (mostly English, Spanish, French and Italian) of the 16th/17th century the character of the servant or other characters of the lower social class have the function to unveil the hidden order of discourse by incorporating a comical wisdom which is only possible to be represented by the minor characters. The period is also characterized as the most productive time in the history of theatre. That's why it seems impossible to discover all these characters by reading (take alone Shakespeare, Lope de Vega or the French théâtre classique) – to make these characters discoverable in a large corpus of plays (like DraCor, + more) would make it possible to see the role of these minor characters more clearly and to analyse the movement and intertextuality in European drama history. Necessary NLP: who is speaking, who is on stage (network), what is this group of characters usually contributing to the speech? The research-driven use of the NLP tools serves both to clarify their potential and to identify the challenges that still exist when using them.

Data used are both plain text and XML (TEI P5) encoded files plus some self-created project-specific Python and XSLT scripts for transforming plain text files to XML and v.v XML encoded and TEI valid files shall be published on Drama Corpora Project[9] afterwards for further scientific, public use. Files are taken from Drama Corpora Project or are already in the public domain.

First, after all dramas have been encoded respectively, and published on Drama Corpora Project, they can be used by literary scholars for different languages as well as literary comparatists. Second, using respective NLP tools helps to further develop and exploit their potential for different languages and language levels, likewise. In addition, one of the case studies focuses on the analysis of verbal aggression against specified targets based on Tweets. The case study foresees two axes for data collection: domain and language. Focused Twitter collections will be created using specific keywords, depending on the domain (e.g. xenophobia or covid19) and the languages (Greek or English), with the support of dedicated NLP technologies.

[9] Drama Corpora Project website: https://dracor.org/ [30.06.2020]

3.1.3 A sample management system for cross-national web surveys

Task 4.1 team launched a web pilot survey in Austria and UK with panellists recruited off the back of the ESS Round 10 pilot. The first wave includes questions on climate change. The web pilot will work as a platform to test the sample management system that is currently under development.

When the test of the sample management system will be completed all the panellists' personal data will be deleted. Survey data will be stored at ESS ERIC (City) for further research. To set up and maintain the web pilot test study panellists contact details have been collected. With the first wave of the test study partners will collect survey data on climate change. The data will be collected through CAWI.

3.1.4 Applying Computer Assisted Translation tools in Social Surveys

Task 4.3 team will conduct a pilot study on the use machine translation (MT) in questionnaire translation using already existing survey questions from ESS and European Values Study[10] (EVS). Team will generate translation-related data, for instance, review (final) and interim translation versions, machine translation output and post-edited version and background information on the translators (participants) to the pilot study. In the process past translations and source questionnaires from ESS and EVS will be used.

3.1.5 Voice recorded interviews and audio analysis

Task 4.4 data collected in the Audio Capture experiment will be collected and processed by NIDI-KNAW through the Generations and Gender Programme. As with the standard GGP consent protocols, respondents are provided with information on the data processing involved within the Generations and Gender Programme and asked to provide informed, active consent for this processing. In the data collection including the audio capture of data, this informed consent explicitly mentions the audio processing elements and that all data will be depersonalized before analysis and only used for statistical processing. The audio capture data is highly sensitive and potentially could lead to the re-identification of respondents and will therefore not be made publicly available. Summary statistics of the audio data, as described in the work description, will be made available alongside the standard GGP survey data file.

3.1.6 Social policy APIs for social surveys

In Task 4.5 of the SSHOC project, the Social Policy Indicator Algorithm is an extension of the existing survey instrument used by the Generations and Gender Programme and will be documented and archived as part of the standard data management procedures of the GGP. The data is not sensitive and integrates the algorithm inherent in existing, publicly available social policy datasets such as the OECD family policy calculator [11]into the Generations and Gender Survey workflow.

[10] European Values Study website: https://europeanvaluesstudy.eu/ [30.06.2020]

[11] OECD Family database: the Family support calculator https://www.oecd.org/els/soc/oecdfamilydatabasethefamilysupportcalculator.htm [30.06.2020]

3.1.7 LOD archaeology case study

Task 5.7 team should create a virtual reconstruction of the Roman theatre in Catania as an example of an actual transition of archaeological data to the cloud. Data will be generated for all stages of the workflow from an archaeological excavation/survey to a 3D reconstruction of the site. It may also generate spatial and temporal norm data and controlled vocabularies. Existing CRM-IBAM survey data and 3D reconstruction data from the Roman theatre in Catania and norm data from spatial and temporal gazetteers will be re-used. Data originates from surveys of the Roman theatre in Catania, and 3D data is produced with the Extended Matrix tools.

3.2 FAIR Data

3.2.1 Findable Data

ESS as a service establishes a pilot for making cross-national survey data FAIR. The ESS data will be freely available and findable in a new cloud-based data repository by use of both different search options and DOI. The aim is to use a within site system allowing for both free text search and filters using a controlled vocabulary. In developing the system, the version numbers (version 0.1, etc.) will be used. In the data files and documentation provided, through the new cloud stored repository and website aim is to use a 4-level version number system. DDI CDI and DDI Lifecycle will be used. Metadata on case level will mostly be either integrated with the survey data (routing information, pre questions and interviewer notes) or published in datasets that will be possible to merge with the main data. This includes data on interviewer, interview situation, timing/length of interview, contact data, data on respondent dwelling situation. It also includes information on sampling and inclusion probability. Data on survey level (both country specific and overall survey specific information) will be included through accompanying structured documentation (Data documentation report, quality reports, source and fieldwork documents and so on).

3.2.2 Openly accessible data

The screening process for repositories for making the translation memories and the translation-related data from the machine translation pilot study openly accessible is currently active. The pilot study on machine translation is conducted using the openly available translation software MateCat [12].

The principle of as open as possible, as closed as necessary will be followed. Files that are controlled for anonymity and can be as anonymous will be openly available for all. Files that are considered as indirectly identifiable will be available upon applications through special license agreed upon with the ESS ERIC DPO. No data will be kept closed. All data will be accessible in a number of different formats (SAS, SPSS, STATA and so on). In addition, aim is to provide a simple online analysis option for those who do not have access to statistical packages. Data can also be downloaded in excel format; however, this is only advisable for smaller subsets of the data. Survey information/documentation will be accessible. No specific software will be provided, except the online analysis option – a guide/tutorial/webinar on how to use this will be developed and be available alongside the tool. For using the translation memory, users will need a Computer-Aided Translation (CAT) tool,

[12] MateCat tool website: https://www.matecat.com/ [29.06.2020]

such as MateCat, that reads TMX format. Software and APIs for online analysis developed by NSD will be Open Source. A cloud repository is being built and the process of applying for Core Trust Seal certification for the ESS data archive is on the run. Data that has restricted access will be available through application and on certain conditions (yet to be agreed upon with the ESS ERIC DPO).

3.2.3 Interoperable data

The translation memory is interoperable as it is designed in standard translation memory format i.e. TMX. CLARIN standards for preserving corpus will be used.

Project aims to store and document existing ESS data to enable interoperability. Use of the DDI Lifecycle and particularly DDI-CDI metadata standards for documentation will support these aims. DDI Lifecycle and particularly DDI-CDI and Open Source APIs for access to data.

3.2.4 Re-usable data

The data of the machine translation pilot study will be made available at the end of the SSHOC period. The translation memories will be made available alongside the deliverable on its use (2021). Most data will be openly available with no license necessary, only registration by use of Orcid[13] or other identification systems. Where licenses are necessary, a simple application form will be used. There is no embargo on ESS data.

3.3 Allocation of resources

The responsibility is held by all consortium members.

3.4 Data Security

All the panellists' contact details are stored at ESS HQ in a protected drive. Besides the task leader no one has access to this information. The panellists' contact details are also stored in the sample management system which is located at Sciences Po (partner of the task). Some pieces of information (like Name, Surname, email addresses) are shared with the survey platform provider Qualtrics[14] for survey operations. The survey data are stored at Qualtrics. At the end of the web pilot study survey data will be downloaded and stored also at ESS HQ for further research useful to develop the sample management system for cross-national surveys. All personal data will be handled in compliance with the GDPR. At the end of the test, all the data will be securely deleted. A certified institution, NSD, hosts the ESS Archive. Documentation as well as data are stored in generic formats.

[13] ORCID website: https://orcid.org/ [30.06.2020]

[14] Qualtrics Experience Management Software website: https://www.qualtrics.com [29.06.2020]

4. Tools and services data

4.1 Data Summary

Development of the tools and services is one of the main aims of the SSHOC project. Data generated, collected and/or re-used for that purpose is presented in this section. Current overview of the SSHOC Tools and Service data per project tasks is listed below and will be updated regularly.

Selected Ontologies and Vocabularies (T3.2)

Data generated from this task serve socio-economic surveys for their long-list questions about respondent´s occupation, religion (In progress), education and industry. Vocabulary and data are available for web surveys and CAPI surveys via Survey Codings[15], a specialised repository that does not need advanced discovery mechanisms. Vocabulary sources are included in an explanatory note on the page. The data are useful for any multi-country survey asking for long-list questions about respondent´s occupation, religion, education and industry.

Making Data Findable by being Citable, task 3.5 Data and Metadata Interoperability Hub and Task 3.6 Making Data Re-usable and actionable (T3.4, T3.5)

All service and tool development activities are required to make results available on GitHub[16] under open software license. Production level services and tools will be made to comply with the GDPR regarding managing user credentials and attributes. With respect to task 3.5 the "interoperability Hub", Drupal based content management application to collect and curate the information on conversion services and tools will be used. The data collected will be backed up regularly by the host ACDH-CH/OEAW[17] who operates the installation. Where the data contains personal information, these will be names of authors and operators, as gathered in publicly available sources.

Preparing tools for the use of Computer Assisted Translation (T4.2)

Task team will produce a multilingual corpus of survey questionnaires, namely the Multilingual Corpus of Survey Questionnaires (MCSQ) using already existing survey questions from ESS, EVS (version Ada Lovelace) and SHARE in a later iteration. Then part of this data will be used in task 4.3: Applying Computer Assisted Translation tools in Social Surveys, to test it as a translation memory in the context of translating survey questionnaires.

The data is collected by transforming PDFs questionnaires into plain text and then building a database (ER model). Output formats are SQL compatible: TMX, CSV and XML. ESS, EVS and SHARE questionnaires will be

[15] Survey Codings repository website: https://www.surveycodings.org/ [29.06.2020]
[16] Github repository: https://github.com/ [30.06.2020]
[17] Austrian Academy of Sciences: https://www.oeaw.ac.at/ [29.06.2020]

re-used. The data are useful to corpus linguists, translation scholars and practitioners, cross-cultural survey methodologists

Semantic annotation of Heritage Science Data (T4.6)

The main purpose of this task is to provide semantic annotations enriched by multiple users and offer a tool for the description of 3D models for various studies in SSH. As a collaborative platform, users can share their observations on the data in an interdisciplinary approach.

The types of data: Digital Asset generate/collect; Photogrammetry Objects; Annotations (user descriptors, text, number, attachments, reports, etc.); Images; Dense 3D point clouds; Geometric descriptors (normal, curvature, ambient occlusions).

The formats of data: .ply for Point Cloud; .json for computed description; .csv for user descriptors.

The data are useful to:

- Actors involved in the conservation-restoration of cultural heritage objects: Curators; Restorers.
- Actors involved in the production of scientific knowledge on cultural heritage objects: Researchers; Researchers in cultural heritage studies; Historians; Historians of art; Archaeologists; Architects; Engineers; Physicists; Chemists.

Hosting and sharing data repositories (T5.2)

Purpose of task 5.2 is the development of a research data repository service on EOSC, for SSH institutes currently without such a facility for their designated communities. Task 5.2 doesn't collect data itself. Source code of the project is available at: https://bitbucket.org/account/user/cessda/projects/DVS and https:/SSH institutes currently without a repository facility for their designated communities.

Issues in providing Open Data in Heritage Science and Archaeology (T5.6)

The main purpose of this task is to document the process of re-formatting two existing data sets to demonstrate how they can be presented as examples of FAIR Heritage Science data sets, as a model for future work. The task is not intended to create new raw data but will create/demonstrate the modelling procedure and a structured vocabulary of the terms used to connect the data sets to external vocabularies.

The data sets may well be presented in several forms, subject to the development of the task, but as a minimum requirement they will be presented via human readable graphical user interfaces, as individually resolvable URIs and as machine actionable structured XML/RDF documents. In addition to the actual datasets some of the modelling processes and examples will also be presented in an interactive and re-usable manner (https://github.com/jpadfield/cidoc-crm.examples).

The two FAIR data sets will be created from two existing, less structured data sets currently accessible via existing web based GUIs: https://research.ng-london.org.uk/iperion/ (currently password protected) and https://cima.ng-london.org.uk/documentation/.

The data produced will be useful for any Heritage Scientists, or Art Historian or Conservators specifically studying the areas covered, they will also be of use to Digital Humanities examining how this type of data might be presented in the future along with any technology researchers examine how the data has been mapped and modelled. The created FAIR datasets will also be directly re-used to test the development of Heritage Science documentation software as part of the work within the task.

Tools used for text processing

1. Machine Translation service frontend for translation between some European languages

   PID: http://hdl.handle.net/11234/1-2922

   Conditions for use: BSD 2-Clause "Simplified" or "FreeBSD" license (http://opensource.org/licenses/BSD-2-Clause)

   Translation service application URI: https://lindat.mff.cuni.cz/services/translation

   Documentation entry point (incl. API description): https://lindat.mff.cuni.cz/services/translation/docs

2. UDPipe multilingual analysis tool

   PID: http://hdl.handle.net/11234/1-1702

   Conditions for use: Mozilla Public License 2.0 license (http://opensource.org/licenses/MPL-2.0)

   UDPipe service application URI: http://lindat.mff.cuni.cz/services/udpipe/

   Documentation entry point (incl. API description): http://ufal.mff.cuni.cz/udpipe.

4.2 FAIR Data

4.2.1 Findable Data

Preparing tools for the use of Computer Assisted Translation (T4.2)

CLARIN repository will be applied to store the MCSQ permanently, and their standards for FAIR data will be followed as they are already validated. Version numbers and names are provided in generating the data, the first one is Version 1 Ada Lovelace and the next one will be Mileva Marić-Einstein. In the data files and documentation provided, common versioning standards are used.

Semantic annotation of Heritage Science Data (T4.6)

The Aïoli platform[18] uses a no-sql document-oriented database which is CouchDB. It allows access to data through standard identification (Hash Standard). Random number of ten characters (a Hash) taken in

[18] AIOLI website: http://www.aioli.cloud/ [30.06.2020]

hexadecimal space (0 to 9 and a to f). It allows a very low collision rate. The Aïoli platform doesn't create metadata but exploits EXIF metadata.

Issues in providing Open Data in Heritage Science and Archaeology (T5.6)

A specific naming convention will be used for the two main datasets, but a variety of naming conventions will be used within the data sets. These will primarily be related to the production of file names and object identifiers used with the systems. National Gallery, published, persistent identifiers will also be used to increase access. The data sets both include a wide variety of text-based data, which can be searched directly, but all of the terms used to tag and connect the data together will also be gathered into a structured vocabulary, which can then be used to facilitate re-use. Metadata for the complete datasets will be relatively limited, however the knowledge stored within the datasets will be fully mapped to public standard ontologies (such as CIDOC CRM[19]), and linked to external public vocabularies (such as Getty Vocabularies as Linked Open Data[20]) to increase the potential for them to be connected to other similar

Tools used for text processing

Both tools are FAIR in the sense that they fulfil the FAIR requirements, except the UDPipe models and the MT translation models (as described at the UDPipe/MT landing pages in the repository, which are reachable directly through the PIDs) are not free for commercial use.

4.2.2 Openly accessible data

Preparing tools for the use of Computer Assisted Translation (T4.2)

CLARIN repository will be applied to make the data permanently accessible. The code will be openly accessible in GitHub repositories. During SSHOC, it is stored in a virtual machine provided by UPF.

Semantic annotation of Heritage Science Data (T4.6)

A user has the possibility to publish his data on the Aïoli platform or to keep them confidential. Users must connect to the Aïoli platform. Make an https query to retrieve the raw data from the database. Retrieve HDD data from the server.

Issues in providing Open Data in Heritage Science and Archaeology (T5.6)

All of the text-based data stored within the current systems will all be made available through the work in SSHOC, via an appropriate re-use licence, as indicated in the Consortium Agreement. Access will be limited via defined re-use licences, such as CC NC-ND-By 4.0[21]. Specific licences will be attached to the data set presentations.

[19]CIDOC CRM website: http://www.cidoc-crm.org/ [30.06.2020]
[20] Getty Vocabularies as Linked Open Data: https://www.getty.edu/research/tools/vocabularies/lod/ [30.06.2020]
[21] Creative Commons website: https://creativecommons.org/licenses/by-nc-nd/4.0/ [30.06.2020]

The data has been provided by a number of different sources so some limitations will be in place for a small subset of the data. The images associated with the text data are not specifically being re-published within SSHOC, as they are already available, but where possible they will still be directly linked to the text data, along with all of the relevant meta-data. The data will be available as a simple download, via an interactive human accessible GUI and via a machine actionable API. The text data will all be ascii text and open-able within a standard text editor. Any linked images will be served via a standard IIIF image server[22] and can be viewed via any compliant IIIF viewer. The primary presentation of the data will be presented on one of the National Galleries public research servers. Additional registration of the data within external repositories will be explored as part of the work of the project.

Tools used for text processing

The repository where these services are stored is LINDAT/CLARIAH-CZ[23], to which the above PIDs are resolved. The repository is CTS certified, and also certified as CLARIN B-type repository. It is also listed in OpenAire. Thus, it is eligible to be used for any resources created by H2020 projects now or in the future. Both the services are also run on (several) servers[24] 24/7 and are accessible from the application pages listed above, or as an API accessible remotely from other scripts or programs, as described in the documentation.

4.2.3 Interoperable data

Preparing tools for the use of Computer Assisted Translation (T4.2)

The MCSQ is interoperable. It is implemented using standards for databases design and implementation.

Semantic annotation of Heritage Science Data (T4.6)

Interoperable standards are used such as JSON, ISO/IEC 21778:2017 (the JSON data interchange syntax), XML and HTTP for data retrieval. The description of the heritage asset within the Aïoli platform is compatible with the Dublin Core standard. The controlled vocabularies in SKOS format are integrated into the Aïoli platform via the Opentheso software (e.g. Art & Architecture Thesaurus (Getty), Thesaurus of Geographic Names (Getty), Iconclass, thesaurus of the French Ministry of Culture). The Task 4.6 focuses on the definition of a mapping mechanism aimed at ensuring the full traceability of the multi-users 3D-annotation process within the CIDOC CRM, as well as within the higher-level ontology introduced by the SSHOC project.

Issues in providing Open Data in Heritage Science and Archaeology (T5.6)

The datasets will be fully mapped to public standard ontologies[25], and linked to external public vocabularies[26] to increase the potential for them to be connected to other similar data sets in the future and to facilitate more complex searching within the data sets.

[22] International Image Interoperability Framework: https://iiif.io/ [30.06.2020]
[23] LINDAT/CLARIAH-CZ repositories: https://lindat.cz, or https://lindat.mff.cuni.cz [30.06.2020]
[24]LINDAT CLARIN CZ website: lindat.mff.cuni.cz [20.06.2020]
[25] such as CIDOC Conceptual Reference Model (CRM): http://www.cidoc-crm.org/ [30.06.2020]
[26] such as Getty Vocabularies as Linked Open Data: https://www.getty.edu/research/tools/vocabularies/lod/ [30.06.2020]

4.2.4 Re-usable data

Preparing tools for the use of Computer Assisted Translation (T4.2)

An academic article about the MCSQ is already accepted, the embargo of the data will last until Q2, 2021.

Semantic annotation of Heritage Science Data (T4.6)

Https authentication is required to access the data. UNIX domain authentication is required to access the servers and there are connection filters. The data life cycle is long: deletion of data is at the user's initiative (there is no automatic deletion).

Issues in providing Open Data in Heritage Science and Archaeology (T5.6)

The processes and modeling created within the project will be licenced under an open licence such as GPL V3, however the actual content of the data sets being produced will be limited by existing licensing restrictions as noted before. A small amount of data related to both data sets will remain restricted as an open licence has not been granted by the original owner. The raw data for one of the data sets is already available, though not formatted correctly, the data for the other data set will be reduced once it has been formatted.

4.3 Allocation of resources

Semantic annotation of Heritage Science Data (T4.6)

Within the framework of the Aïoli platform, the SSHOC project allows to align the development of the platform to the CIDOC-CRM standard, which will allow, in the long run, to make the data produced by the Aïoli platform FAIR. The goal is to introduce a mechanism to make all the public scenes of the Aïoli platform directly FAIR. The SSHOC's funding is oriented in this direction.

The Aïoli project is linked to the activities of the 3D SHS consortium of the TGIR Huma-Num, in charge of defining guidelines for the perennial conservation of 3D representations. For the moment, the repository of a 3D scene annotated by the Aïoli platform for long-term archiving is not yet defined in detailed technical specifications (in particular due to the lack of a standard for the description of 3D annotations). However, the results of a 3D digitization produced by the Aïoli platform (3D photogrammetric correlation in the form of a 3D point cloud and a set of oriented photographic images) can already be deposited in the 3D conservatory (3D SHS consortium Huma-Num) for permanent archiving via CINES. Concerning the costs, there is no information available because, for the moment, they are covered by the national action carried out by the CNRS via the TGIR Huma-Num.

Issues in providing Open Data in Heritage Science and Archaeology (T5.6)

The costs for making the data FAIR relates to the time required to model the data, this time is covered by the SSHOC project. The long-term presentation of the datasets is being absorbed into the ongoing costs of the

National Gallery. Additional presentations of models and mappings will be presented via the free service on GitHub. The management of the core data sets will be covered by the National Gallery.

Tools used for text processing

There was no cost to SSHOC associated with making them FAIR or Open Source, since their licences have been selected as Open Source / FAIR before the SSHOC project started.

4.4 Data security

Preparing tools for the use of Computer Assisted Translation (T4.2)

Partners participating in the task implementation have already data security procedures operational in their institutions (e.g. back-ups, access through a VPN). They adhere to the GDPR. Sensitive data is not collected within this task and for the machine translation pilot study, contact details of participants will be stored separately from the research data.

Semantic annotation of Heritage Science Data (T4.6)

Data is protected from eavesdropping attacks: https encrypted/secure channel. To access the Aïoli platform, an account with a username and password is required. The implementation of restriction policies at the server level is that the connection to the Aïoli platform requires VPN. Within the laboratory, a replication system is set up with replication servers.

Issues in providing Open Data in Heritage Science and Archaeology (T5.6)

There is no sensitive data included in the data set, security of the data will be insured via multiple back-ups and by incorporating the data into the National Galleries on-going off-site data backup procedure.

The data will be stored by the National Gallery, but additional external data repositories will be explored as part of related research activities external to the project.

5. SSHOC Marketplace data

5.1 Data Summary

5.1.1 Marketplace Metadata

The SSH Open Marketplace primary data is metadata. As an aggregator, the Marketplace collects metadata from third party catalogues, which are publicly available sources. These metadata undergo a curation and enrichment process and are subsequently exposed to the public via an online catalogue and an API. The aim and the context of this enrichment is described in detail in Deliverable D7.1. This document also contains a description of the format that will be used to store the curated metadata.

Since this harvesting and curation process is still ongoing, it is hard to provide exact numbers about the number of resulting metadata descriptions, but current estimations range between a few thousands and a few millions.

5.1.2 User requirements data

In 2019, user stories were created to inform the requirements and specification process of the SSH Open Marketplace (cf. Annex 3 of the D7.1 "System Specification of the SSH Open Marketplace"). To create these user stories, 22 interviews were conducted with SSH researchers, targeted as the main end-user group of the (future) website. Some of these interviews were recorded to facilitate the production of interviews' summaries. User stories were extracted from the interviews' summaries. To allow the collection and processing of personnel data following GDPR regulation, a consent form needed to be approved by the interviewees. After some explanation about the interview and how their data would be collected and stored, an explicit question to allow WP7 members to process personal data - recordings and summary of the interviews - collected during these interviews was added to the consent form. Based on a discussion with the coordinator of the SSHOC project, it was decided that these materials should be stored in a DARIAH repository and DARIAH was indicated as data controller to the interviewees.

5.1.3 User account data

The Marketplace will also store some minimal data about users that choose to login. This user account data is specified in more detail in Deliverable D7.1, table 21.

5.2 FAIR Data

5.2.1 Findable Data

The Marketplace is mainly a portal intended to be used by humans and machines to find tools, services, datasets and training materials. Thus, discoverability and findability are one of the main goals of the Marketplace. Actually, by building links between metadata collected from various sources and the creation of

normalized content supported by community-driven curation, the Marketplace provides a common access to a set of different resources. Every entry from the Marketplace will be assigned an identifier.

Local identifiers (e.g. https://marketplace.sshoc.eu/tools/43) and if available Wikidata identifiers (e.g. https://www.wikidata.org/wiki/Q28405731) will be used and in the future DOIs if applicable. If other relevant identifiers exist, they will be integrated in the metadata.

About findability, the Marketplace provides a search engine for humans and a classical REST API for machines. Finally, findability will be enhanced by the integration of the citation mechanism specified and developed in task 3.4 from WP3.

As the Marketplace doesn't host data, naming conventions are not entirely relevant. However, Task 7.4 is elaborating a guide dedicated to the curation team in which all conventions are detailed. As the Marketplace is mostly an aggregator for existing material coming from different sources, there is a strong need for curation to harmonize naming conventions. As an example, standard representations for dates are used.

The marketplace uses standard vocabularies from different ontologies used in SSH like TADIRAH, NEMO and MORES. The alignment is done during the curation process and provides a common access through common keywords.

The content of the Marketplace will be refreshed regularly by harvesting different sources. As it was mentioned before, the Marketplace is an aggregator of metadata. Metadata are curated (curation team) and enriched (manual & automatized enrichment) by the Marketplace's team. The Marketplace uses a specific model, which was specified and developed by the SSHOC project, to structure data.

5.2.2 Openly accessible data

Generally, all the data collected is meant to be made available online freely through the Marketplace web application. However, this data is subject to curation and entries not fulfilling certain criteria (e.g. considered obsolete or incomplete) may be hidden from public access and only accessible to the moderators of the Marketplace.

The source code of the application is available on Gitlab: https://gitlab.gwdg.de/sshoc, divided into multiple repositories containing the code for individual components: Java-based backend (https://gitlab.gwdg.de/sshoc/sshoc-marketplace-backend) and JavaScript/react based frontend. (https://gitlab.gwdg.de/sshoc/sshoc-marketplace-frontend).

The application exposes all the data also via a publicly accessible API, which is documented using Swagger under: https://marketplace-api.sshoc.eu/swagger-ui/index.html?url=/v3/api-docs. The access to the application is generally possible without any restrictions also to anonymous users. Authentication is needed for users who wish to actively contribute to the content of the marketplace. Authentication will be implemented using the EOSC AAI, i.e. via Federated Identity, where no passwords, or other sensitive information will be stored and managed on the side of the application.

5.2.3 Interoperable data

The Marketplace metadata will be accessible via its API that will allow serialisation as JSON, with a flexible data format as described in Deliverable D7.1, including the possibility to refer to external vocabularies as inventorized in table 23 of Deliverable D7.1 (which is based on the broader vocabulary inventory as provided in Deliverable D3.1).

5.2.4 Re-usable data

As an aggregator, the Marketplace uses when available the license of the original resource. The contents created by the curation team, which are pure metadata, are totally opened as they are designed to facilitate the dissemination. The resources created by the Marketplace are opened and usable immediately without an embargo. About harvested resources, it depends on the nature of the resources: the Marketplace mainly proposes tools which are generally free to use.

The content of the Marketplace is based on existing resources that are harvested regularly. If the resource no longer exists in the original, the curation committee will decide what to do with the entry. Considering existing entries, the sustainability model of the Marketplace (DOA, Task 7.4) ensures the availability of created metadata on the long run. The curation process of the Marketplace, which is community driven, will be based on curation guidelines that will be produced by Task 7.4 and guarantee the quality of the Marketplace's content. The choice of different sources to be harvested by the Marketplace is done at another level by the governing body of the Marketplace and is based on the relevance of the content.

5.3 Allocation of resources

(Meta)data collected and created in WP7 are managed following the work plan of the SSHOC Grant Agreement. Task 7.1 efforts contributed to design the conceptual model and the system architecture of the SSH Open Marketplace, and a large part of Task 7.3 efforts, dedicated to the SSH Open Marketplace interoperability, is also dedicated to these questions. Task 7.4 will address the detailed curation workflow and precise the sustainability plan of the SSH Open Marketplace, including post-project data management aspects.

5.4 Data security

The data collected for the marketplace is backed up regularly by the host OEAW (ACDH-CH). The web servers at OEAW (ACDH-CH) are run in a professional setting including services to check if a website stays online. OEAW (ACDH-CH) hosts many websites, therefore having an excellent expertise on data storage, security and website maintenance. There are established workflows to rely on.

5.5 Ethical aspects

There are no ethical nor legal issues concerning the SSHOC Marketplace data.

6. SSHOC user communities data

6.1 Ethnic and migration studies

6.1.1 Data Summary

Ethnic and migration studies (Data Communities) is building on and supporting the work undertaken by ETHMIGSURVEYDATA, a COST Action[27] and data user community that brings together researchers from all sectors (academic, think tanks, government, civil society organizations, private companies) to improve access, usability, dissemination, and standards of the multiple and scattered quantitative survey data that exist on the integration of ethnic and migrant minorities (EMMs). The SSHOC project will therefore work with ETHMIGSURVEYDATA to compile, document, and archive a large amount of data from various survey-based/including studies conducted in Europe (35 countries that have formally joined ETHMIGSURVEYDATA[28]) on EMMs' integration, so that this data can be made available on a data hub[29] that will respect, whenever possible and feasible, the FAIR (findable accessible, interoperable, reusable) principles. More concretely, this means that the Task 9.2 and ETHMIGSURVEYDATA will co-develop 2 main components of the aforementioned data hub:

- EMM Survey Registry: A free, online, and user-friendly tool that will allow users to search for and discover specific surveys using the survey-level metadata compiled and documented by Task 9.2 and ETHMIGSURVEYDATA, as well as survey-level metadata shared and added to the registry by data producers themselves.
- EMM Question Data Bank (pilot version): A pilot version of the EMM-dedicated component of the CESSDA-led Euro Question Bank (EQB) to test the feasibility of setting up a full-scale version that allows users to search for specific question items, as well as learn about existing survey questionnaires through the questionnaire and question-level metadata compiled and documented by Task 9.2 and ETHMIGSURVEYDATA.

Data generated

Task 9.2 (and ETHMIGSURVEYDATA) have been conducting a comprehensive search of existing surveys on EMMs' integration in the 35 countries participating in ETHMIGSURVEYDATA. The completion of this task will enable Task 9.2 (and ETHMIGSURVEYDATA) to generate the following types of data using the dataset(s), questionnaire(s), and technical documentation of the identified surveys:

[27] COST actions website: https://www.cost.eu/ [30.06.2020]

[28] more on COST Action website: https://www.cost.eu/actions/CA16111/#tabs|Name:overview [30.06.2020]

[29] Ethmig Survey Data hub website: https://ethmigsurveydatahub.eu/ [30.06.2020]

- survey-level metadata (e.g. the key features of the survey, the target population, sampling methods, sample sizes, data collection information, data availability, data ownership and distribution) for the identified surveys;
- questionnaire and question-level metadata (e.g. the key features of the questionnaire, the key questionnaire topics/concepts, the questionnaire design and implementation, questionnaire availability, questionnaire ownership and distribution) for the identified surveys with accessible questionnaires; and
- for each of the identified surveys with accessible questionnaires, extracted (and translated) text of the questionnaire used (including the specific question items).

The data generated by Task 9.2 (and ETHMIGSURVEYDATA) will subsequently be used to develop the EMM Survey Registry and the pilot iteration of the EMM Question Data Bank. Specifically, each of these components will require the use of a particular type/types of generated data:

- EMM Survey Registry: This component requires the use of the survey-level metadata.
- EMM Question Data Bank (pilot version): This component requires the use of the questionnaire and question-level metadata and the extracted (and translated) questionnaire text.

Individuals outside of Task 9.2 and the ETHMIGSURVEYDATA membership (i.e. external parties) will be able to share and add their own survey information (as survey-level metadata, following the compilation process undertaken by Task 9.2 and ETHMIGSURVEYDATA) to the EMM Survey Registry through an online form. As such, Task 9.2 will be involved in screening, reviewing, and managing the survey-level metadata that has been submitted via the EMM Survey Registry to ensure that these new survey records are consistent with the standards and processes followed by Task 9.2 and ETHMIGSURVEYDATA.

Data collected - internal

The survey, questionnaire, and question-level metadata will be generated based on information that is provided in and can be gleaned from existing survey dataset(s), the survey questionnaire(s), and technical documentation.

The accessibility of the survey dataset(s), survey questionnaire(s), and technical documentation will inevitably vary from survey to survey; as such, they could be accessed, for example, through data archives/repositories, project/study websites, institutional websites, and the researcher or research team directly. To generate the questionnaire text, Task 9.2 (and ETHMIGSURVEYDATA) will be using the original survey questionnaire(s) (all languages that have been made available). Any translations of the extracted text will also be based on the original survey questionnaire(s).

It should be noted that the data collection process for the survey-level metadata also includes a protocol for obtaining the necessary consent to store and use the metadata on the EMM Survey Registry. Specifically, the data management/science experts (i.e. those from the SSHOC project consortium) and data archives/repositories have advised Task 9.2 (and ETHMIGSURVEYDATA) that:

- surveys deposited to a data archive/repository require explicit and written consent from the data archive/repository to store and reuse the metadata on the EMM Survey Registry;
- surveys accessed by Task 9.2 and ETHMIGSURVEYDATA by contacting the researcher or research team directly is a form of implicit consent and no further consent is needed to store and reuse the metadata on the EMM Survey Registry; and
- surveys accessible through an online platform that is not a data archive/repository (e.g. a project/study website, institutional website) do not require consent to store and use the metadata as the survey and its metadata are considered "public" information.

For the questionnaire and question-level metadata that will be generated by Task 9.2 (and ETHMIGSURVEYDATA), as well as the extracted (and translated) questionnaire text, a consent protocol will also be identified and adopted (and potentially embedded into the data collection process) if the data management/science experts advise that this is necessary.

Data collected - external

This part refers to data collected by parties with no affiliation with Task 9.2 or ETHMIGSURVEYDATA but is regulated by Task 9.2). As mentioned above, data producers outside of Task 9.2 and ETHMIGSURVEYDATA will also be able to share and add survey-level metadata about their respective survey to the EMM Survey Registry using an online form.

They will be using their own survey dataset(s), survey questionnaire(s), and technical documentation to provide the requested survey-level metadata. As such, Task 9.2 will not be involved in collecting the survey-level data, as it will be the responsibility of the data producers themselves. Whenever a data producer submits their survey's metadata to the EMM Survey Registry, they will be presented with a disclaimer/waiver that:

- confirms that they give Task 9.2 and ETHMIGSURVEYDATA with consent to store and use their survey's metadata; and
- details how Task 9.2 and ETHMIGSURVEYDATA will be storing, handling, etc. their survey's metadata.

Size and volume of the generated data - internal

For the survey-level metadata, each of the 35 countries will produce the following final products/outputs: (i) one Excel file (fully quality controlled and cleaned) that documents the survey-level metadata that has been compiled for each of the identified surveys for that country and (ii) one file that is a transformation of the aforementioned Excel file into a format that is readable and accessible through a statistical program (e.g. do file for STATA, .sav or .sps file for SPSS).

For the questionnaire and question-level metadata, as well as the extracted (and translated) questionnaire text, each of the pilot participating countries will generate one record in a data documentation tool (e.g. Colectica, GESIS Question Editor) for each of their identified surveys with an accessible questionnaire. This tool will also need to facilitate integration into the EQB.

The size and volume of the different data files (i.e. those that have been identified above) will vary. The details are presented/described below:

- Survey-level metadata: The size or volume of the individual Excel file and the file that is readable and accessible through a statistical program will depend on the number of identified surveys for that country and the amount of metadata that has been compiled for each of the surveys. It should be noted that Task 9.2 and ETHMIGSURVEYDATA anticipate that, on average, a country will identify roughly 50 unique surveys.
- Questionnaire and question-level metadata: The size or volume of each record will depend on the amount of metadata that has been compiled for the specific survey questionnaire.
- Extracted (and translated) text from questionnaires: The size or volume of each extracted text will depend on the length of the survey questionnaire itself or the quantity of text that needs to be extracted.

Size and volume of the generated data - internal

Starting in 2020, the EMM Survey Registry will allow data producers to directly share and add survey-level metadata for their respective survey for an indefinite period. Since it will be the data producers themselves who will be compiling and documenting the survey-level metadata onto the online form, it is expected that each new survey record will contain detailed information. The frequency at which the online form will be used is unknown and will depend on how effectively Task 9.2 can encourage data producers to add and share their survey information to the registry.

6.1.2 FAIR Data

All in all, the data generated by Task 9.2, ETHMIGSURVEYDATA, and the external parties— as they will be made accessible via the data hub—will enable researchers and other relevant stakeholders to improve how they access, (re)use, share, and work with quantitative survey data on EMMs' integration.

Resharing the generated data

Table 1 describes how each type of data generated by Task 9.2 (and ETHMIGSURVEYDATA) will be made accessible to user communities interested in working with and learning about quantitative surveys on EMMs' integration.

For any survey dataset, survey questionnaire, and/or technical documentation that was used to generate the data types identified in Table 1, Task 9.2 (and ETHMIGSURVEYDATA) will provide information on how to access these sources (e.g. providing a URL to the data archive where the survey and its dataset(s), questionnaire(s), and technical documentation have been deposited; identifying the party or parties to contact to access or enquire about the survey and its dataset(s), questionnaire(s), and technical documentation).

Table 1: Accessibility of Data Generated by Task 9.2 (and ETHMIGSURVEYDATA) to User Communities

Type	Access Conditions	Location	Temporality	Licenses	Restrictions
Survey-level metadata	Open	EMM Survey Registry	Metadata will gradually become available starting in fall/winter 2019	FAIR-compliant open license	The metadata will only be made available via the registry if the necessary usage consent has been obtained. No (re)usage restrictions will be placed on any metadata made available via the registry, though proper citation/referencing will be requested.
Questionnaire and question-level metadata	Open	EMM Question Data Bank (pilot version)	Metadata will gradually become available starting in 2020	FAIR-compliant open license	The metadata will only be made available via the question data bank if the necessary usage consent has been obtained. Restrictions will likely not be placed on any metadata made available via the question data bank. Proper citation/referencing will also likely be requested.
Extracted (and translated) questionnaire text	Open	EMM Question Data Bank (pilot version)	Data will gradually become available starting in 2020	FAIR-compliant open license	The data will only be made available via the question data bank if the necessary usage consent has been obtained. Restrictions will likely not be placed on any metadata made available via the question data bank. Proper citation/referencing will also likely be requested.

Identification of the generated data

The EMM Survey Registry will assign a unique identifier to each survey and its corresponding survey-level metadata. This will allow an individual to quickly locate a specific survey in the EMM Survey Registry, as they will be able to search by the unique identifier.

The pilot version of the EMM Question Data Bank may also utilize an identification process. The decision to implement an identification process will be determined during the design and development phase of the specific component, in collaboration with CESSDA/GESIS (i.e. the parties responsible for developing the EQB).

6.1.3 Allocation of resources

For the datahub, Task 9.2 (and ETHMIGSURVEYDATA) selected Dot Design Web S.R.L to provide hosting services from 2019-04-01 to 2021-03-31 (with the option to renew). This company was selected as their services sufficiently met the technical, security, and protection needs for the data hub. For the actual development of

the data hub, only the EMM Survey Registry component is currently underway; Youngminds is the IT company the project has contracted to design and develop the EMM Survey Registry. For both the hosting services and the development of the EMM Survey Registry, the costs have been shared between Task 9.2 and ETHMIGSURVEYDATA. For the pilot version of the EMM Question Data Bank, Task 9.2 and ETHMIGSURVEYDATA will need access to some sort of metadata documentation tool that will allow integration with the EQB. Once a specific tool is selected, decisions regarding costs will be made. Sciences Po team Task 9.2, leads and oversees the overall data management work.

6.1.4 Data security

The survey-level metadata that will be compiled by Task 9.2 (and ETHMIGSURVEYDATA) will be stored in different ways: (i) the Excel file versions of this metadata will be stored on the Sciences Po institutional Google DRIVE, (ii) the versions of this metadata that can be read and accessed through a statistical program will also be stored on Sciences Po institutional Google DRIVE, and (iii) the metadata that has been imported into the EMM Survey Registry will be stored directly onto a secure server that is paid for and managed by Task 9.2 and ETHMIGSURVEYDATA. Moreover, Task 9.2 (and ETHMIGSURVEYDATA via its country representatives) will be using the survey dataset(s), survey questionnaire(s), and/or technical documentation to produce the survey-level metadata. As such, if the survey dataset(s), survey questionnaire(s), and/or technical documentation cannot be accessed, viewed, and analyzed directly on an online platform/service, the individual compiling and documenting the survey-level metadata will likely store such data onto a secure personal/work computer. As for the survey-level metadata that will be generated by external parties, they will be able to do this directly on an online form that is made available through the EMM Survey Registry. Once a data producer submits a completed form, a new record will be created and stored on the secure EMM Survey Registry server.

Finally, the data generated for the pilot version of the EMM Question Data Bank will be stored as part of the EQB. For any survey questionnaire that needs to be accessed in order to produce the questionnaire and question-level metadata, as well as to extract (and translate) the questionnaire text, they will be stored on an EU-based, GDPR-compliant, and secure online platform called, MyCore.

6.1.5 Ethical aspects

This project will use secondary data/sources (i.e. the survey dataset(s), survey questionnaire(s), and/or technical documentation) to produce the survey-level metadata, questionnaire-level metadata, and extracted (and translated) text from survey questionnaires. As the secondary data/sources used by Task 9.2 (and ETHMIGSURVEYDATA) could contain personal or individual-level data, Task 9.2 (and ETHMIGSURVEYDATA) will ensure that such data is accessed via a secure network, is used only after obtaining the necessary permissions from the data owner/distributor, and is used in full compliance with the data usage provisions set forth by the data owner/distributor.

6.2 Electoral studies

6.2.1 Data Summary

This Data Community Project (T9.3) aims to generate an Open Research Knowledge Graph in the field of Electoral Studies to be included in the SSHOC project. It starts with the subfield of electoral behaviour and motivations of citizens (other subfields, e.g., behaviour of political parties and elites, or aggregate outcomes, can be added later). Taxonomies and ontologies underlying the Knowledge Graph will be developed with active involvement of the user community. The project will harvest-integrate-interlink-analyse national, European and cross-national data on citizens' electoral behaviour (mainly survey data) and relevant additional data (e.g. contextual data) from existing (open) sources and repositories. This task builds upon existing foundational work that has been conducted by e.g., the European Voter and the COST-TEV projects, and the European Election Studies. In this project Linked (Open) Data principles and technologies will be used, as well as W3C standards RDF and SPARQL. The overall results will be integrated in an election studies analytics dashboard to access analytics and visualisations in an easy to use manner, to export results, and to query the data (for more advanced users) via a data API for further analysis using common stat packs such as STATA, SPSS or R. The project and some of its services will be provided as part of the SSHOC Marketplace. To maximise the actual use of the outcomes by the user community, and to enhance that the resulting Knowledge Graph will be self-sustaining and continuously updated, the project will be characterised by a demand and user-driven approach.

Task 9.3 does not involve major primary data collection. Instead, it relies heavily on existing data (mainly survey data) that have been collected by various research groups in the field of electoral studies. Such data, as collected by, e.g., National Election Study teams in many countries, by the European Election Studies team, or similar organisations, are almost invariably made available for free to the research community by way of being archived in and curated by various data archives, such as GESIS (in Germany), DANS (in the Netherlands) , AUSSDA (in Austria), and their equivalents in other countries. Additional data generation will take place in the context of the development of the aforementioned Knowledge Graph, which requires expert coding of research publications to provide a machine learning training base for classification of scientific literature, which, in turn, will be one of the building blocks of the Knowledge Graph.

The T9.3 will not collect new data, other than expert coding of research publications as a prerequisite for the development of a machine-learned algorithm. These expert codes will not produce a data set but will constitute an intermediate product in the development of the Knowledge Graph. The project relies on existing data which are curated by established data archives in Europe. These data are mainly survey data, with their associated meta-data. The re-use will be particularly at the meta-data level and consist of interlinking meta-data information in the form of a so-called Knowledge Graph. Because the aim of the project is to contribute by building infra-structural tools for the research community, the re-use will not consist of analyses aimed at answering substantive research questions in the domain of electoral studies.

The origin of the data to be re-used are the principal investigators c.q. research teams of the surveys involved. These are National Election Study teams in various European countries, the team of the European Election Studies, occasionally individual principal investigators, and similar. These teams and investigators have almost

invariably been awarded public funds (from governments, research councils, occasionally from charitable organisations) to collect primary data on the condition that these data are to be professionally documented, archived and made available to the entire scientific community. This condition is fulfilled through archiving in established data archives.

Size of new data to be collected is nil, as no new primary data collection is intended (see above). The size of the data to be re-used is considerable. At its current maximum the group of studies involved will be approximately 200 surveys, varying in size between 100Mb and 2Gb (with a median size of ~400Mb).

The re-use of data is focussed on the construction of a research tool (the Knowledge Graph, see above) that is in first instance meant for academic researchers in the field of electoral studies. It is foreseen that this tool will also be useful for others who have an interest in empirical answers to substantive questions with respect to the electoral behaviour of citizens, and who may work in government, media, education, think-tanks and industry

6.2.2 FAIR Data

For Data Community Project Electoral Studies, no new primary data collection will be undertaken. Data that will be (re-)used are all already openly accessible (and FAIR), and already archived in and curated by established data archives.

In Data Community Project Electoral Studies project specific ontologies will be developed which are mostly at a more detailed level than commonly used ontologies, and that can therefore be linked to those.

6.2.3 Allocation of resources

For Data Community Project Electoral Studies, the Task Leader (Cees van der Eijk) will be responsible for data management (which pertains mainly to re-use of data openly available and archived/curated in established data archives).

6.2.4 Data security

This heading is not applicable for Data Community Project Electoral Studies because no primary data collection will be undertaken, as indicated in Section 6.2.1.

6.3 Heritage Science and Humanities datasets

6.3.1 Data Summary

Purpose of data generation, collection and use is to foster innovative techniques in scientific study, interpretation and preservation of Cultural Heritage objects, by using complementary knowledge produced by a range of disciplines, by laboratories and institutions across Europe.

Due to the complexity of the research questions and the inherent characteristics of the experiments on Cultural Heritage artefacts, it is expected to have an heterogeneous data landscape with a mix of structured, tabular, unstructured and semi-structured data.

The main type of data will be scientific, explored by users and access providers in the context of the IPERION HS project[30]. Each of the 52 participating facilities may manage several data sets. In total more than 100 different kinds of data sets are expected within the partnership. The main difficulty lies in identifying and documenting the various possible sets, which can be split into the following categories:

- completion of an existing dataset by means of newly created data
- creation of a new data set
- new conditions attached to existing data sets facilitating their re-use.

A general overview of such data sets and their management at facility level is provided:

- Digitized data: all data (raw and processed) are either digital or digitized.
- Formats: most of the data is stored in proprietary formats (readable and non-human readable formats) and/or "standard" formats (docx, tiff, pdf, jpg, xlsx etc.).
- Raw data: these are obtained through scientific studies, after being processed by researchers themselves.

Valuable information and metadata generated by the HS community will be reused in various contexts and research scenarios, across disciplines, actions, research and services. A culture of opening and reusing digital data will support researchers in preserving, protecting and enhancing the significance of tangible and intangible cultural heritage as well as the long-term sustainability for heritage data.

Open directories and repositories of digital data will be created hosting information about the available instruments and archives, as well as scientific project results. These will be made available online. There is a strong commitment in the HS community to effectively manage the generated research data. SSHOC partners will build on the track record of the previous integrating activity (IPERION CH) where a Data Management Plan (DMP, IPERION CH Deliverable n. 2.2) was developed and rigorously maintained. The IPERION HS DMP will extend the existing document and adapt it. The goal of the IPERION HS DMP is to define common practices and recommendations in terms of curation, storage, formatting, dissemination and licensing of data used in the frame of IPERION HS.

Open directories and repositories of digital data will be created hosting information about the available instruments and archives, as well as scientific project results. These will be made available online. Curation of IPERION HS data will be carried out, and the associated costs borne by, the E-RIHS ERIC, which will be established in time to inherit and manage the IPERION HS data.

[30]IPERION HS website: http://www.iperionhs.eu/ [30.06.2020]

6.3.2 FAIR Data

Data produced by the Heritage Science community will be compliant to the rules agreed within the DIGILAB Working Group, launched on June 1st, 2020, and will be interoperable with the EOSC PID Policy. The Digital Platform of E-RIHS (DIGILAB)[31] is still in the construction phase. The implementation rules will be discussed and decided in a specific task force within the DIGILAB Working Group, launched on June 1st, 2020. Search filters will be available to foster re-use of data.

By investigating and comparing practices for documenting data and processes (including the core set of metadata fields/terms required to describe it) a standardised and semantically mapped cataloguing system based on metadata will be created for the future E-RIHS DIGILAB platform, as a state-of-the-art research tool that can help users find the most relevant data sources.

The current DIGILAB policy is based on stable requirements of the EU Commission, such as the Open Access EU strategy, the EOSC and the implementation of the FAIR principles, and with good practices such as those adopted for example by DANS-KNAW, it is anticipated that any implementation will need to comply with such general policies[32]. Access to the E-RIHS DIGILAB data will be based on a Wide access mode. The possibility of a Market-driven access mode in accordance with the planning of E-RIHS ERIC commercial activities, may be also considered, according to the global E-RIHS strategies to be developed by the project. Wide access promotes a broad digital access to scientific data and digital services provided by E-RIHS to users wherever they are based and following the FAIR and Open Access principles[33]. Datasets registered in the E-RIHS DIGILAB Catalogue and stored in local repositories may belong to one of the following categories:

- Open Access Dataset (CC0). No restriction applies for accessing the data. Dataset metadata used in the Catalogue always belong to this category.
- Open Access to Registered Users Dataset. The dataset is made available to all registered users of the E-RIHS DIGILAB, without further restriction.
- Restricted-Access Dataset. The dataset is exclusively available to registered users who have received permission from the data owner or repository manager.
- Dataset with Restricted-Access Data. Registered users may access the dataset, with the exclusion of restricted access data, for which a specific permission is required[34].

Data will be made accessible by the implementation of open data resources fostering virtual access to data and tools for heritage research and searchable registries of multidimensional images, analytical data and documentation from large academic as well as research and heritage institutions. Wide access mode will be

[31] E-RIHS website; on DIGILAB: The new platform for Heritage Science: http://www.e-rihs.eu/ercim-news-digital-humanities/ [28.06.2020]

[32] E-RIHS project Access Policy http://www.e-rihs.eu/wp-content/uploads/2020/02/D5.1_User-strategy-and-access-policies.pdf [30.06.2020]

[33] E-RIHS project Access Policy http://www.e-rihs.eu/wp-content/uploads/2020/02/D5.1_User-strategy-and-access-policies.pdf [30.06.2020]

[34] E-RIHS project Access Policy http://www.e-rihs.eu/wp-content/uploads/2020/02/D5.1_User-strategy-and-access-policies.pdf [30.06.2020]

the default for digital activities on E-RIHS DIGILAB. However, due to specific user requirements and scenarios, particular research questions and - eventually - depending on the kind of data the researcher needs to carry on his/her research, custom tools may be required in addition to the single digital entry point[35].

Open directories and repositories of digital data will be created hosting information about the available instruments and archives, as well as scientific project results. These will be made available online.

E-RIHS will strengthen the integration and interoperability of the facilities with both Operation Level Agreement (OLA) and Service Level Agreement (SLA), still to be defined within the context of IPERION HS project.

E-RIHS supports an Open Access policy to data. However, access limitations may be necessary for copyright reasons, personal and privacy-sensitive data protection, and to protect legitimate Intellectual Property Rights (IPR). E-RIHS abides by the DARIAH Heritage Data Re-Use Charter.

The Access Board will be an IPERION HS internal advisory body composed of access experts helping the integration of user services. The AB will serve as a forum to ensure continued interaction and exchange of experiences between platforms on common issues, such as IPR, interoperability and data management. The E-RIHS strategy and policies to enable users' access to its physical and virtual research facilities are reported in a deliverable describing different modes according to the physical or virtual nature of the facilities involved[36]

User access in E-RIHS is currently regulated as follows:

- Anonymous E-RIHS DIGILAB users may only access the Catalogue and Open Access datasets;
- Registered users can access Open Access Datasets, Open Access to Registered Users Datasets and Datasets with Restricted-Access Data with the exclusion of the restricted-access data.

The E-RIHS DIGILAB Authentication, Identification and Authorization System will provide a federated identity system to all participating data managers and to all the categories of users[37]. IPERION HS will contribute cutting-edge services to heritage science research and joint innovative research for improved interoperability not limited to data, but including sample and reference materials, methods and instruments. The networking activities aim at reinforcing integration of the group and at creating a sense of belonging for heritage science researchers which will take advantage of the RI services. The objective is to increase the level of integration of heterogeneous data related to heritage artefacts to fully represent their complexity in the digital domain and to match the research needs of the heritage science community. This will constitute a pilot action in IPERION HS towards the development of a Heritage Data Integration suite allowing semantic linkage of data and related annotations, vocabulary terms, attributes, based on real-world research test-beds encompassing selected

[35] E-RIHS project Access Policy http://www.e-rihs.eu/wp-content/uploads/2020/02/D5.1_User-strategy-and-access-policies.pdf [30.06.2020]

[36] E-RIHS project Access Policy http://www.e-rihs.eu/wp-content/uploads/2020/02/D5.1_User-strategy-and-access-policies.pdf [30.06.2020]

[37] E-RIHS project Access Policy http://www.e-rihs.eu/wp-content/uploads/2020/02/D5.1_User-strategy-and-access-policies.pdf [30.06.2020]

analytical technologies (e.g., data from XRF, XRPD, Raman, OCT, 14C dating) used by researchers, to define common workflows to be used for interoperable descriptions of produced datasets.

The goal of the E-RIHS RI is to promote global interoperability through alignment of protocols and good practices and exchange of reference materials; foster interoperability and reuse of datasets and reference collections, created or maintained by the project. Creating usable guidelines and developing standards and recommendations for data dissemination will be a continuous project activity.

Conditions for data re-use must be specified in each dataset metadata. Such conditions may refer to general licensing schemes (e.g. Creative Commons) or to specific re-use conditions established by the data owner/depositor as well as by legal constraints, e.g. a non-transferable use license. A "re-use document" summarizing legal re-use constraints should be available in each participating repository and referenced in each dataset. E-RIHS abides by the DARIAH Heritage Data Re-Use Charter.

Furthermore, IPERION HS will be investigating best practice for preparing data for sharing and re-use, including managing/storing and formatting the data output, license requirements etc., to ensure a shared consensus on description and data interoperability. This action will work towards the design of a shared repository for E-RIHS. Conditions for data re-use must be specified in each dataset metadata. Such conditions may refer to general licensing schemes (e.g. Creative Commons) or to specific re-use conditions established by the data owner/depositor as well as by legal constraints, e.g. a non-transferable use license. A "re-use document" summarizing legal re-use constraints should be available in each participating repository and referenced in each dataset[38].

Limitations may be necessary for copyright reasons, personal and privacy-sensitive data protection, and to protect legitimate Intellectual Property Rights (IPR). E-RIHS abides by the DARIAH Heritage Data Re-Use Charter[39].

The quality system proposed to be adopted by E-RIHS for the quality assessment of prospective new partners and their services and for the quality audit of existing E-RIHS partners and their services are described in a specific document. It also outlines the process to grant external organizations, services, projects and proposals the affiliation to E-RIHS, or its support. All such procedures are based on a modular operation: the evaluation of the candidate's internal processes, of its scientific excellence and of the quality of its services and eventual suitability for E-RIHS.

6.3.3 Allocation of resources[40]

In order for E-RIHS to fulfil the promise of DIGILAB, governments, funders, organisations and institutions will need to assess how best to move to a funding model that supports making their data open. This will not be a

[38] D5.1 Access Policy http://www.e-rihs.eu/wp-content/uploads/2020/02/D5.1_User-strategy-and-access-policies.pdf
[39] https://datacharter.hypotheses.org/
[40] Within E-RIHS PP a Report on Data Management Policy (D3.3) to address financial aspects of data policy and management was elaborated

straightforward process and will require long-term commitment to change across a range of sectors and may greatly differ across countries.

An exact evaluation of costs for making data fair is still not available, since the DIGILAB digital infrastructure has not been implemented. The DIGILAB Working Group will support this kind of evaluation providing updated reports. E-RIHS currently plans to earmark €100,000.00 annually for digital infrastructure management and curation (DIGILAB)[41]. A recent survey indicated that the funding models used by the E-RIHS partner institutions include both National/Regional Governmental Funding and Project Funding (also in combination); Direct Charging and/or Subscription Funding or Commercial Partnership are - for the moment - less widespread[42].

There will be tailored DMPs for each platform (mobile, fixed etc) within E-RIHS. During the preparatory phase there will be a Working group to manage the process. The future ERIC will have a specific figure to address this issue. A recent survey indicates that within the Heritage Science community the preservation and dissemination of HS data is already supported within the current funding model. However, the Long Term Sustainability for Heritage Science Data is currently under discussion within the DIGILAB Working Group (established on June 1st, 2020) as one of the pillars of the future DIGILAB. It will also be one of the main entries in the E-RIHS Financial Plan. The exact amount of resources needed, though, has not been determined yet (see above).

6.4 "Historical high-quality company-level data for Europe" design study

6.4.1 Data Summary

EURHISFIRM "Historical high-quality company-level data for Europe" is a design study to build a world- class research infrastructure (RI) compliant to the FAIR (findable, accessible, interoperable, reusable) data principles. The project aims to increase the accessibility and usability of historical company-level data (financial, governance, and geographical) and to expand the available pool of this data.

The recent economic crisis, usually called the Great Recession, has drawn comparisons with the Great Depression in terms of economic and historical impact. While the causes are complex and spread across various social, historical, economic factors and beyond, examining the financial markets remains a high priority to fully understand the cause and effects[43].

The weak empirical foundations of the models used to analyse structural and cyclical changes have become obvious in the recent fierce debates on how to foster economic growth and job creation. One of the main reasons for this uncertainty is the lack of high-quality, long-term and FAIR data on European companies for testing these models.

[41] referring to E-RIHS project D3.1 Preliminary Financial Plan

[42] referring to E-RIHS project D3.3 Data Management Policy

[43] Economic Crisis in Europe: Causes, Consequences and Responses: http://ec.europa.eu/economy_finance/publications/pages/publication15887_en.pdf [30.06.2020]

An in-depth analysis of existing company-level data and historical serial sources is carried out for three main types of information related to firm characteristics: a) financial data (stock market data such as securities issued, prices, dividends and coupons, number of traded securities, corporate events such as (reverse) splits, mergers, balance sheets and income statements); b) information on the companies' governance (e.g. evolution of the juridical status, directors, voting and governance rules), and geographical data (e.g. location of headquarters, subsidiaries, and production units). The origin, types, and size of the existing sources and data will become clearer as the project progresses. WP4 prepares an inventory of existing data and sources.

It delivers in-depth knowledge on the type, quality, and other key characteristics of the available data and sources such as yearbooks of companies and stock exchange lists. It produces accurate data and sources documentation according to the chosen common documentation standard (DDI Lifecycle[44]). Specific attention will be paid to data semantics, a scientific challenge particularly relevant to historical data. WP4 explicitly recognizes the methodological challenge of ensuring standardised approaches whilst allowing for idiosyncrasies of the diverse data types from various countries across time.

The project concerns two main data formats: digitised (stored in databases) and raw (not yet digitised and not yet transformed into databases). A few large stand-alone long-term databases have been built by both the academic community and by private companies (e.g. the London Share Prices Database, the Global Financial Data database)[45], but interoperability remains low.

The main goal of EURHISFIRM is twofold: 1) designing the infrastructure to be used by academics and other stakeholders to deposit and connect their data, as well as to 2) inspire new projects of data collection with the next-generation data extraction and enrichment platform developed from EURHISFIRM.

Many members of the consortium have already run extensive data collections. All of them are committed to reverse and integrate their data into the infrastructure to create a first pool of data big enough to raise interest within the "data collectors" community. This gravitational pull will attract already existing data to make them re-usable within the infrastructure once they are documented according to the established data format.

6.4.2 FAIR Data

In order to render the data findable for future users, the EURHISFIRM must select the appropriate metadata format. A number of standards have been under study and the optimal method for the type of data EURHISFIRM envisions has been chosen as the DDI Lifecycle due to its compatibility with historical datasets, especially in dealing with various elements of the data that may change in format and content over time[46].

[44] DDI Alliance website: https://www.ddialliance.org/ [29.06.2020]

[45] It is worthwhile to note the exceptions of the SCOB database at the University of Antwerp and the Data for Financial History
Database at the Paris School of Economics which have been built in a coordinated way (both institutions belong to the EURHISFIRM consortium).

[46] See deliverable EURhisFIRM project D4.1: Information system and documentation standards (author: J Poukens) https://zenodo.org/record/3246455#.XvtOXJMzZUM [30.06.2020]

Unique identifiers will be assigned to datasets stored within the infrastructure and updated in case of several versions of the concerned dataset. Search by keywords will be provided.

EURHISFIRM aims to design an RI of open-access data under the constraint of a sustainable business model. As a research infrastructure, EURHISFIRM aims to become the reference repository location for historical company-level data. The software, method(s) and possibly license(s) required will be decided with the project progression, based on their abilities to provide open access to potential users under the constraint of a sustainable business model. The data, associated metadata, documentation and code will be deposited within the EURHISFIRM infrastructure.

To ensure ethical use of data, as well as to comply to the General Data Protection Regulation (GDPR)[47], the design will enforce that any data that may reveal personally identifiable data will be properly handled.

As the interoperability of historical European company-level is currently low, EURHISFIRM aims to create an RI design to specifically overcome this obstacle.

In the EURHISFIRM project, the consortium will create a common data model to overcome these interoperability challenges in historical European company-level data. The current model developed so far describes a system in which the local data sources will be treated through data integration gateways and then integrated within a common access system through which data users can consume the data. The ideal goal is to increase the interoperability of historical financial data with as many other European countries as possible. The legal details concerning the data license will become more known with the progression of WP3's work.

6.4.3 Allocation of resources

The costs for conforming to the FAIR principles established within the EURHISFIRM project are under calculation. In case of data that do not conform to the FAIR principles, the costs will be paid by institutions willing to deposit data. Data preservation policy will be decided by the governance structure of the EURHISFIRM infrastructure.

6.4.4 Data security

Although EURHISFIRM designs the RI by envisioning an open-access data system within a sustainable business model, proper data security measures are high priorities to ensure that the data are used and maintained with proper handling. These issues are to be examined and agreed upon with establishment of the proper technology infrastructure. These are to be studied and elaborated in future versions of this document.

[47] GDPR: https://ec.europa.eu/info/law/law-topic/data-protection_en [30.06.2020]

7. Other data

7.1 Project management data

In order to successfully facilitate the SSHOC project implementation and monitor the progress of the project activities and objectives, project management data will be produced, collected and archived.

Project management data contain contracts (GA, CA), amendment files, deliverables and other major research results/documents, internal reports, periodic reports, financial statements, financial summaries (updated, filled cost templates), contact list, meeting minutes, agendas, presentations, signature lists, media files or other material (project logo, project ppt template, project dissemination materials, press releases, etc.) and other related documents.

In order to obtain the structured documentation managing, exchange and archiving of the project data, provide access to all project partners and enable successful communication, CESSDA ERIC as the project coordinator established:

1. Internal Project Document Repository for keeping records - hosted in Coordinator owned Google Shared Drive. Location: https://drive.google.com/

   Google Drive (and Google Shared Drive as alternative) is agreed and used as a project document repository and basis for project management, administration and collaboration. It contains all important information needed for ongoing work and collaboration across WPs.

2. Collaboration platform for communication and day-to-day collaboration - hosted by Coordinator Basecamp account. Location: https://3.basecamp.com/4165192/

   This platform is primarily used for project team day to day communication and secondary cloud-based storage solution for the collaborative creation, management and versioning of documents i.e. an addition to storing the documentation in the official Project Document Repository.

3. SSHOC Wiki "Guidelines" for informing partners in detail on rules and procedures in SSHOC.

   In order to facilitate timely and comprehensive communication about all project rules, procedures, both established in official EC documents, and internal project related ones, including the changes in the respective areas, project CO produced an online project Wiki site "SSHOC Guidelines", available internally to all project partners. Location: https://sites.google.com/cessda.eu/sshocwiki/.

   The SSHOC Wiki site contains project policies and guidelines based on the project contractual documents, decisions of the project bodies and identified best practices. The content will evolve throughout the project lifetime.

All project partners have access to all information stored in Internal Project Document Repository, Collaboration platform and SSHOC Wiki. Coordinator grants additional access based on the written request by the project team member.

List of persons having access to SSHOC project repository and collaboration platform is placed in the project Contact list (in a form of a Google sheet stored in the Implementation folder within the SSHOC Document repository). Research or sensitive data should be managed following the specific legal regulative and established best practices at the organisations responsible for the collection and processing of sensitive data. Storing of such data on Google Drive shall be avoided.

The archiving of the project documentation, after the project ends, will follow the procedure described in the GA.

7.2 Training, dissemination and outreach data

SSHOC project training, dissemination and outreach data consist of the social media data, analysis, statistics, webinars participants, publications and articles, SSHOC Web Platform (approach to Market place), events, polls, stakeholders mapping database and other data.

The SSHOC stakeholder landscape analysis was built upon a stakeholder mapping exercise during which each project partner shared their knowledge of and some general information about stakeholder organizations known within their existing networks. For this purpose, a stakeholder mapping database was created that identifies specific organisations within each stakeholder group. This database is GDPR compliant and was set up as a Google spreadsheet within the SSHOC Shared Drive for internal SSHOC project use only and in order to be able to track reach by stakeholder groups. No personal data has been collected, only general information about stakeholder organizations. The collected information include the geographical location of identified stakeholder organizations, the discipline(s) they belong to, the specific SSHOC WPs targeting the stakeholder category they belong to, and networks, channels and tools that project partners already have in place and which connect them to these organizations. This spreadsheet is a living document populated further during the project life.

Several types of events are organised by WP6, in collaboration with relevant SSHOC WPs: raising awareness workshops, raising awareness webinars, training workshops, training webinars, training bootcamps, a mid-term Stakeholder Forum and a final conference.

Participants of face-to-face events register via the SSHOC website, or the site of the conference organisers, when the SSHOC event is co-located with a larger event. Participants list are generated and archived for reporting purposes only, accessible to project partners.

In the case of online events, for which LIBER provides its webinar platform, data collection is done according to LIBER's Institutional Privacy Policy[48]. LIBER collects and uses personal information pursuant to its legitimate

[48] LIBER Privacy Policy: https://libereurope.eu/liber-europe-privacy-policy / [30.06.2020]

interest in organizing and running the relevant event. It may share the collected personal information internally (e.g., with LIBER staff and the SSHOC project partners), but for the purpose of organizing and running the respective event only. If personal information needs to be shared with vendors, sponsors or third party contractors, LIBER always asks for permission to do this in the registration form. However, this hasn't been needed for any of the SSHOC online events run by LIBER so far.

Event management data is not all publicly accessible. It consists of internal communication among event organisers, pre-event announcements, workshop/webinar registration and participation lists, workshop photographs, post-event satisfaction survey and its results, post-event blog posts and reports. In order to pursue the open science principles and the highest level of possible re-use of the data, SSHOC tries to avoid producing sensitive data (e.g., by enabling anonymous replies to post-event survey), and openly share all data that can be made freely available (e.g., post-event blogspot) in order to maximize the outreach of engagement and training events. However, when it is not possible to avoid producing and using sensitive data (e.g., registration lists, photographs), it is processed according to the GDPR regulation (e.g., no sharing, consent obtained).

In terms of surveys, project partners are surveyed on their planned work and WP needs in terms of engagement and training. Answers are used for organising relevant work. Survey data and reports are saved on project internal Shared drives. Personal data of project partners exist in documents available on the SSHOC internal shared drives for internal project communication and organisation of work. Post-event evaluation surveys targeting external stakeholders are anonymous. Data and text are archived for reporting purposes and improving future work.

Any presentations and recording of events are made live with licence and approval of the speaker. Personal data of the speaker (e.g. email) are included by the speakers themselves. Engagement and training materials data consists of workshop/webinar presentation slides and webinar recordings. This data is produced in order to provide knowledge and skills transfer among researchers and thus promote science advancement. Slides and recordings are made openly available for later use in order to expand the reach of training activities (e.g. webinar recording[49] and presentation slides[50]). To this end, well-established platforms will be used, namely Zenodo and YouTube, and clearly state applicable licence for re-use, DOIs, version history and pre-prepared citation formats. This helps us align the data with the FAIR principles.

Polling tools are used in engagement and training activities, either integrating in an online event platform, or as separate software. The tool currently used, that is not embedded in an online event platform, is Mentimeter and is anonymous. Data are stored for reporting and following up in project work in the project's internal shared folders.

SSHOC is fostering the SSH Training Community. Trainers from inside and outside the project subscribe to the Training Community by filling in a form[51] entering email, name, surname, stakeholder type, organisation name,

[49] Example of SSHOC webinar on Youtube: https://www.youtube.com/watch?v=X6bFGJpMjVQ [30.06.2020]
[50] Example of SSHOC presentation slides on Zenodo: https://zenodo.org/record/3694223#.XvtwS5MzYjh [30.06.2020]
[51] SSHOC Training Community form: https://www.sshopencloud.eu/join-ssh-training-community [30.06.2020]

role in organisation, domain, country. The box to read, understand and agree to the SSHOC Privacy Policy should be ticked to become a member. Members of the Training Community are added to an email group list trainingcommunity@sshopencloud.eu. Members of the Training Community are added to the separate Gdrive set up exclusively for the community by the SSHOC project coordinator to share information on upcoming events, planning, minutes of the meetings. The member list of the Training Community and access to the separate Google drive is managed by WP2 and WP6 during the project. Members can cancel their participation at any given time. At the end of the project the member list will be dissolved or continued based on the wishes of the individual members.

Information about existing publicly available training materials have been collected into the Inventory of training materials (D6.7). For the purposes of collecting and curating the data a web-application based on Drupal was used, which has also been adopted for continuous collection of information about training material in Task 6.4 and is available under SSHOC Training Toolkit on SSHOC website[52]. Metadata on training material (sources as well as selected items from these sources) is collected in this database. A custom data model was created for this purpose. The data was gathered almost manually. Only some enrichment of it was done automatically. It is expected that there will be in the end some hundred datasets on sources and items. There is also an API available, where this data can be gathered. Re-use of this data is therefore easily possible and was already done between T6.3 and T6.4. The database may be also of relevance for the Open Marketplace (finding new sources for ingest). The data collected here is backed up regularly by the host ACDH-CH/OEAW. Where the data contains personal information, it is just names of authors, as gathered in publicly available sources. Additionally, for internal curators there are user accounts created in the Drupal database. These user accounts hold an email address and a name. Activities of these accounts are logged for the case of restoring changes that were done wrongly in the data.

When DMP guidelines overlap with those of other stakeholders that are involved in the organisation of specific events, the priority will be to align and respect all applicable DMP guidelines.

[52] SSHOC Training Toolkit: https://training-toolkit.sshopencloud.eu/ [30.06.2020]

8. References

- AIOLI website: http://www.aioli.cloud/ [30.06.2020]
- Annika Schwabe, Julian Ausserhofer, Leonardo Marino, Marieke Willems, Vasso Kalaitzi, Irena Vipavc Bvrar, … Silvana Muscella. (2019). SSHOC D2.1 Overall Communication and Outreach Plan (approved 18 nov 2019) (Version v1.0). Zenodo https://zenodo.org/record/3595936#.Xvtl7ZMzZUM
- Austrian Academy of Sciences: https://www.oeaw.ac.at/ [29.06.2020]
- Broeder, Daan, Trippel, Thorsten, Degl'Innocenti, Emiliano, Giacomi, Roberta, Sanesi, Maurizio, Kleemola, Mari, … Ďurčo, Matej. (2019). SSHOC D3.1 Report on SSHOC (meta)data interoperability problems (Version v1.0). Zenodo. https://zenodo.org/record/3569868#.XvtkBJMzZUM
- CIDOC Conceptual Reference Model (CRM): http://www.cidoc-crm.org/ [30.06.2020]
- CIDOC CRM website: http://www.cidoc-crm.org/ [30.06.2020]
- COST actions website: https://www.cost.eu/ [30.06.2020]
- Creative Commons website: https://creativecommons.org/licenses/by-nc-nd/4.0/ [30.06.2020]
- DDI Alliance website: https://www.ddialliance.org/ [29.06.2020]
- Drama Corpora Project website: https://dracor.org/ [30.06.2020]
- E-RIHS project Access Policy http://www.e-rihs.eu/wp-content/uploads/2020/02/D5.1_User-strategy-and-access-policies.pdf [30.06.2020]
- E-RIHS website; on DIGILAB: The new platform for Heritage Science: http://www.e-rihs.eu/ercim-news-digital-humanities/ [28.06.2020]
- Economic Crisis in Europe: Causes, Consequences and Responses: http://ec.europa.eu/economy_finance/publications/pages/publication15887_en.pdf [30.06.2020]
- Ethmig Survey Data hub website: https://ethmigsurveydatahub.eu/ [30.06.2020]
- EURhisFIRM project D4.1: Information system and documentation standards (author: J Poukens) https://zenodo.org/record/3246455#.XvtOXJMzZUM [30.06.2020]
- European Social Survey European Research Infrastructure – ESS ERIC website: https://www.europeansocialsurvey.org/about/ [30.06.2020]
- European Values Study website: https://europeanvaluesstudy.eu/ [30.06.2020]
- Generations & Gender Programme website: https://www.ggp-i.org/about/ [30.06.2020]
- Getty Vocabularies as Linked Open Data: https://www.getty.edu/research/tools/vocabularies/lod/ [30.06.2020]
- Github repository: https://github.com/ [30.06.2020]

- H2020 Programme Guidelines on FAIR Data Management in Horizon 2020: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf [accessed 30.06.2020]
- H2020 templates: Data management plan v2.0 – 15.02.2018 https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm#A1-template [accessed 24.06.2020]
- International Image Interoperability Framework: https://iiif.io/ [30.06.2020]
- IPERION HS website: http://www.iperionhs.eu/ [30.06.2020]
- Laure Barbot, Yoan Moranville, Frank Fischer, Clara Petitfils, Matej Ďurčo, Klaus Illmayer, … Sotiris Karampatakis. (2019). SSHOC D7.1 System Specification - SSH Open Marketplace (Version 1.0). Zenodo. https://zenodo.org/record/3547649#.XvtrgpMzZUM
- LIBER Privacy Policy: https://libereurope.eu/liber-europe-privacy-policy/[30.06.2020]
- LINDAT CLARIN CZ website: lindat.mff.cuni.cz [20.06.2020]
- LINDAT/CLARIAH-CZ repositories: https://lindat.cz, or https://lindat.mff.cuni.cz [30.06.2020]
- Malter, F. and A. Börsch-Supan (Eds.) (2017). SHARE Wave 6: Panel innovations and collecting Dried Blood Spots. Munich: MEA, Max Planck Institute for Social Law and Social Policy.
- MateCat tool website: https://www.matecat.com/ [29.06.2020]
- OECD Family database: the Family support calculator https://www.oecd.org/els/soc/oecdfamilydatabasethefamilysupportcalculator.htm [30.06.2020]
- ORCID website: https://orcid.org/ [30.06.2020]
- Qualtrics Experience Management Software website: https://www.qualtrics.com [29.06.2020]
- Registration procedure on ESS website: https://www.europeansocialsurvey.org/user/new [30.06.2020]
- Regulation of the European Parliament- General Data Protection Regulation: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN [30.06.2020]
- Schmidutz, D. (2019): "Report on the Feasibility of a 'Broad Consent' Strategy with Regard to the Storage and Use of Biological Samples – Considerations Regarding the Inclusion of Biological Samples in European Population-Based Social Surveys." Deliverable D6.13 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.
- SHARE Methodology Volumes and Compliance Profiles can be accessed via the SHARE homepage: http://www.share-project.org/data-documentation/methodology-volumes.html [30.06.2020]
- SSHOC Gitlab: https://gitlab.gwdg.de/sshoc [30.06.2020]

- SSHOC Grant Agreement (no.823782)
- SSHOC Overall Communication and Outreach Plan (D2.1) - https://sshopencloud.eu/d21-sshoc-overall-communication-and-outreach-plan
- SSHOC presentation slides on Zenodo: https://zenodo.org/record/3694223#.XvtwS5MzYjh [30.06.2020]
- SSHOC Training Community form: https://www.sshopencloud.eu/join-ssh-training-community [30.06.2020]
- SSHOC Training Toolkit: https://training-toolkit.sshopencloud.eu/ [30.06.2020]
- SSHOC webinar on Youtube: https://www.youtube.com/watch?v=X6bFGJpMjVQ [30.06.2020]
- Survey Codings repository website: https://www.surveycodings.org/ [29.06.2020]
- Survey of Health, Ageing and Retirement in Europe (SHARE) website: http://www.share-project.org/organisation/share-eric.html [30.06.2020]
- Weiss, L. M. & Börsch-Supan, A. (2019). Influence of fieldwork conditions and sample quality on DBS values. in: Health and socio–economic status over the life course. A. Börsch-Supan, J. Bristle, K. Andersen-Ranberg et al. (Eds.) Berlin/Boston, de Gruyter: 359.