Research and Innovation Action

Social Sciences & Humanities Open Cloud

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

Deliverable 4.19 Mapping of two indicative selected standards to the SSHOCro

Dissemination Level	PU
Due Date of Deliverable	31/12/20 (M24)
Actual Submission Date	15/12/2020
Work Package	WP 4 - Innovations in data production
Task	T4.7 Modeling the SSHOC data life cycle
Type	Report
Approval Status	Waiting EC Approval
Version	V1.0
Number of Pages	p.1 – p.61

Abstract:
The report concerns the mapping of two selected metadata standards used to document social science research –namely DDI (Document Discover Interoperate) Codebook and CMDI (Component Metadata Infrastructure) –to the SSHOC Reference Ontology (SSHOCro). SSHOCro is a common meta-level schema based on CIDOC CRM, that aims at providing a semantic interoperability framework for the description of the data life cycle used by social scientists and humanities researchers. The task includes the integration and harmonization of the DDI and CMDI metadata schemas and the transformation of selected cases from social sciences and humanities repositories, documented with the DDI and CMDI to the SSHOCro ontology.

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

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
0.0	27/11/2020	first draft	AK, CB, ET
0.1	30/11/2020	WP4 leader review	Diana Zavala-Rojas
0.2	30/11/2020	WP4 partner review	Maurizio Sanesi
0.3	01/12/2020	WP4 partner review	Isabelle Cao
0.4	09/12/2020	Address partner comments	Eleni Tsouloucha
0.5	10/12/2020	Address partner comments	Chryssoula Bekiari
1.0	14/12/2020	Address CO comments	Chryssoula Bekiari, Eleni Tsouloucha

Author List

Organisation	Name	Contact Information
FORTH	Eleni Tsouloucha	tsoulouha@ics.forth.gr
FORTH	Athina Kritsotaki	athinak@ics.forth.gr
FORTH	Chryssoula Bekiari	bekiari@ics.forth.gr
FORTH	Maria Theodoridou	maria@ics.forth.gr

Executive Summary

This report documents the work undertaken in the frame of the SSHOC project Task 4.7 Modeling the SSHOC data life cycle. Specifically, SSHOC Deliverable 4.19 concerns the mapping of two selected metadata standards used to document social science research –namely DDI Codebook and CMDI –to the SSHOC Reference Ontology (SSHOCro). The mapping involves the integration and harmonization of DDI and CMDI metadata schema and the transformation of selected cases from social sciences and humanities repositories, documented with the DDI and CMDI to SSHOCro.

SSHOCro is a common meta-level schema based on CIDOC CRM, that aims at providing a semantic interoperability framework for the description of the Social Sciences and Humanities data life cycle, by offering a conceptual model that can describe at a generic level the real-world lifecycle of data produced by the generic workflow of processes of collection, connection, interpretation and the auxiliary activities of storing, publishing and finding data –as it actually takes place in the various domains of Social Sciences and Humanities. Its development has, in fact, been informed by data lifecycle management practices in use, in said disciplines. In practical terms, the use of such a model and schema for the research community is twofold: (a) it can be applied as a standard to be used in the step of devising and implementing metadata capture scheme for tracking the data lifecycle in individual projects, institutions and disciplines; (b) it is a canonical form or target schema that can provide a model for a single knowledge base for cross–domain tools and services (e.g., resource discovery, browsing, and data mining). In this case mappings must be produced to relate DDI or CMDI concepts or relationships (source schemata) to SSHOCro concepts (target schema) in a way that facts described in terms of the above source schemata can automatically be translated into descriptions in terms of the target schema (SSHOCro). This is the mapping definition process and the output of this task is the mapping, i.e., a collection of mapping rules. The present report describes the mapping definition process between the DDI / CDMI and SSHOCro along with their resulting mapping rules.

This task includes the following activities:

1. interpreting conceptualizations expressed in DDI and CMDI and of concepts necessary to explain the intended meaning of DDI and CMD attributes and relationships, especially those related to the data life cycle, in terms of SSHOCro v.0.1. Any conflicts occurred in the harmonization process with the existing version of SSHOCro have been resolved on the SSHOCro side producing a new version of SSHOCro (v.1.1.3). The output of this task is the listing of the entities, relationships and attributes defined in DDI and CMDI which shows how the same information can be expressed using SSHOCro. These listings are the mappings at schema level which can be served as an intellectual definition of the relationship between DDI and CMD with SSHOCro. Also, they are in a format that could be turned more or less mechanically into an algorithm to automatically transform data structured following the one form into data in the other form, i.e., they can be used to implement an automatic data translation.

2. transforming selected cases found in social science repositories into the SSHOCro v1.1.3. The tool used for the transformation of the data of the selected cases was the X3ML (3M) Toolkit; a set of small, open source, micro services that follow the Synergy Reference Model of data provision and aggregation (SRM), which defines a consistent set of business processes, user roles, generic software components and open interfaces that form a harmonious whole. It is based on experience and evaluation of national and international information integration projects. It is an initiative of the CIDOC CRM Special Interest Group (CRM SIG), a Working Group of CIDOC, the International Committee for Documentation of the International Council of Museums (ICOM). The X3ML Toolkit allows data experts to transform their internal structured data and other associated contextual knowledge to other schemas. Fields or elements from a source database (Source Nodes) are aligned with one or more entities described in the target schema so that the data from an entire system can be transformed. The purpose of this is typically for publication on the Web and in particular meaningful integration with other data also transformed to the same target schema.

Abbreviations and Acronyms

ADP	Analyze data! Deposit study! Promote science! -Social Science Data Archives
CESSDA	Consortium of European Social Science Data Archives
CIDOC	International Documentation Committee
CIDOC CRM	CIDOC Conceptual Reference Model
CLARIN	Common Language Resources and Technology Infrastructure
CMDI	Component Metadata Infrastructure
CRMsci	Scientific Observation Model
CD	Dublin Core
DDI	Document Discover Interoperate
EMM Survey Registry	Ethnic and Migrant Minority Survey Registry
E-RIHS	European Research Infrastructure for Heritage Science
EOSC	European Open Science Cloud
FORTH	Foundation for Research and Technology -Hellas
FSD	Finnish Social Science Data Archive
ICPSR	Inter-university Consortium for Political and Social Science
LINDAT/CLARIAH-CZ	Digital Research Infrastructure for Language Technologies, Arts and Humanities
NESSTAR/SoDaNet	Networked European Social Science Tools and Resources/Social Data Network
PARTHENOS	Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies
RDF/S	Resource Description Framework Schema
SND	Swedish National Data Service
SO	Scholarly Ontology
SSH	Social Sciences and Humanities
SSHOC	Social Sciences and Humanities Open Cloud
SSHOCro	SSHOC Reference Ontology
URI	Uniform Resource Identifier

Table of Contents

1. Introduction

The purpose of this deliverable is to show to the social scientists and humanities researchers how to relate and combine their data documented with the DDI and CMDI (source schema) with the use of SSHOCro (target schema) with minimal loss of meaning and perspective of the data.

The SSHOC Reference Ontology (SSHOCro) proposes an ontological model and RDF Schema to be used as a top-level ontology for organizing knowledge and information found distributed across various primary sources of information in the Social Sciences and Humanities Open Cloud (SSHOC). It aspires to provide a semantic interoperability framework for the description of the SSHOC data lifecycle, by offering a conceptual model that can be used in order to describe at a generic level the real-world lifecycle of data produced by the generic workflow of processes of collection, connection, interpretation and the auxiliary activities of storing, publishing and finding data –as it actually takes place in the various domains of Social Sciences and Humanities. In practical terms, the use of such a model and schema for the research community is twofold: it can be applied as a standard to be used in the step of devising and implementing metadata capture scheme for tracking the data lifecycle in individual projects, institutions and disciplines; it can also be used to map, transform and integrate existing data across projects, institutions and disciplines into interoperable pools of information for reuse and exploitation. Within this frame, SSHOCro proposes an ontological model that tries to capture the tools and services used by research communities across the Social Sciences and Humanities disciplines at each point in the data lifecycle, the kind of data they generate/capture, how and by whom are the ensuing data maintained, used, published and archived and under what conditions. In that sense, SSHOCro assumes the event-centric approach of digital provenance models, which allows tracing the intermediate results (data) of the processes involved in the research workflows in SSH. In this context, keeping track of the processes involved in the data lifecycle amounts to associating each stage with a set of activities performed in it.

To this end, an (automated) transformation of each instance (specific SSH research case) documented with DDI or CMD into an instance documented with SSHOCro has been defined. The approach undertaken includes the following stages: (i) interpretation of source schema as semantic model,[1] (ii) mapping each element of that to an equivalent path in target schema, such that (iii) each instance of an element of the semantic source schema can be converted into a valid instance of the target schema (SSHOCro) with the same meaning. Sometimes during the stage (i) & (ii) in order to produce valid equivalent paths to a target schema, it is required to expand the target schema, this process is called harmonization of two semantic schemas.

This procedure is called *Mapping* and the purpose of performing it is twofold; on the one hand, it forms an intellectual definition of the relation between the models. On the other, the mapping stands as an intermediate format that can be used to more or less automatically transform data from the one form to the other –i.e., they can be used to implement an automatic data translation.

The mapping can provide a common layer from which to access several ontologies and exchange information in a semantically sound manner.[2]

The intellectual definitions of the relation between the models are self-explanatory SSHOCro-compatible propositions. Thus, they can be merged into huge knowledge pools and the document boundaries can be ignored.

The work done in this task and presented in this report includes the following steps:

1. selection of the specifications of versions and profiles for DDI and CMDI (see section: Selection process-2.1.1& 2.2.1)
2. selection of the tags to be mapped into SSHOCro (see section: Selection process-2.1.2 & 2.2.2)
3. updating of SSHOCro, following the harmonization process between DDI and CMDI on the one hand, and SSHOCro on the other, (see section "SSHOCro Harmonization")
4. conceptual mapping process: establishing correspondences between selected elements of the DDI and elements/full paths of SSHOCro (see section: DDI Codebook to SSHOCro).
5. conceptual mapping process: establishing correspondences between selected elements of CMDI and SSHOCro (see section: CMDI to SSHOCro)
6. implementation of the mappings between selected research datasets documented with CMDI/DDI and SSHOCro (see Appendix)

2. Selection process

This section offers documentation on (i) the version of DDI Codebook and CMDI schema that served as the source for the mappings to SSHOCro, (ii) the available profiles that were taken into consideration upon deciding the elements of each schema that were mapped to SSHOCro, and (iii) the repositories that the actual metadata instances that were used for the mapping came from.

2.1 DDI Codebook (2.0 & 2.5)

The Data Documentation Initiative (also known as DDI)[3] is an international standard for describing surveys, questionnaires, statistical data files, and social sciences study-level information. This information is described as metadata by the standard. It is, in essence, a way of formatting the documentation for a social science data file. It is used to document datasets and it allows search in many social science data archives around the world. DDI focuses on both microdata and aggregated data. It has its strength in microdata –data on the

[1] It is a conceptual data model in which semantic information is included. This means that the model describes the meaning of its instances. Such a semantic data model is an abstraction that defines how the stored symbols (the instance data) relate to the real world.

[2] For an example illustrating the mapping procedure briefly mentioned above, see: Bekiari et al. FRBR object-oriented; Definition and Mapping from FRBRer, FRAD and FRSAD; pp.105-146.

[3] DDI documentation: https://ddialliance.org/explore-documentation; [10.06.2020]

characteristics of units of a population, such as individuals or households, collected by a census or a survey for example.[4]

The DDI has emerged as a de facto standard for documenting data in the social and behavioural sciences across the full life cycle. DDI actually has two specifications; the original DDI (versions 1.0-2.5, now called DDI-Codebook) and the life-cycle based DDI (versions 3.0-3.3, now called DDI-Lifecycle). Both are able to describe microdata sets, with a wealth of related metadata (questions, variables, concepts, categories, codes, etc.).[5]

2.1.1 Schema

The selected schema for this task is the DDI Codebook 2.5 XML schema. The documentation, amendments for the changes implemented from the previous version, migration guidelines are available through the DDI Alliance website.[6] DDI Codebook 2.5 is backwards compatible with previous versions.

DDI Alliance offers links to specifications of metadata profiles for DDI Codebook, such as the CESSDA Mandatory and Recommended Elements for DDI-Codebook[7] and the DDI-Lite.[8] Though both are outdated in many ways, their core metadata elements are, in fact, used by many archives.

2.1.2 Instances: DDI Codebook: XML-records selected as use cases for the mappings

DDI Codebook XML records documenting social science research can be accessed from a large number of online resources. The records studied for the purposes of the current deliverable, all come from the Finnish Social Science Data Archive (FSD) Data Catalogue, DataverseNo, the Ethnic Minority Migration (EMM) Survey Registry, the Inter-university Consortium for Political and Social Research (ICPSR) and the Swedish National Data Service (SND) Research Data Catalogue.

[4] Martinez, *The Data Documentation Initiative (DDI) and Institutional Repositories*, 3.

[5] Arofan, *The Data Documentation Initiative (DDI): An Introduction for National Statistical Institutes*, 2.

[6] DDI Codebook 2.5, available through the DDI alliance website: https://ddialliance.org/Specification/DDI-Codebook/2.5/; [10.06.2020]

[7] CESSDA Mandatory and Recommended Elements for DDI Codebook: https://www.cessda.eu/content/download/709/6350/file/CESSDA [10.06.2020]

[8] DDI-Lite elements: https://ddialliance.org/sites/default/files/ddi-lite.html [10.06.2020]

Finish Social Science Data Archive (FSD)[9]

FSD is a national service resource for scientific research and teaching. It archives, promotes and disseminates digital research data for research, teaching and learning purposes. Aila Data Service is the FSD's online data service which includes an online data catalogue that contains study descriptions of archived data in English and Finnish. The service offers more than 1,500 datasets -the vast majority of which (app. 1,300) are quantitative.

DataverseNo[10]

DataverseNo is a national, generic repository for open research data from researchers from Norwegian research institutions.

Ethnic Minority Migration (EMM) Survey Registry [11]

The Ethnic and Migrant Minority (EMM) Survey Registry is a database of quantitative surveys that have been undertaken with EMM (sub)samples across Europe and beyond. Survey-level metadata is available for each of the surveys included in the EMM Survey Registry. It includes EMM-specific surveys and general population surveys with a substantive EMM (sub)sample.

Inter-university Consortium for Political and Social Science (ICPSR)[12]

ICPSR maintains a data archive of more than 250,000 files of research in the social and behavioural sciences. It hosts 21 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields.

Swedish National Data Service (SND) Research Data Catalogue[13]

The SND has a primary function to support the accessibility, preservation, and re-use of research data and related materials. SND's national research data catalogue lists app. 2000 studies spanning a wide range of subject areas. It describes data that are accessible in the catalogue itself, as well as data accessible via another portal or actor.

Records from the FSD are encoded in DDI Codebook 2.0, whereas records form the other repositories in DDI Codebook 2.5. The update to version 2.5 does not result in interoperability problems between versions.

[9] Aila Data Service: https://services.fsd.uta.fi/catalogue/index?lang=en&study_language=en [10.03.2020]

[10] DataverseNo: https://dataverse.no/ [17.03.2020]

[11] The Ethnic and Migrant Minority (EMM) Survey Registry is a free online tool that allows users to search for and learn about existing quantitative surveys to EMM populations through the compiled survey-level metadata. For more information: https://ethmigsurveydatahub.eu/emmregistry/; [30.05.2020]

[12] ICPSR: https://www.icpsr.umich.edu/web/pages/index.html [10.10.2020]

[13] SND Research Data Catalogue (SND): https://snd.gu.se/en/search/content; [08.10.2020]

Most of the records consulted are about studies dated after 2010 –and none predates 2000.

To ensure that all major elements of a DDI file were represented in the sampled records, there was a bias towards selecting records that document quantitative studies, the data of which were publicly available – meaning that there were metadata describing the data in terms of variables as well.

The criterion for selecting these repositories over other free online resources was guided by **consistency constraints** –such as the need of uniformity in terms of the DDI elements that would ultimately be mapped to SSHOCro; by **formatting constraints** –such as the need for the metadata mapped to SSHOCro to appear in DDI-XML records and not as .html pages or as .json files; by **linguistic constraints** –i.e., the metadata values should appear in English.

The criteria mentioned above resulted in excluding repositories such as the UK Data Service[14] and NESSTAR/SoDaNet[15] (.html page), CESSDA Data Catalogue[16] (exports in. json), and the ADP Data Catalogue[17] (metadata values in Slovene).

2.1.3 DDI Codebook tags selected for the mapping

The DDI components/elements mapped to SSHOCro form a concatenation of **DDI-Lite** and **CESSDA Mandatory and Recommended Elements for DDI-Codebook**, that appear in the sample of records examined. Some **elements appearing in the metadata descriptions of studies examined that are not listed by any profile** are **added** to these. In their turn, these elements fall under two broad categories:

(a) **non-core elements** consistently used in records by FSD/ EMM Survey Registry/SND and thus form some sort of guideline by the repository for data providers to follow, or
(b) **new tags** that were included in the DDI Codebook 2.5 specification –which is predated by both profiles mentioned above. The DDI tags falling under this particular category are embedded under complex elements <**sampleFrame**> and <**targetSampleSize**>, both consistently appearing in records from EMM Survey Registry.

In general, the decision to include or exclude an element from the mapping to SSHOCro process was affected by whether they consistently appeared in the examined records.

In what follows, the selected elements are presented in their context, i.e., the component they are expected to appear in. This practice results in certain elements appearing multiple times (for instance <**citation**>).

Aside from the DDI elements listed below, their attributes have been mapped to SSHOCro as well.

[14] UK Data Catalogue: https://beta.ukdataservice.ac.uk/datacatalogue/studies/; [29.04.2020]

[15] NESSTAR/SoDaNet: http://nesstar-server.sodanet.gr/webview/; [28.04.2020]

[16] CESSDA Data Catalogue: https://datacatalogue.cessda.eu/; [28.04.2020]

[17] ADP Catalogue: https://www.adp.fdv.uni-lj.si/opisi/; [30.04.2020]

The elements marked in blue fonts represent additional, non-core, elements appearing in the DDI-xml records from FSD, EMM Survey Registry, SND and ICPSR.

codeBook

docDscr {citation {titlStmt {titl, parTitl, altTItl, IDNo}, prodStmt {producer, copyright, prodDate, prodPlace}, biblCit, holdings}, notes}

stdyDscr {citation {titlStmt {titl, parTitl, altTItl, IDNo}, rspStmt {AuthEnty, othID}, prodStmt {producer, prodDate, copyright}, distStmt {distrbtr, **distDate**, depositr, depDate}, serStmt {serName, serInfo}, verStmt {version}, biblCit, holdings}, stdyInfo {subject {keyword, topClas, abstract}, sumDscr {timePrd, collDate, nation, universe, dataKind}}, method {dataColl {timeMeth, dataCollector, sampProc, collMode, collSitu, resInstru, **sources**, weight, cleanOps, stdyClas, sampleFrame {txt, universe, unitType}, targetSampleSize {targetSize}}, anlyInfo {respRate}, stdyClas}}, dataAccs {setAvail {avlStatus, accsPlac, **collSize**}, useStmt {restrctns, citeReq, deposReq, disclaimer}}, othrStdyMat {relMat, relStdy, relPubl, othRefs}}

fileDscr {fileTxt {fileName, dimensns {caseQnty, varQnty}, fileType}}

dataDscr {var {labl, qstn {qstnLit}, valrng {range}, sumStat, catgry {catValu, labl, catStat,}}}

2.2 CMDI

The Component Metadata Infrastructure[18] (CMDI) was developed within the context of the European project CLARIN[19] infrastructure with input from other initiatives and experts. The scope was to enable the flexible construction of interoperable metadata schemas suitable for, but not limited to, describing language resources. The metadata schemas based on these principles can be used to describe resources at different levels of granularity (e.g., descriptions on the collection level or on the level of individual resources). In ISO 24622-1:2015 the component metadata model has been standardized [ISO 24622-1 (2015) and ISO 24622-2 (2019].

This component-based approach or Component Metadata Infrastructure (CMDI; Broeder et al 2009) is based on well-defined formal schemas and explicit semantics by using registries.

A CMDI component registry is the service where a CMD specification can be registered and accessed. It is the central place of metadata components and profiles for purposes of reuse and sharing. The registry contains all the component metadata (CMD) instances which are collected and made available via central catalogues. All the required components are combined into one profile specific for the type of resources. It contains around 1000 components and 200 profiles. Modellers can browse and search a registry for components and profiles that are suitable or come close to meeting their requirements. Existing component registries, e.g., the CLARIN

[18] CMDI: http://www.clarin.eu/cmdi [06.05.2020]

[19] CLARIN: https://www.clarin.eu/; [06.05.2020]

(common language resources and technology infrastructure) share and distribute their metadata in CMDI format.

2.2.1. Schema

CMDI relies on a model of reusable building blocks, the components, which group together metadata elements that can potentially be reused in a different context. Each metadata record is expressed as an XML file, including a link to the profile on which it is based.

The form of a profile depends on what is deemed the proper description given (i) a language resource type, (ii) the context of the project it stemmed from or (iii) the repository that it will be stored in.

Consequently, CMDI offers a large set of concepts/elements that serve to create profiles, with potentially very little overlap among them. However, this can result in metadata descriptions that are incomparable in terms of content, as well as a proliferation of profiles (Broeder et al. 2012).

Only partial mappings from CMDI to SSHOCro are performed, which take into account a set of CMDI records adhering to the clarin.eu:cr1:p_1403526079380 profile/schema, i.e., the one used for LINDAT/CLARIN repository resources.[20] In its turn, this profile encapsulates a core set of metadata components that will serve as the basis for the DDI-CMDI conversion –one of the actions to be undertaken by SSHOC Task 3.5 *Data and Metadata Interoperability Hub*. If the minimal metadata captured by the said profile are judged adequate for the DDI-CMDI conversion, they could also be used for mappings with the SSHOCro.

2.2.2 Instances of CMDI: XML-Records selected as use cases for the mappings

Use cases studied for the purpose of this deliverable have all been retrieved from LINDAT/CLARIAH-CZ. LINDAT/CLARIAH-CZ is an open digital repository/library for **linguistic data and tools** (defined as the primary research outputs) that the users can **deposit**, **find**, **store** and **cite**. It is a shared distributed infrastructure that aims at making language resources, technology and expertise available to the humanities and social sciences research communities.

The ultimate objective of CLARIN ERIC (which LINDAT/CLARIAH-CZ is part of) is to advance research in humanities and social sciences by giving researchers unified single sign-on access to a platform which integrates language-based resources and advanced tools at a European level.

[20] The definition of the LINDAT_CLARIN profile/schema can be found on the CMDI component registry: https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu%3Acr1%3Ap_1403526079380&registrySpace=public; [06.05.2020]

2.2.3 CMDI tags selected for the mappings

The CMDI components describing the resource that was mapped to SSHOCro derive from the LINDAT_CLARIN profile,[21] i.e., the ones adhering to the clarin.eu:cr1:p_1403526079380 profile/schema.[22]

Only the basic components –namely cmd:bibliographicInfo, cmd:dataInfo, cmd:licenseInfo and cmd:relationsInfo (if existing) –plus their immediate constituents are presented. There are subtle differences depending on the type of the resource being described. Not all the fields are present in the instances that were used for the mapping. The mapping to SSHOCro will take into account the more deeply embedded components/elements as well.

cmd:LINDAT_CLARIN

cmd:bibliographicInfo {cmd:projectUrl, cmd:version, cmd:titles, cms:authors, cmd:dates, cmd:identifiers, cmd:funds, cmd:contactPerson, cmd:publishers}
dataInfo {cmd:type, cmd:detailedType, cmd:description, cmd:languages, cmd:keywords, cmd:links, cmd:sizeInfo, cmd:formats, cmd:requirements, cmd:genres, cmd:annotationInfo}

cmd:licenseInfo {cmd:license}

cmd:relationsInfo cmd:relations}

3. SSHOCro harmonization

SSHOCro proposes an ontological model and RDF schema to be used as a top-level ontology for organizing knowledge and information found distributed across various primary sources of information in the Social Sciences and Humanities Open Cloud (SSHOC). This document describes the use of SSHOCro in order to map, transform and integrate existing data into interoperable pools of information for reuse and exploitation.

SSHOCro is modelled as an extension of CIDOC CRM, the ISO standard ontology for Cultural Heritage data, from which it inherits its event-centric orientation. CIDOC-CRM provides a common and extensible semantic framework that any procedural information can be mapped to.

[21] LINDAT/CLARIAH-CZ repository https://lindat.mff.cuni.cz/repository/; [06.05.2020]
LINDAT/CLARIAH-CZ Repository Home About metadata:
https://lindat.mff.cuni.cz/repository/xmlui/page/metadata; [06.05.2020]

[22]CMDI Component Registry
https://catalog.clarin.eu/ds/ComponentRegistry/#/?registrySpace=published&itemId=clarin.eu:cr1:p_1403526079380; [06.05.2020]

Due to the harmonization process, mentioned in the introduction, the SSHOCro ontology was updated, with additions, corrections and general improvements. The current mappings are based on **SSHOCro v1.1.3** –an update from SSHOCro v1.0 that can be found on Zenodo (Bekiari, et al. 2020).

The document for **SSHOCro v1.1.3** and the corresponding rdf (SSHOCro_version1_1_3.rdf), have been updated to include the following changes:

1) **Identifiers** have been added to the labels of SSHOCro classes and properties. SSHOCro classes and properties linking them to one another have been given both a name and an identifier following the conventions:
   - Class labels are preceded by the letters **SHE** and are named using noun phrases (nominal groups) using title case (initial capitals) and identified by numbers. For instance: **SHE3 Project Activity**.
   - Property labels are preceded by the letters **SHR** and identified by numbers.
2) Addition of a new property **SHR25 provided by (provided)** [D: SHE4 Service; R: E39 Actor]
3) Update of the existing property **SHR21 incorporates (is incorporated in)**: changed the classification of the property as a subproperty of **P165 incorporates (is incorporated in)** instead of **P106 is composed of (forms part of)** that was before.
4) Addition of the new property **SHR26 has version (is version of)** [D: SHE1 Dataset; R: SHE1 Dataset] and declared as subproperty of **SHR20 has derivative (is derivative of)**.
5) Simplification of the labels of the properties SHR2 consists of data preparation (data preparation part of) and **SHR3 consists of data interpretation (data interpretation part of)**, that changed to **SHR2 consists of** and **SHR3 consists of**, respectively.
6) Simplification of the labels of the properties **SHR14 collection follows (is followed by collection)**, **SHR15 preparation follows (is followed by preparation)** and **SHR16 interpretation follows (is followed by interpretation)** that changed into **SHR14 follows (is followed by), SHR15 follows (is followed by)** and **SHR16 follows (is followed by)**, respectively.

The last updated version of SSHOCro v.1.1.3 can be found in the supplementary zipped file, attached in this report (see Linked Files, below).

4. DDI Codebook to SSHOCro

This chapter defines the mapping between the DDI Codebook model[23] and SSHOCro.

The idea is to establish mappings of the entities/relations/attributes defined in DDI Codebook (v2.0 and 2.5) to an equivalent path in SSHOCro. These are self-explanatory SSHOCro-compatible propositions consisting of long paths of classes and properties of SSHOCro.

Following the method described in the introduction of interpretation of the source schema as semantic model, the mapping of a particular DDI element to SSHOCro would occur as many times as necessary, given its possible positions in the DDI document. Thus, DDI elements are mapped to more than one path in SSHOCro –depending on their position in the DDI document or on the presence of attributes that further specify them.

For instance, elements like <**universe**>, <**txt**>, <**notes**> all correspond to free-text descriptions ultimately rendered by CIDOC-CRM paths E73 Information Object -P3 has note: E62 String. However, the fact that they can appear in multiple parts of the DDI document adds to the semantics of each of their instances, a fact which is captured by longer SSHOCro paths in their mappings.

On the other hand, DDI complex element <**citation**>, which encodes bibliographic information at the appropriate level (for instance, the bibliographic information concerning a DDI metadata record, a particular study/survey, publications in the context of a survey, datasets and studies influenced from the documented dataset, projects that used it etc.), despite varying to some extent regarding the detail in which its component elements are elaborated upon, assumes a more-or-less constant structure which is rendered by the same SSHOCro classes and properties, irrespective of its position in the DDI document. What changes on each iteration is the set of elements that need to be documented, according to the DDI guidelines.

In what concerns attributes, they too get translated in the SSHOCro style, as paths consisting of classes linked by properties. For instance, there exists a certain homogeneity, in terms of where the path rendering **URI** attributes should end –namely E1 CRM Entity –**P1 is identified by: E41 Appellation -P2 has type: E55 Type** {URI}. However, the semantics of the element the attribute specifies (and its position in the DDI document) have an impact on the class instantiating the domain of the property, as well as on its overall position in the path.

Thus, for <**holdings**> and <**accPlac**>, the full path is **SHE1 Dataset –P1 is identified by: E41 Appellation - P2 has type: E55 Type** {URI}, whereas for <**distrbtr**> the full path is rendered by **SHE1 Dataset -SHR19i was stored by: SHE9 Data Storage - P14 carried out by: E39 Actor -P76 has contact point: E51 Contact Point -P2 has type: E55 Typ**e {URI} and for <**serStmt**> it is rendered by **SHE1 Dataset -P106i forms part of: E73 Information Object -P1 is identified by: E42 Identifier -P2 has type: E55 Type** {URI}.

[23] DDI Codebook 2.5: https://ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html [10.06.2020]

4.1 DDI Codebook to SSHOCro

Mappings from DDI elements to the respective SSHOCro classes and properties are listed below. Given that the meaning of the DDI elements is determined to some extent by their position in the DDI document, and taking into account their large number, the mappings are grouped as follows:

(1) DDI elements listed under **Document Description**, representing the **bibliographical metadata** for the DDI document –i.e., the content elements of <**citation**>.

(2) DDI elements listed under **Study Description** have been split into three subsections. Study Description constitutes the bulk of the DDI record. The subsections are grouped as follows:
   a. One section designated to representing the **bibliographical information** for a given study, i.e., the content elements of <**citation**>, and the **scope** of the study, i.e., the content elements of <**stdyInfo**>
   b. One section designated to representing the **methods** followed in a research/survey/data collection, i.e., the content elements of <**method**>.
   c. One section designated to representing the **conditions** and **terms of use** for a given collection, i.e., <**dataAccs**>, and **other material relevant for the study description**, i.e., <**OthStdyMat**>.

(3) DDI elements listed under **File Description** documenting any accompanying files for a given study.

(4) DDI elements listed under **Variable Description**, where each variable gets documented

(5) **DC** elements used to represent **bibliographical information** in parts of the DDI document –other than the Study Description.

DDI elements listed under **other materials related to the study** (<**OthMat**>), where links to related reports and publications are supplied, were not mapped to SSHOCro, because they are not corroborated by the examined records. Where necessary, the relevant information is provided at the level of Study Description, under <**othStdyMat**>.

In what follows, the SSHOCro classes or paths corresponding to the DDI elements are marked in **boldface**. It is possible that a path rendering DDI element/attribute **a**, extends a path that has been already declared in the mapping process to express DDI element **b**. In that particular case, the full path for **a.** will be given, i.e., the one including **b.**, but only the part exclusively referring to **a.** will be marked in **boldface**. For instance, DDI element <**IDNo>** and its attribute <IDNo.**agency**> are respectively expressed in SSHOCro as follows:

- <**IDNo>** -> **SHE1 Dataset** (–P1 is identified by: E42 Identifier)
- <IDNo.**agency**> - > SHE1 Dataset –P1 is identified by: E42 Identifier **–P140i was attributed by: E15 Identifier Assignment -P14 carried out by: E39 Actor**

Marked in {curly brackets} are the types assigned to the SSHOCro classes to that are to be taken into account by the transformation algorithm. For instance, a parallel title and an alternative title both resolve to an instance of **E35 Title** in the CIDOC-CRM universe: E35 Title **-P139 has alternative form: P35 Title**. To establish that the one is a translation whereas the other is a name commonly used to refer to a given work (but not its official title), one needs to add the appropriate type by **-P2 has type: E55 Type –**be it {Translation} or {Other Appellation}. The fact that these specifications remain constant and have to be assigned ad hoc, is captured by the information in the {curly brackets}.

4.1.1 Document Description –Metadata for the DDI document

The document description forms the metadata of the DDI document and lists the sources used in its creation. As stated in the definition, "It consists of bibliographic information describing the DDI-compliant documents itself as a whole. [It] can be considered the wrapper or header whose elements uniquely describe the full contents of the compliant DDI file".

Bibliographic information includes title information, statement of responsibility, production and distribution information, series and version information, text of a preferred bibliographic citation, and notes (if any).

LIST OF MAPPINGS –DOCUMENT DESCRIPTION

DDI Tag	Unit of Information (embedding)	Condition	SSHOCro
IDNo	codebook/ docDscr/ citation/ titlStmt/ IDNo		**SHE1 Dataset** (–P1 is identified by: E42 Identifier) and SHE1 Dataset -P67 refers to: SHE1 Dataset
IDNo **@agency**	codebook/ docDscr/ citation/ titlStmt/ IDNo @agency	@agency= " "	SHE1 Dataset –P1 is identified by: E42 Identifier **–P140i was attributed by: E15 Identifier Assignment -P14 carried out by: E39 Actor**
titl	codebook/ docDscr/ citation/ titlStmt/ titl		SHE1 Dataset **-P102 has title: E35 Title**
parTitl	codebook/ docDscr/ citation/ titlStmt/ parTitl		SHE1 Dataset - P102 has title: E35 Title **- P139 has alternative form** {P139.1 has type = "Translation"}**: P35 Title**
altTitl	codebook/ docDscr/ citation/ titlStmt/ altTitl		SHE1 Dataset - P102 has title: E35 Title **- P139 has alternative form** {P139.1 has type = "Other Appellation"}**: P35 Title**

producer	codebook/ docDscr/ citation/ prdStmt/ producer		SHE1 Dataset **–P94i created by: E65 Creation –P14 carried out by** {P14.1 in the role of: E55 Type = "Producer"}: **E39 Actor**
producer **@abbr**	codebook/ stdyDscr/ citation/ prodStmt/ producer @abbr	@abbr = ""	SHE1 Dataset –P94i created by: E65 Creation –P14 carried out by: E39 Actor-**P1 is identified by: E41 Appellation** -P2 has type: E55 Type {"Abbreviation"}
copyright	codebook/ docDscr/ citation/ prdStmt/ copyright		SHE1Dataset **–P104 is subject to: E30 Right –P2 has type: E55 Type** {"Copyright"} and **E30 Right –P3 has note: E62 String**
prodDate	codebook/ docDscr/ citation/ prdStmt/ prodDate		SHE1 Dataset **-P94i created by: E65 Creation -P4 has timespan: E52 Time-Span**
prodDate **@date**	codebook/ docDscr/ citation/ prdStmt/ prodDate @date	@date= " "	SHE1 Dataset -P94i created by: E65 Creation -P4 has timespan: E52 Time-Span -**P82 at some time within: E61 Time Primitive**
prodPlace	codebook/ docDscr/ citation/ prdStmt/ prodPlace		SHE1 Dataset **-P1 is identified by: E41 Appellation -P2 has type: E55 Type** {"Address"}
biblCit	codebook/ docDscr/ citation/ biblCit		SHE1 Dataset **-P67i is referred to by: E73 Information Object -P3 has note: E62 String**
biblCit **@format**	codebook/ docDscr/ citation/ biblCit @format	@format = ""	SHE1 Dataset -P67i is referred to by: E73 Information Object -P3 has note: E62 String **-P2 has type: E55 Type**
holdings **@location**	codebook/ docDscr/ citation/ holdings @location	@location = ""	SHE1 Dataset -**P1 is identified by: E41 Appellation** -P2 has type: E55 Type {"Location"}

holdings **@URI**	codebook/ docDscr/ citation/ holdings @URI	@URI = ""	SHE1 Dataset -**P1 is identified by: E41 Appellation [value of URI attribute]** **E41 Appellation**-P2 has type: E55 Type {"URI"}
notes	codebook/ docDsc/ notes		SHE1 Dataset -**P67i is referred to by: E33 Linguistic Object –P2has type: E55 Type** {"Comment"} and **E33 Linguistic Object -P3 has note: E62 String**
notes **@xml:lang**	codebook/ docDsc/ notes @xml:lang	@xlm:lang = ""	SHE1 Dataset - P67i is referred to by: E33 Linguistic Object **-P72 has language: E56 Language**

Note that the relation **SHE1 Dataset –P67 refers to: SHE1 Dataset** captures the relation between the metadata of the DDI document to the metadata of the study it describes.

4.1.2 Study Description

According to its definition, the Study Description consists of information about the data collection, study, or compilation that the DDI-compliant documentation file describes. This section includes information about how the study should be cited, who collected or compiled the data, who distributes the data, keywords about the content of the data, summary (abstract) of the content of the data, data collection methods and processing, etc.

4.1.2.1 Bibliographical Metadata Of The Study; Scope Of The Study

The bibliographical information of the study is conveyed by the elements headed by <**citation**> in DDI, a repeatable element appearing at multiple sections of a DDI document. Some of its content elements have already been mentioned in the context of Document Description, and their mappings are essentially the same in the Study Description too. What changes in this respect, is the detail with which these elements are described.

The Scope of the Study is conveyed by the elements headed by <**stdyInfo**>. The section contains information about the scope of the study/survey/collection's scope across several parameters –including its substantive content, spatial and temporal dimension.

LIST OF MAPPINGS –STUDY DESCRIPTION: <CITATION>; <STDYINFO>

DDI Tag	Unit of Information (embedding)	Condition	SSHOCro
IDNo	codebook/ stdyDscr/ citation/ titlStmt/ IdNo		**SHE1 Dataset** (-P1 is identified by: E42 Identifier)
IDNo **@agency**	codebook/ stdyDscr/ citation/ titlStmt/ IDNo @agency	@agency= ""	SSHE1 Dataset –P1 is identified by: E42 Identifier **–P140i was attributed by: E15 Identifier Assignment -P14 carried out by: E39 Actor**
titl	codebook/ stdyDscr/ citation/ titlStmt/titl		SHE1 Dataset **- P102 has title: E35 Title**
parTitl	codebook/ stdyDscr/ citation/ titlStmt/ parTitl		SHE1 Dataset - P102 has title: E35 Title **- P139 has alternative form** {P139.1 has type = “Translation”}**: P35 Title**
altTitl	codebook/ stdyDscr/ citation/ titlStmt/ altTitl		SHE1 Dataset - P102 has title: E35 Title **- P139 has alternative form** {P139.1 has type = “Other Appellation”}**: P35 Title**
AuthEnty	codebook/ stdyDscr/ citation/ rspStmt/ AuthEnty		SHE1 Dataset **-94i was created by: E65 Creation -P14 carried out by** {P14.1 in the role of: E55 Type = “Primary Investigator”}**: E39 Actor**
OthId	codebook/ stdyDscr/ citation/ rspStmt/ OthID		SHE1 Dataset **-94i was created by: E65 Creation -P14 carried out by** {P14.1 in the role of: E55 Type = “Contributor”}**: E39 Actor**
OthId **@affiliation**	codebook/ stdyDscr/ citation/ rspStmt/ OthID @affiliation	@affiliation = ""	SHE1 Dataset -94i was created by: E65 Creation -P14 carried out by: E39 Actor **- P107i is current or former member of: E74 Group**
producer	codebook/ stdyDscr/ citation/ prodStmt/ producer		SHE1 Dataset **-SHR18i was published by: SHE8 Publication –P14 carried out by** {P14.1 in the role of: E55 Type = “Publisher”}**: E39 Actor**

copyright	codebook/ stdyDscr/ citation/ prodStmt/ copyright		SHE1 Dataset **-P104 is subject to right: E30 Right –P2 has type: E55 Type** {“Copyright”} and **E30 Right-P3 has note: E2 String**
distrbtr	codebook/ stdyDscr/ citation/ distSmt/ distrbtr		SHE1 Dataset -**SHR18i was published by: SHE8 Publication-P14 carried out by** {P14.1 in the role of: E55 Type = “Distributor”}**: E39 Actor -P1 is identified by: E41 Appellation**
distrbtr **@abbr**	codebook/ stdyDscr/ citation/ distSmt/ distrbtr @abbr	@abbr = ""	SHE1 Dataset -SHR18i was published by: SHE8 Publication -P14 carried out by: E39 Actor **-P1 is identified by: E41 Appellation** -P2 has type: E55 Type {“Abbreviation”}
distrbtr **@URI**	codebook/ stdyDscr/ citation/ distSmt/ distrbtr @URI	@URI = ""	SHE1 Dataset -SHR18i was published by: SHE8 Publication- P14 carried out by: E39 Actor **-P76 has contact point: E51 Contact Point -P2 has type: E55 Type** {“URI”}
distDate	codebook/ stdyDscr/ citation/ distSmt/ depDate		SHE1 Dataset **-SHR18i was published by: SHE8 Publication -P4 has time-span: E52 Time-Span –P82 at some time within: E61 Time Primitive**
distDate **@date**	codebook/ stdyDscr/ citation/ distSmt/ depDate @date	@date = ""	SHE1 Dataset -SHR18i was published by: SHE8 Publication-P4 has time-span: E52 Time-Span **-P82 at some time within: E61 Time Primitive**
depositr	codebook/ stdyDscr/ citation/ distSmt/ depositr		SHE1 Dataset -SHR19i was stored by: SHE9 Data Storage - **P14 carried out by** {P14.1 in the role of: E55 Type = “Depositor”}**: E39 Actor**
depositr **@affiliation**	codebook/ stdyDscr/ citation/ distSmt/ depositr @affiliation	@affiliation = ""	SHE1 Dataset -SHR19i was stored by: SHE9 Data Storage - P14 carried out by: E39 Actor-**P107i is current or former member of: E74 Group**

depDate	codebook/ stdyDscr/ citation/ distSmt/ depDate		SHE1 Dataset **-SHR19i was stored by: SHE9 Data Storage-P4 has time-span: E52 Time-Span –P82 at some time within: E61 Time Primitive**
depDate **@date**	codebook/ stdyDscr/ citation/ distSmt/ depDate @date	@date = ""	SHE1 Dataset -SHR19i was stored by: SHE9 Data Storage-P4 has time-span: E52 Time-Span **-P82 at some time within: E61 Time Primitive**
serStmt	codebook/ stdyDscr/ citation/ serStmt		SHE1 Dataset **-P106i forms part of: E73 Information Object**
serStmt **@Uri**	codebook/ stdyDscr/ citation/ serStmt @URI	@URI = ""	SHE1 Dataset -P106i forms part of: E73 Information Object **-P1 is identified by: E42 Identifier** -P2 has type: E55 Type {"URI"}
serName	codebook/ stdyDscr/ citation/ serStmt/ serName		SHE1 Dataset -P106i forms part of: E73 Information Object **-P102 has title: E35 Title**
serName **@abbr**	codebook/ stdyDscr/ citation/ serStmt/ serName @abbr	@abbr = ""	SHE1 Dataset -P106i forms part of: E73 Information Object: **P2 has type: E55 Type.**
serInfo	codebook/ stdyDscr/ citation/ serStmt/ serInfo		SHE1 Dataset -P106i forms part of: E73 Information Object -P67i is referred to by: **E73 Information Object –P2 has type: E55 Type** {"Short Description"} and **E73 Information Object –P3 has note: E62 String**
version	codebook/ stdyDscr/ citation/ verStmt/ version		SHE1 Dataset **-P1 is identified by: E42 Identifier** -P2 has type: E55 Type {"Version"}
version **@date**	codebook/ stdyDscr/ citation/ verStmt/ version @date	@date = ""	SHE1 Dataset -P1 is identified by: E42 Identifier**-P140 was attributed by: E15 Identifier Assignment -P4 has timespan: E52 Time-Span -P82 at some time within: E61 Time Primitive**

biblCit	codebook/ stdyDscr/ citation/ biblCit		SHE1 Dataset **-P67i is referred to by: E73 Information Object -P3 has note: E62 String**
keyword	codebook/ stdyDscr/ stdyInfo/ subject/ keyword		SHE1 Dataset --P129 is about: **E73 Information Object** -P2 has type: E55 Type {“Keyword”}
keyword **@vocab**	codebook/ stdyDscr/ stdyInfo/ subject/ keyword @vocab	@vocab= ""	SHE1 Dataset --P129 is about: E73 Information Object **-P106i forms part of: SHE1 Dataset** -P2 has type: E55 Type {“Vocabulary”}
topClas	codebook/ stdyDscr/ stdyInfo/ subject/ topClas		SHE1 Dataset --P129 is about: **E73 Information Object** -P2 has type: E55 Type {“Topic classification”}
topClas **@vocab**	codebook/ stdyDscr/ stdyInfo/ subject/ topClas @vocab	@vocab= ""	SHE1 Dataset –P129 is about: **E73 Information Object -P106i forms part of: SHE1 Dataset** -P2 has type: E55 Type {“Vocabulary”}
abstract	codebook/ stdyDscr/ stdyInfo/ abstract		SHE1 Dataset -**P67i is referred to by: E73 Information Object -P3 has note: E62 String**
timePrd	codebook/ stdyDscr/ stdyInfo/ sumDscr/ timePrd		SHE1 Dataset -**P129 is about: E4 Period -P4 has timespan: E52 Time-Span**
timePrd **@date**	codebook/ stdyDscr/ stdyInfo/ sumDscr/ timePrd @date	@date =“”, **IF** timePrd @event= **“start”**	SHE1 Dataset -P129 is about: E4 Period -P4 has timespan: E52 Time-Span – **P79 beginning is qualified by: E62 String**
timePrd **@date**	codebook/ stdyDscr/ stdyInfo/ sumDscr/ timePrd @date	@date =“”, **IF** timePrd @event= **“end”**	SHE1 Dataset -P129 is about: E4 Period -P4 has timespan: E52 Time-Span – **P80 end is qualified by: E62 String**

timePrd **@date**	codebook/ stdyDscr/ stdyInfo/ sumDscr/ timePrd @date	@date ="", **IF** timePrd @event= **"single"**	SHE1 Dataset -P129 is about: E4 Period -P4 has timespan: E52 Time-Span –**P82 at some time within: E61 Time-Primitive**
collDate	codebook/ stdyDscr/ stdyInfo/ sumDscr/ collDate		SHE1 Dataset **-P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -P4 has timespan: E52 Time-Span**
collDate **@date**	codebook/ stdyDscr/ stdyInfo/ sumDscr/ collDate @date	@date ="", **IF** collDate **@event= "start"**	SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -P4 has timespan: E52 Time-Span **-P79 beginning is qualified by: E62 String**
collDate **@date**	codebook/ stdyDscr/ stdyInfo/ sumDscr/ collDate @date	@date ="", **IF** collDate **@event= "end"**	SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -P4 has timespan: E52 Time-Span **-P80 end is qualified by: E62 String**
collDate **@date**	codebook/ stdyDscr/ stdyInfo/ sumDscr/ collDate @date	@date ="", **IF** collDate **@event= "single"**	SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -P4 has timespan: E52 Time-Span **-P82 at some time within: E61 Time Primitive**
collDate **@cycle**	codebook/ stdyDscr/ stdyInfo/ sumDscr/ collDate @cycle	@cycle = ""	SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -P4 has timespan: E52 Time-Span **-P2 has type: E55 Type**
nation	codebook/ stdyDscr/ stdyInfo/ sumDscr/ nation		SHE1 Dataset **-P129 is about: E53 Place-P1 is identified by: E41 Appellation** and E53 Place -P2 has type: E55 Type {"Country Name"}

nation **@abbr**	codebook/ stdyDscr/ stdyInfo/ sumDscr/ nation @abbr	@abbr= ""	SHE1 Dataset -P129 is about: E53 Place-P1 is identified by: E41 Appellation **-P139 has alternative form** {P139.1 has type: E55 Type = "Country Code"}**: E41 Appellation**
geogCover	codebook/ stdyDscr/ stdyInfo/ sumDscr/ geogCover		SHE1 Dataset -**P129 is about: E53 Place -P2 has type: E55 Type** {"Geographic Area"}
anlyUnit	codebook/ stdyDscr/ stdyInfo/ sumDscr/ anlyUnit		SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 data Collection **-P9 consists of: S4 Observation -O8 observed: S15 Observable Entity -P2 has type: E55 Type**
universe	codebook/ stdyDscr/ stdyInfo/ sumDscr/ universe		SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 data Collection **-P9 consists of: S4 Observation -O8 observed: S15 Observable Entity -P67i is referred to by: E73 Information Object -P3 has note: E62 String**
universe **@clusion**	codebook/ stdyDscr/ stdyInfo/ sumDscr/ universe @clusion	@clusion = ""	SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 data Collection -P9 consists of: S4 Observation -O8 observed: S15 Observable Entity -P67i is referred to by: E73 Information **Object -P2 has type: E55 Type**
dataKind	codebook/ stdyDscr/ stdyInfo/ sumDscr/ dataKind		SHE1 Dataset **-P2 has type: E55 Type**

Regarding the mapping of <timePrd>, <collDate> and their attributes @event and @date to SSHOCro:

Given their definitions, the two elements link a data collection to two distinct time intervals; <**timePrd**> corresponds to the time period covered by a data collection, whereas <**collDate**>refers to the actual time that the process of collecting one's data took place.

Therefore, SSHOCro needs to represent **time** as a research object –i.e. a time interval during which the events of interest unfolded (see <**timePrd**>). This is achieved through **P129 is about** [D: **E89 Propositional Object**, R: **E1 CRM Entity**].[24] It is a very high-level property, used to express the topic of propositional objects. The fact that its range is set to E1 CRM Entity, means that a proposition can be about practically anything.[25]

SSHOCro also needs to anchor each phase of a data-driven research workflow –typically defined in the course of a project –to the timespan during which it occurred (see <**collDate**>).

Since both elements refer to time intervals through their **@date** attributes, **collDate@date** and **timePrd@date** have both been rendered by means of using **P82a_begin_of_the_begin/ P82b_end_of_the_end**, in 3M. The range of **P82a/b** is set to **xsd:dateTime**.

The discrepancy between the proposed theoretical mapping in SSHOCro above (through **P79 beginning is qualified by**; **P80 end is qualified by**) and its implementation in RDF (through **P82a_begin_of_the_begin**; **P82b_end_of_the_end**) is due to the way the CRM envisions time –i.e. as an interval rather than an instant. Upon specifying two distinct points in time when an event/period (a) started and (b) was no longer ongoing, one has defined a time interval during which the temporal trace of the event/period at hand is (properly) included.

The top-level properties of the CIDOC CRM relating to temporal entities support the documentation dates as timespans or dimensions, mereological relations between temporal entities as well as a complete suite of topological relations.

When the dates delimiting/bounding a temporal entity are known, this can be documented by instantiating the ***P4 has time-span*** property and then by creating an instance of **E52 Time-Span**. Dates should be recorded as instances of **E61 Time Primitive** and related to the timespan through either **P81 ongoing throughout** or **P82 at some time within**.

Property **P82 at some time within** describes the narrowest known outer bounds of the temporal extent of an **E52 Time-Span**, i.e. that the described temporal phenomenon is definitely ongoing "at some time within" this interval. It is the default for historical dates, where it may be given in years for events of a much smaller duration. The actual mode of encoding the documented date lies outside the scope of the CIDOC CRM universe, which

[24] It has been declared in SSHOCro that **SHE1 Dataset** isA **E89 Propositional Object**.

[25] From the definition of CIDOC CRM, it follows that every CIDOC CRM class isA **E1 CRM Entity** (i.e., **E4 Period** isA **E1 CRM Entity** as well)

defines this with a "primitive class", i.e. **E61 Time Primitive**. It is replaced in the official RDF version by the properties "**P82a_begin_of_the_begin**" and "**P82b_end_of_the_end**", to be used together.

"**P82a_begin_of_the_begin**" should be instantiated as the latest point in time the user is sure that the respective temporal phenomenon is indeed not yet happening. It constitutes a lower limit to the beginning of the indeterminacy/fuzziness marking the beginning of the described temporal phenomenon.

"**P82b_end_of_the_end**" should be instantiated as the earliest point in time the user is sure that the respective temporal phenomenon is indeed no longer ongoing. It constitutes an upper limit to the end of the indeterminacy /fuzziness marking the end of the described temporal phenomenon.

It always holds that "**P82a_begin_of_the_begin**" is before "**P82b_end_of_the_end**".

Regarding the mapping of <universe> and <anlyUnit> to SSHOCro/ CIDOC-CRM:

Element <**universe**> relates to the overall population sampled in the course of a study, to produce the observations and the data points that formed the basis of the study. It refers to the exact population that the sample generalizes over ("universe" and "population" being co-extensional) and is rendered as a short description in the DDI.xml files.

Given that the populations that the results of various studies generalize over are extremely diverse; that they are not confined to individual persons and groups but can extend to items (such as documents, photographs)[26] and events (like armed conflicts, deaths etc.)[27] or other measurements; then this means that both the sampled entities and the populations they were sampled from (and are supposed to reflect) are represented as extremely high-level entities, suchlike they can be observed through scientific methods. Observable entity serves as a cover term for entities that exist in time (endurants) as well as entities that unfold in time (perdurants).

The analysis proposed here uses elements from the definition of *CRMsci; An extension of CIDOC-CRM that supports scientific observation* –namely classes S4 Observation, S15 Observable Entity and relation O8 observed.

[26] see Hermann, Roberto Rivas; Jensen, Are, 2020, "Replication Data for: Contingencies of circular economy; Discourse hegemony and institutionalization in Norway", https://doi.org/10.18710/CZRKNZ, DataverseNO, V1
The universe sampled at the course of this study corresponds to documents:
<universe>Publically available documents in the servers of the Norwegian government (www.regjeringen.no) and the Research Council of Norway</universe>

[27] see Kjeksrud, Stian, 2019, "Replication Data for: Using force to protect civilians", https://doi.org/10.18710/FZAVCN, DataverseNO, V1
The universe sampled at the course of this study corresponds to military conflicts (i.e. a sum of events).
<universe>United Nations military efforts to protect civilians from ten UN missions in African conflicts as reported by the United Nations Secretary-General to the United Nations Security Council from 1999 to 2017</universe>

The path documenting the relation between the universe and the dataset that was produced by studying a proper and (hopefully) representative subpart thereof is rendered as follows:

> SHE1 Dataset –P94i was created by: E65 Creation –P10 falls within: SHE5 Data Collection –P9 consists of: S4 Observation – O8 observed: S15 Observable Entity –**P67i is referred to by: E73 Information Object** - **P3 has note: E62 String (<universe>)** and **E73 Information Object** - **P2 has type: E55 Type (universe.clusion =“”)**

In a similar vein, element <**anlyUnit**> describes the basic unit of analysis or observation of a given study, i.e. the kinds of the entities that were sampled from a population of interest, which indicates that the path must, once again, go through **S4 Observation**;

> SHE1 Dataset –P94i was created by: E65 Creation –P10 falls within: SHE5 Data Collection –P9 consists of: S4 Observation – O8 observed: S15 Observable Entity –**P2 has type**: **E55 Type (<anlyUnit>)**.

Based on its definition, an instance of S4 Observation is a specification of E7 Activity and as such, it inherits all its properties –P9 consists of (forms part of) being one of them. Whether instances of S4 Observation only form part of the data collection activity (SHE5 Data Collection) or they can form part of the other stages of the research activity is not something that can be deduced by the DDI 2.0-2.5 Codebook metadata records, to the extent that they focus on documenting the data collection activity.

4.1.2.2 Methods Used In A Data Collection

The section describes the methodology and processing involved in a data collection. For the most part, the elements listed under <**method**> concern the process of data collection and provide information on the time methods observed; on the sampling procedure followed; on the sampling frame used for identifying the population from which the sample was taken; and the calculation of the desired size of the sample, given the targeted population. Incidentally, the two last pieces of information, conveyed by DDI elements <**sampleFrame**> and <**targetSampleSize**> respectively, are related to activities that precede – and guide – the actual data collection.

The emphasis placed on the process of the data collection is a direct corollary of the fact that the quality of the data depends on the actual conditions it was collected under. Furthermore, a large part of the studies documented using DDI refer to survey and census data, where the manipulation of the data –post collection – seems a little superficial –at least in the context of the survey/census.

However, there are a few elements that suggest some post-processing and some level of analysis is described by the DDI methods; For instance, elements <**cleanOps**> and <**respRate**> indicate that documenting the processes necessary to yield good data and to determine how representative of the population are one’s observations, respectively. In a similar vein, element <weight> serves as a guideline on how to manipulate the data to draw conclusions, once it has been collected and cleaned. Nonetheless, the methods are skewed in the direction of the data collection process.

List Of Mappings –Study Description: methods

DDI Tag	Unit of Information (embedding)	Condition	SSHOCro
timeMeth	codebook/ stdyDscr/ method/ dataColl/ timeMeth		SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -**SHR23 used method: SHE11 Method -P2 has type: E55 Type** {“Time Dimension”} and **SHE11 Method –P3 has note: E62 String**
timeMeth **@method**	codebook/ stdyDscr/ method/ dataColl/ timeMeth @method	@method= ""	SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -**SHR23 used method: SHE11 Method** -P2 has type: E55 Type {“Time dimension”}
data Collector	codebook/ stdyDscr/ method/ dataColl/ dataCollector		SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -**P14 carried out by** {P14.1 in the role of: E55 Type = “Data Collector”}**: E39 Actor**
dataCollector @affiliataion	codebook/ stdyDscr/ method/ dataColl/ dataCollector @affiliation	@affiliation=“”	SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -P14 carried out by {P14.1 in the role of: E55 Type = “Data Collector”}: E39 Actor **–P107i is current or former member of: E74 Group**
sampProc	codebook/ stdyDscr/ method/ dataColl/ sampProc		SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -**SHR23 used method: SHE11 Method -P67i is referred to by: E73 Information Object -P3 has note: E62 String** -P2 has type: E55 Type {“Sampling Method”}
collMode	codebook/ stdyDscr/ method/ dataColl/ collMode		SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -**SHR23 used method: SHE11 Method** -P2 has type: E55 Type {“Data Collection Method”}

collSitu	codebook/ stdyDscr/ method/ dataColl/ colSitu		SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection **-P67i is referred to by: E73 Information Object -P3 has note: E62 String**
resInstru	codebook/ stdyDscr/ method/ dataColl/ resInstru		SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection **-P125 used object of type: E55 Type**
sources	codebook/ stdyDscr/ method/ dataColl/ sources		SHE1 Dataset –**P130 shows features of: E73 Information Object**
srcDocu	codebook/ stdyDscr/ method/ dataColl/ sources/srcDocu		SHE1 Dataset –P130 shows features of: E73 Information Object –**P3 has note: E62 String**
txt	codebook/ stdyDscr/ method/ dataColl/ sampleFrame/ txt		SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -P9 consists of: S4 Observation - **P15 was influenced by: E7 Activity –P2 has type: E55 Type** {"Sampling"} and **E7 Activity -P3 has note: E62 String**
universe	codebook/ stdyDscr/ method/ dataColl/ sampleFrame/ universe		SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -P9 consists of: S4 Observation - P15 was influenced by: E7 Activity -**P9 consists of: S21 Measurement -O24 measured: S15 Observable Entity -P67i is referred to by: E73 Information Object –P2 has type: E55 Type** {"Target Population"} and **E73 Information Object -P3 has note: E62 String**
universe **@clusion**	codebook/ stdyDscr/ method/ dataColl/ sampleFrame/ universe @clusion	@clusion=""	SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -P9 consists of: S4 Observation - P15 was influenced by: E7 Activity -P9 consists of: S21 Measurement -O24 measured: S15 Observable Entity -P67i is referred to by: E73 Information Object **-P2 has type: E55 Type**

unitType	codebook/ stdyDscr/ method/ dataColl/ sampleFrame/ frameUnit/ unitType @numberOfUnits		SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -P9 consists of: S4 Observation -P15 was influenced by: E7 Activity -P9 consists of: S21 Measurement -O24 measured: S15 Observable Entity **-O12 has dimension: E54 Dimension -P91 has unit: E58 Measurement Unit**
unitType **@numberOfUnits**	codebook/ stdyDscr/ method/ dataColl/ sampleFrame/ frameUnit/ unitType	@numberOfUnits=""	SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -P9 consists of: S4 Observation -P15 was influenced by: E7 Activity **-P9 consists of: S21 Measurement -O24 measured: S15 Observable Entity -O12 has dimension: E54 Dimension -P90 has value: E60 Number**
sampleSize	codebook/ stdyDscr/ method/ dataColl/ targetSampleSize		SHE1 Dataset -P94i was created by: E65 Creation -P10 falls within: SHE5 Data Collection -P9 consists of: S4 Observation -P15 was influenced by: E7 Activity -P9 consists of: S21 Measurement -O24 measured: S15 Observable Entity **–P2 has type: E55 Type** {"Target Sample"} and S15 Observable Entity **–O12 has dimension: E54 Dimension -P90 has value: E60 Number**
respRate	codebook/ stdyDscr/ method/ anlyInfo/ respRate		SHE1 Dataset **-SHR11i was connected by: SHE6 Data Prep/Con**[28] **–SHR23 used method: SHE11 Method –P2 has type: E55 Type** {"Calculation"} and **SHE11 Method -P67i is referred to by: E73 Information Object -P3 has note: E62 String**
cleanOps	codebook/ stdyDscr/ method/ dataColl/cleanOps		SHE1 Dataset **-SHR11i was connected by: SHE6 Data Prep/Conn -SHR23 used method: SHE11 Method**

[28] SHE6 Data Prep/Con is a shorthand for SSHOCro class SHE6 Data Preparation Connection

weight	codebook/ stdyDscr/ method/ dataColl/ weight		SHE1 Dataset **–SHR12i was interpreted by**: **SHE7 Data Interpretation -SHR23 used method: SHE11 Method -P2 has type: E55 Type** {"Weighting"}
stdyClas	codebook/ stdyDscr/ method/ stdyClas		SHE1 Dataset -**P67i is referred to by: E73 Information Object -P3 has note: E62 String**
stdyClas	codebook/ stdyDscr/ method/ stdyClas	@type= ""	SHE1 Dataset -P67i is referred to by: E73 Information Object **-P2 has type: E55 Type**

4.1.2.3 Terms Of Use; Related Material

This subsection concerns the access conditions and terms of use for a data collection and documenting other materials that are related to the study description in general, represented by composite elements <**dataAccs**> and <**othStdyMat**>, respectively.

Of the content elements of <**dataAccs**> listed below that are mapped to SSHOCro, only <**restrctn**> forms part of the DDI-Lite profile. <**accsPlac**>, <**avlStatus**> form part of the CESSDA DDI Recommended Elements. The remaining elements are consistently found in DDI records obtained from FSD.

In what concerns the contents of <**othrStdyMat**> listed below, they are not recommended elements by any profile, but simply occur quite regularly in DDI records obtained from FSD and EMM Survey Registry.

List Of Mappings –Study Description: <dataAccs>; <othStdyMat>

DDI Tag	Unit of Information (embedding)	Condition	SSHOCro
avlStatus	codebook/ stdyDscr/ dataAccs/ setAvail/ avlStatus		SHE1 Dataset -**P67i is referred to by: E73 Information Object –P2 has type: E55 Type** {"Data Availability"} and **E73 Information Object -P3 has note: E62 String**

accsPlac **@URI**	codebook/ stdyDscr/ dataAccs/ setAvail/ accPlac @URI	@URI = ""	SHE1 Dataset **-P1 is identified by: E41 Appellation** -P2 has type: E55 Type {“URI”}
accsPlac	codebook/ stdyDscr/ dataAccs/ setAvail/ accPlac		SHE1 Dataset **-P1 is identified by: E41 Appellation –**P2 has type: E55 Type {“Place Appellation”}
restrctn	codebook/ stdyDscr/ dataAccs/ useStmt/ restrctn		SHE1 Dataset -P104 is subject to right: E30 Right P67i is referred to by: E73 Information Object -P106is composed of: **E73 Information Object P2 has type: E55 Type** {“Restrictions”} and **E73 Information Object** –**P3 has note: E62 String**
citReq	codebook/ stdyDscr/ dataAccs/ useStmt/ citReq		SHE1 Dataset -P104 is subject to right: E30 Right P67i is referred to by: E73 Information Object -P106is composed of: **E73 Information Object –P2 has type: E55 Type** {“Citation Requirement”} and **E73 Information Object –P3 has note: E62 String**
deposReq	codebook/ stdyDscr/ dataAccs/ useStmt/ deposReq		SHE1 Dataset -P104 is subject to right: E30 Right P67i is referred to by: E33 Linguistic Object -P106is composed of: **E73 Information Object –P2 has type: E55 Type** {“Deposit Requirement”} and **E73 Information Object –P3 has note: E62 String**

disclaimer	codebook/ stdyDscr/ dataAccs/ useStmt/ disclaimer		SHE1 Dataset -P104 is subject to right: E30 Right P67i is referred to by: E33 Linguistic Object -P106is composed of: **E73 Information Object –P2 has type: E55 Type** {“Disclaimer”} and **E73 Information Object -P3 has note: E62 String**
relStdy	codebook/ stdyDscr/ othStdyMat/ relStdy		SHE1 Dataset -**P130 shows features of: SHE1 Dataset**
relMat	codebook/ stdyDscr/ othStdyMat/ relMat		SHE1 Dataset -**P129i is subject of: E7 Information Object -P3 has note: E62 String**
relPubl	codebook/ stdyDscr/ othStdyMat/ relPubl		SHE1 Dataset -**P167i is referred to by: E73 Information Object -P2 has type: E55 Type** {“Publication”} and **E73 Information Object -P3 has note: E62 String**
othRefs	codebook/ stdyDscr/ othStdyMat/ othRefs		SHE1 Dataset -**P167i is referred to by: E73 Information Object: E55 Type** {“Supporting Material”} and **E73 Information Object -P3 has note: E62 String**

4.1.3 File Description –Metadata for the data files comprising the collection.

List Of Mappings: File Description

DDI Tag	Unit of Information (embedding)	Condition	SSHOCro
fileName	codebook/ fileDscr/ fileTxt/ fileName		**SHE1 Dataset -P1 is identified by: E41 Appellation**
fileName **@ID**	codebook/ fileDscr/ fileTxt/ fileName @ID	@ID= “”	SHE1 Dataset **-P1 is identified by: E42 Identifier**

caseQnty	codebook/ fileDscr/ fileTxt/ dimensns/ caseQnty		SHE1 Dataset **-P43 has dimension: E54 Dimension -P90 has value: E60 Number** and **E54 Dimension -P91 has unit: E58 Measurement Unit** {"Total Number of Observations"}
varQnty	codebook/ fileDscr/ fileTxt/ dimensns/ varQnty		SHE1 Dataset **-P43 has dimension: E54 Dimension -P90 has value: E60 Number** and **E54 Dimension -P91 has unit: E58 Measurement Unit** {"Total Number of Variables"}
fileType	codebook/ fileDscr/ fileTxt/ fileType		SHE1 Dataset **-P2 has type: E55 Type**

4.1.4 Data Description –Metadata for the variables

List of Mappings –Data Description

DDI Tag	**Unit of Information (embedding)**	**Condition**	**SSHOCro**
var	codebook/dataDscr/ var		SHE1 Dataset –P2 has type: E55 Type {"Variable"}
var **@name**	codebook/dataDscr/ var@name	@name=""	SHE1 Dataset –**P1 is identified by: E41 Appellation** –P2 has type: E55 Type {"Short Label"}
var **@files**	codebook/dataDscr/ var@files	@files=""	SHE1 Dataset –**P106i forms part of: SHE1 Dataset** –P2 has type: E55 Type {"File"}
var **@intrvl**	codebook/dataDscr/ var@intrvl	@intrvl=""	SHE1 Dataset –**P2 has type: E55 Type** {"continuous"} | {"discrete"}

var **@dcml**	codebook/dataDscr/ var@dcml	@dcml=""	SHE1 Dataset –P43 has dimension: E54 Dimension –P90 has value: E60 Number, and **E54 Dimension –P91 has** unit: E58 Unit {"Number of Decimal Points"}
labl	codebook/dataDscr/ var/labl		SHE1 Dataset –P1 is identified by: E41 Appellation –P2 has type: E55 Type {"Short Label"} and E41 Appellation **–P3 has note: E62 String**
qstnLit	codebook/dataDscr/ var/qstn/qstnLit		SHE1 Dataset –P2 has type: E55 Type {"Variable"} and SHE1 Dataset –**P3 has note: E62 String**
range	codebook/dataDscr/ var/valrng/range		SHE1 Dataset **–P43 has dimension: E54 Dimension**
range **@min**	codebook/dataDscr/ var/valrng/range@min	@min=""	SHE1 Dataset –P43 has dimension: E54 Dimension **–P90 has value: E60 Number –P2 has type: E55 Type {"min"}**
range **@max**	codebook/dataDscr/ var/valrng/range@max	@max=""	SHE1 Dataset –P43 has dimension: E54 Dimension **–P90 has value: E60 Number –P2 has type: E55 Type {"max"}**
sumStat	codebook/dataDscr/ var/sumStat		SHE1 Dataset –**P16i was used for: S6 Data Evaluation –O10 assigned dimension: E54 Dimension –P90 has value: E60 Number** and **S6 Data Evaluation –P10 falls within: SHE7 Data Interpretation**
sumStat **@type**	codebook/dataDscr/ var/sumStat@type	@type=""	SHE1 Dataset –P16i was used for: S6 Data Evaluation –O10 assigned dimension: E54 Dimension **–P91 has unit: E58 Measurement Unit**

catgry	codebook/dataDscr/ var/catgry		**SHE1 Dataset –P2 has type: E55 Type** {"Response Documentation"} and **SHE1 Dataset –P106i forms part of: SHE1 Dataset –P2 has type: E55 Type** {"Variable"}
catValu	codebook/dataDscr/ var/catgry/catValu		SHE1 Dataset **–P16i was used for: S6 Data Evaluation –P10 falls within: SHE7 Data Interpretation** and **S6 Data Evaluation–O11 described: S15 Observable Entity –P2 has type: E55 Type** {"Response Value"}, and S15 Observable Entity **–P3 has note: E62 String**
labl	codebook/dataDscr/ var/catgry/labl		SHE1 Dataset –P16i was used for: S6 Data Evaluation –O11 described: S15 Observable Entity **–P3 has note: E62 String**
catStat	codebook/dataDscr/ var/catgry/catStat		SHE1 Dataset –P16i was used for: S6 Data Evaluation –O11 described: S15 Observable Entity **–P43 has dimension: E54 Dimension –P90 has value: E60 Number**
catStat **@type**	codebook/dataDscr/ var/catgry/ catStat@type		SHE1 Dataset –P16i was used for: S6 Data Evaluation –O11 described: S15 Observable Entity –O12 has dimension: E54 Dimension **–P91 has unit: E58 Unit**

4.1.5 DC elements used in DDI Files

Records appearing on the EMM Survey Registry make extensive use of DC metadata terms to represent bibliographic information for the documented research. In principle, one can exclusively use DC elements to document social science research in a DDI record (Martinez 2008, 19). However, it seems that even repositories that resort to the practice of using DC tags, only do so to document bibliographical information for the DDI document or other documents referred in it. The description at the level of the study resorts to the more expressive DDI tags.

Hence, DC tags used, appear either at the level of (i) Document Description (codebook/docDscr/citation) or (ii) other materials relating to the study description (codebook/stdyDscr/othrStdyMat/citation). Not only core elements are invoked, but extended metadata terms as well.

A list of the DC tags used in the DDI records and their mapping to SSHOCro can be found below:

LIST OF MAPPINGS –METADATA IN DC

Dublin Core tag	**Unit of Information (embedding)**	**Condition**	**SSHOCro**
dcterms: spatial	codebook/ docDscr/ citation/ dcterms:spatial		SHE1 Dataset **-P129 is about: E53 Place**
dc: contributor	codebook/ docDscr/ citation/ dc:contributor		SHE1 Dataset **-P94i was created by: E65 Creation -P14 carried out by** {P14.1 in the role of: E55 Type = "Contributor"}**: E39 Actor**
dcterms: temporal	codebook/ docDscr/ citation/ dcterms:temporal		SHE1 Dataset -**P129 is about: E4 Period -P4 has timespan: E52 Time-Span -P82 at some time within: E61 Time Primitive**
dcterms: modified	codebook/ docDscr/ citation/ dcterms: modified		SHE1 Dataset **– P15 was influenced by: E7 Activity -P2 has type: E55 Type** {"Modification"} and **E7 Activity -P4 has timespan: E52 Time-Span -P82 at some time within: E61 Time Primitive**
dc: coverage	codebook/ docDscr/ citation/ dc: coverage		SHE1 Dataset -P129i is subject of: E73 Information Object **-P2 has type: E55 Type** {"Coverage"} and **E73 Information Object** -**P3 has note: E62 String**
dcterms: conformsTo	codebook/ stdyDscr/ othStdyMat/ relMat.ID="relMat _technical"/ dcterms: conformsTo		SHE1 Dataset -P129i is subject of: **E33 Linguistic Object** -P2 has type: E55 Type {"Encoding Type"} and **E33 Linguistic Object** -P3 has note: E62 String

dc: language	codebook/ stdyDscr/ othStdyMat/ relMat/ citation/ dc: language		SHE1 Dataset -P129i is subject of: **E33 Linguistic Object -P72 has language: E56 Language**
dcterms: available	codebook/stdyDscr/ othrStdyMat/ relMat/ citation/ dcterms:available		SHE1 Dataset – **SHR18i was published by: SHE8 Publication -P4 has timespan: E52 Time-Span -P82 at some time within: E61 Time Primitive**

Mappings of the property <**dcterms:available**> and <**dcterms:temporal**> to SSHOCro conform to its definitions in DC as opposed to its use in the DDI files in EMM.

A declared subproperty of <**date**>, <**dcterms:available**> should be used to indicate the date that the resource became/is going to become available.[29] Instead, DDI records from the EMM Survey Registry use it to refer to the terms and conditions under which the survey can be accessed.

4.2 Discussion points

DDI Codebook is a very complex standard. It offers a very large set of elements from which subsets may be used to create a profile. The sum of elements used and the detail in which social science research is documented, varies from one repository to the other. However, the core set that was extracted from DDI-Lite, and CESSDA Mandatory and Recommended Elements and the addition of a few extra elements appearing on the examined records, made it possible to re-express a large set of DDI Codebook elements in SSHOCro – despite the fact that the latter assumes a dynamic, event-centric perspective.

The mapping was driven by the effort to uncover workflow patterns from the DDI documents, such that they would correspond to the discrete stages observed in data-driven research. Despite their very clear structure – metadata for the record, metadata of the study, metadata of any supplementary files, metadata of the variables –DDI records do not distinguish among the separate stages of a research workflow.

One could argue that the workflow is alluded to by the information relayed at each major element of the DDI document. The manipulation of the collected data documented in the <dataDscr> section, for instance, can be said to evoke an analysis/interpretation stage in the workflow. However, as it is, there is no direct reference to the processes involved in data manipulation. Rather, the information made available is about the observed values for each of the documented variables together with some basic statistics covering their distributions. Furthermore, not all DDI records offer a description of the variables used in a research.

The only way to infer the stage of a research workflow that a given action was a part of is by a close inspection of the <**method**> section –and specifically its subsection <**dataColl**>, where descriptions are offered regarding

[29] For the definition of the property, see: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/terms/available/; [20.10.2020]

all aspects of the processes involved in the generation and processing of a given data collection. However, there are a number of problems with that:

(i) **not all concepts lumped under "methodology" involve detailed and context specific problem-solving procedures** –see, for instance, elements like <**dataCollector**> or <**sources**>;

(ii) **not all actual methods form part of one and the same stage in the research workflow**: [30]
For instance, elements <**collMode**>, <**frquenc**>, <**resInstru**>, <**sampProc**>, <**sampleFrame**>, <**targetSampleSize**>, and <**timeMeth**>, all contained in <**dataCollection**>, clearly refer to the data collection process. In SSHOCro terms, they refer to instances of SHE11 Method used during the SHE5 Data Collection stage.
According to their definitions, <**actMin**> and <**cleanOps**> form part of a pre-processing stage—i.e. the point at which the data is prepared for the analysis –and would map to SHE6 Data Preparation Connection, whereas element <**weight**> refers to actual data manipulation methods and would map to SHE7 Data Interpretation.
<**ConOps**> and <**deviat**> can't be attributed to any particular stage in the research workflow –as they characterize the data forming a collection as a whole.

(iii) The **data collection process is overrepresented in most of the examined records to the detriment of the other stages in the research workflow**, which creates the impression that the documentation almost stops once the collection has been achieved.

The effort to link DDI elements with the process that they document lead to large paths, determined by the definition of the DDI elements and the values they assumed across the records examined.

SSHOCro is an event-centric ontology in which linked entities are mediated through events. It is these events that are ascribed temporal properties –not the entities participating in them. That being the case, the time-dimensional attributes listed in DDI elements, had to be connected to some appropriate SSHOCro class –i.e. one encoding a temporal dimension of some sort. The fact that the mediating activity would remain unnamed and implicit in the DDI, meant that it would be rendered through a very high-level appropriate class in SSHOCro, which would subsequently link to one of the major stages of the research workflow observed in SSHOCro.

Given the genericity of the processes described, the nature of entities that participated in them and the kind of statements that needed to be made about them, it seemed reasonable to make use of CRMsci classes and properties in the mappings. Furthermore, some of the documented processes involved entities of a very different nature –like physical objects, documents, people, or events. Seeing as these entities need to be collectively referred to by an appropriate class, S15 Observable Entity seemed like a good fit, as it can be used to characterize aggregations of very dissimilar entities. As for the documented processes, these include S4

[30] Out of the DDI elements listed in this paragraph, the ones in **blue fonts** have not been mapped to the SSHOCro for lack of mention in the DDI Codebook records examined.

Observation, S6 Data Evaluation, and E21 Measurement. The referred properties include: O8 observed, O10 assigned dimension, O11 described, O12 has dimension, and O21 measured.

5. CMDI to SSHOCro

This chapter includes the mapping of CMDI to SSHOCro.

The idea is to establish mappings of the entities/relations/attributes defined in CMDI to an equivalent path in SSHOCro. These are self-explanatory SSHOCro-compatible propositions consisting of long paths of classes and properties of SSHOCro

The mapping from CMDI to SSHOCro takes into account a set of CMDI records adhering to the clarin.eu:cr1:p_1403526079380 profile/schema, i.e. the one used for LINDAT/CLARIN repository resources. In its turn, this profile encapsulates a core set of metadata components that will serve as the basis for the DDI-CMDI conversion.

The approach of the mapping takes into account the values from the actual CMDI records according to the specific profile. Following the general model of CMD framework, a mapping for the generic elements of CMDI apart from the elements of the individual profile used (which are the LINDAT components) is also provided.

The SSHOCro classes or paths corresponding to the CMI elements are marked in boldface.

Marked in {curly brackets} are the types assigned to the SSHOCro/CIDOC-CRM classes to ensure that no information is lost during the mapping process. For instance, a cmd:MdProfile is interpreted in the SSHOCro/ CIDOC-CRM universe as: SHE10 Tool - P2 has type: E55 Type {"profile"}.

A notation, which uses the "value" and "content" along with quotation marks, means that the thing that should instantiate a given class is the very string of characters recorded in the field or subfield within the quotation marks.

SSHOCro models information regarding knowledge processes, the life-cycle of creating, using and finding data –amongst other actions –in the various domains of social sciences and humanities, while CMDI is a format developed in order to exchange and record descriptive information mainly about objects and not processes.

5.1 List of CMDI Mappings

Mappings from CMDI elements to the respective SSHOCro/CIDOC-CRM classes and properties are listed below.

CMDI tag	Unit of Information (embedding)	Condition	SSHOCro -CIDOC-CRM
cmd:CMD			**SHE1 Dataset** -P2 has type: E55 Type {“metadata”}
cmd:Header	cmd:CMD/cmd:Header[31]		SHE1 Dataset – **P94i was created by: E65 Creation**
cmd:MdCreationDate	cmd:CMD/cmd:Header/ cmd:MdCreationDate		SHE1 Dataset – P94i was created by: E65 Creation - **P4 has timespan: E52 Time-Span –P82 at some time within: E61 Time Primitive**
cmd:MdProfile	cmd:CMD/ cmd:Header/ cmd:MdProfile		SHE1 Dataset – P94i created by: E65 Creation – **P16 used specific object: SHE10 Tool - P2 has type: E55 Type** {“profile”}
cmd:MdSelfLink	cmd:CMD/ cmd:Header/ cmd:MdSelfLink		SHE1 Dataset - **P1 is identified by:E42 Identifier**
cmd:MdCollectionDisplayName	cmd:CMD/ cmd:Header/ cmd:MdCollectionDisplayName		SHE1 Dataset – **P106i forms part of: SHE1 Dataset - P2 has type: E55 Type** {“CMDI Collection”}
cmd:ResourceProxy	cmd:CMD/ cmd:Resources/ cmd:ResourceProxyList/ cmd:ResourceProxy		SHE1 Dataset –**P129 is about: SHE1 Dataset**
cmd:ResourceProxy.id	cmd:CMD/ cmd:Resources/cmd:ResourceProxyList/ cmd:ResourceProxy		SHE1 Dataset –P129 is about: SHE1 Dataset **- P1 is identified by: E42 Identifier**

[31] This column shows the current context node (or the current position) of nested elements represented in a path, as it is structured in the xml file.

cmd:ResourceRef	cmd:CMD/cmd:Resources/cmd:ResourceProxyList/cmd:ResourceProxy/cmd:ResourceRef		SHE1 Dataset –P129 is about: SHE1 Dataset **- P1 is identified by: E41 Appellation – P2 has type: E55 Type** {“Access Point”}
cmd:ResourceType	cmd:CMD/cmd:Resources/cmd:ResourceProxyList/cmd:ResourceProxy/cmd:ResourceType		SHE1 Dataset –P129 is about: SHE1 Dataset **- P2 has type: E55 Type**
cmd:ResourceType.mimetype	cmd:CMD/cmd:Resources/cmd:ResourceProxyList/cmd:ResourceProxy/cmd:ResourceType		SHE1 Dataset –P129 is about: SHE1 Dataset - P2 has type: **E55 Type -- P2 has type: E55 Type** {“media type”}
cmd:JournalFileProxyList	cmd:CMD/cmd:Resources/cmd:JournalFileProxyList		SHE1 Dataset –**P129 is about: SHE1 Dataset– P2 has type: E55 Type** {“Journal file”}
cmd:Components/cmd:LINDAT_CLARIN	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN		SHE1 Dataset –**P129 is about: SHE1 Dataset/E33 Linguistic Object**
cmd:title	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:bibliographicInfo/cmd:titles/cmd:title		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object **- P1 is identified by: E35 Title**
author	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:bibliographicInfo/cmd:authors/ author		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object- **P94i was created by: E65 Creation - P14 carried out by** { P14.1 in the role of: E55 Type=”author”}**: E39 Actor**
firstName	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:bibliographicInfo/cmd:authors/author/firstName		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object- P94i was created by: E65 Creation - P14 carried out by { P14.1 in the role of: E55 Type=”author”}: E39 Actor **- P1 is identified by:E41 Appellation – P2 has type: E55 Type** {“first name”}

lastName	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:bibliographicInfo/cmd:authors/author/lastName		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object- P94i was created by: E65 Creation - P14 carried out by {P14.1 in the role of: E55 Type= ”author”}: E39 Actor- **P1 is identified by:E41 Appellation – P2 has type: E55 Type** {“last name”}
cmd:projectUrl	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:bibliographicInfo/cmd:projectUrl		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object - **SHR6i was used by project:SHE3 Project Activity - P1 is identified by:E41 Appellation– P2 has type: E55 Type** {“url”}
cmd:dates /cmd:dateIssued	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:bibliographicInfo/cmd:dates/cmd:dateIssued		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object-– **P94i was created by: E65 Creation: P2 has type: E55 Type** {“Publication”} and **E65 Creation: P4 has timespan: E52 Time-Span –P82 at sometime within: E61 Time Primitive**
cmd:publishers/cmd:publisher	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:bibliographicInfo/cmd:publishers/cmd:publisher		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object-– **P94i was created by: E65 Creation: P2 has type: E55 Type {“Publication”} and E65 Creation: P14 carried out by** {P14.1 in the role of: E55 Type=”publisher”}: **E39 Actor**
cmd:identifiers/cmd:identifier	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:bibliographicInfo/cmd:identifiers/cmd:identifier		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object-– **P1 is identified by:E42 Identifier**
cmd:identifiers/cmd:identifier. type	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:bibliographicInfo/cmd:identifiers/cmd:identifier		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object-– P1 is identified by:E42 Identifier- **P2 has type: E55 Type**

cmd:funds/ funding	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/ cmd:bibliographicInfo/ cmd:funds/ funding		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object-–- **SHR6i was used by project:SHE3 Project Activity –P9 consists of:E7 Activity- P2 has type: E55 Type** {“funding”}
projectName	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/ cmd:bibliographicInfo/ cmd:funds/funding/ projectName		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object-– SHR6i was used by project:SHE3 Project Activity- **P1 is identified by:E35 Title**
code	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/ cmd:bibliographicInfo/ cmd:funds/funding/ code		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object-– SHR6i was used by project:SHE3 Project Activity - **P1 is identified by:E42 Identifier**
organization	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/ cmd:bibliographicInfo/ cmd:funds/funding/ organization		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object-–- SHR6i was used by project:SHE3 Project Activity – P9 consists of:**E7 Activity - P2 has type: E55 Type** {“funding”} **and E7 Activity: P14 carried out by:E40 Legal Body**
fundsType	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/ cmd:bibliographicInfo/ cmd:funds/funding/ fundsType		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object-–- SHR6i was used by project:SHE3 Project Activity – P9 consists of:**E7 Activity- P2 has type: E55 Type** {“funding”} **and E7 Activity: P2 has type: E55 Type**
contactPerson	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/ cmd:bibliographicInfo/ contactPerson		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object- **P94i was created by: E65 Creation - P14 carried out by: E39 Actor**
firstName	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/ cmd:bibliographicInfo/ contactPerson/ firstName		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object- P94i was created by: E65 Creation - P14 carried out by: E39 Actor -- **P1 is identified by:E41 Appellation – P2 has type: E55 Type** {“first name”}

lastName	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:bibliographicInfo/contactPerson/lastName		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object- P94i was created by: E65 Creation - P14 carried out by: E39 Actor **-- P1 is identified by:E41 Appellation – P2 has type: E55 Type** {"last name"**}**
email	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:bibliographicInfo/contactPerson/email		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object- P94i was created by: E65 Creation - P14 carried out by: E39 Actor **– P76 has contact point: E51 Contact Point**
affiliation	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:bibliographicInfo/contactPerson/affiliation		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object- P94i was created by: E65 Creation - P14 carried out by: E39 Actor **–P107i is current or former member of: E40 Legal Body**
cmd:type	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:dataInfo/cmd:type		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object **- P2 has type: E55 Type**
cmd:detailedType	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:dataInfo /cmd:detailedType		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object - P2 has type: E55 Type – **P127i has narrower term: E55 Type**
cmd:description	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:dataInfo/cmd:description		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object **– P3 has note: E62 String**
cmd:languages/cmd:language	cmd:CMD/cmd:Components/cmd:LINDAT_CLARIN/cmd:dataInfo/cmd:languages/cmd:language		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object **–P72 has language: E56 Language**

cmd:code	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/ cmd:dataInfo/ cmd:languages/ cmd:language /cmd:code		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object **–**P72 has language: E56 Language **- P1 is identified by:E42 Identifier**
cmd:name	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/ cmd:dataInfo/ cmd:languages/ cmd:language /cmd:name		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object **–**P72 has language: E56 Language **- P1 is identified by: E41 Appellation**
cmd:keywords/ cmd:keyword	cmd:CMD/ cmd:Components/cmd:LINDAT_CLARIN/cmd:dataInfo/ cmd:keywords/ cmd:keyword		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object **–P129 is about: E1 CRM Entity**
cmd:links/ cmd:link	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/ cmd:dataInfo/ cmd:links/cmd:link		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object **–P67i is referred to by: E73 Information Object - P1 is identified by: E41 Appellation - P2 has type: E55 Type** {“access point”}
cmd:sizeInfo/size	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/ cmd:dataInfo/ cmd:sizeInfo/ size		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object–**P43 has dimension: E52 Dimension**
size	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/ cmd:dataInfo/ cmd:sizeInfo/size/size		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object –P43 has dimension: E52 Dimension **– P90 has value:E60 Number**
unit	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/cmd:dataInfo/ cmd:sizeInfo/ size/unit		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object –P43 has dimension: E52 Dimension **–P91 has unit: E58 Measurement Unit**

cmd:licenseInfo/ cmd:license	cmd:CMD/ cmd:Components/ cmd:LINDAT_CLARIN/ cmd:licenseInfo/ cmd:license		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object **–P104 is subject to: E30 Right**
cmd:uri	cmd:CMD/ cmd:Components/cmd:LINDAT_CLARIN/cmd:licenseInfo/ cmd:license/ cmd:uri		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object –P104 is subject to: E30 Right **- P1 is identified by : E42 Identifier** {“uri”}
cmd:formats /cmd:format/ cmd:medium	cmd:CMD/ cmd:Components/cmd:LINDAT_CLARIN/cmd:dataInfo/ cmd:formats/cmd:format/cmd:medium		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object **– P2 has type: E55 Type** {“format”}
cmd:name	cmd:CMD/ cmd:Components/cmd:LINDAT_CLARIN/cmd:dataInfo/cmd:formats/cmd:format/cmd:name		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object – P2 has type: E55 Type **-P1 is identified by : E41 Appellation**
cmd:description	cmd:CMD/ cmd:Components/cmd:LINDAT_CLARIN/cmd:dataInfo/cmd:formats/cmd:format/ cmd:description		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object – P2 has type: **E55 Type -– P3 has note: E62 String**
cmd: requirements/ cmd:requirement	cmd:CMD/ cmd:Components/cmd:LINDAT_CLARIN/cmd:dataInfo/cmd:requirements /cmd:requirement		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object - **P3 has note: E62 String**
cmd:genres/ cmd:genre	cmd:CMD/ cmd:Components/cmd:LINDAT_CLARIN/cmd:dataInfo/cmd:requirements /cmd:requirement		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object **– P2 has type: E55 Type** {“genre”}
cmd: annotationInfo/ cmd: annotationType	cmd:CMD/ cmd:Components/cmd:LINDAT_CLARIN/cmd:dataInfo cmd: annotationInfo/ cmd: annotationType		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object **– P2 has type: E55 Type** {“annotation type”}

cmd: version	cmd:CMD/ cmd:Components/cmd:LINDAT_CLARIN/cmd:bibliographicInfo/cmd: version		SHE1 Dataset –P129 is about: SHE1 Dataset/E33 Linguistic Object - **P1 is identified by: E42 Identifier -P2 has type: E55 Type** {"version"}

5.2 Discussion points

One of the important things of the mapping procedure is to distinguish the resources, specifically the metadata record (CMDI:CMD) itself which has its own documentation (creation information, id etc.: CMDI:Header) and the actual resource which has a detailed documentation (the component) or other potentially described resources/datasets, and they are related with an aboutness relation (metadata record is about resource/dataset).

A common problem in the mapping process is the ambiguous semantics of the expressions and structures used in CMDI (Ďurčo et al. 2018). A representative example is the cmd:ResourceProxy which is part of the metadata record (and not of the component) and may express three different semantics, such as a) different access points for the same resource, b) items of a collection and c) different resources that may be part or not of the component. The relation to be assigned between all these is even more difficult to understand the way it is accomplished. It seems that there is no clear distinction between the sources and the direct relationship of one source to another. In most of the examples, the relationship information was not even declared.

A similar problem that shows that a complete correspondence from CMDI to SSHOCro cannot be achieved is the case of the CMD element "relationsInfo''. This element is described as an element that groups information on the relations of the resource being described with other resources. It includes the element "relationType'' which specifies the type of relation between resources and additionally, the related resource to link with. The problem is that an abstract and non-semantically explicit relationship between resources cannot be mapped to an explicit, specified relationship from SSHOCro. Especially, when the type of the relation depends on accidental verbal expressions or values from terminologies that the users use instead of particular semantics links. Almost every entity and relation can be assigned a classification however that indicates a particular semantic interpretation, requiring concepts from an external semantic model or vocabulary - the terms used in a domain and in natural language to express propositions and the conceptual structure that lies behind these expressions are two different things. This distinction between linguistic and conceptual levels is also represented in a mapping procedure. An important part of the mapping is interpreting relations. Often semantics are built into values. It is suitable if we know a priori all the possible values, but this is not the case here, since the type of the relations is not a controlled or fixed /defined vocabulary, which can be interpreted in a semantic statement, in the mapping procedure. In that sense, this specific element has not been mapped to SSHOCro. Typing of relationships does not help to semantic integration and using/querying the data in practice.

Additionally, missing values from the fields do not help to understand the semantics. Misleading is also when the values are not inserted in the right field. For example, inside the description of a funding there is reference to the project name and the code, which is information actually belonging to the project and not to the funding, although it is contained in the funding – it seems that in xml, different semantically elements are concatenated. Another problem is the practice of having very generic elements without specifying exactly the definition and their role, such as Contact Person (defined as "Person to contact in case of issues with the submission….e.g. one of the authors **or** the submitter" – this means that is two different concepts, the author or the submitter without knowledge of the condition that justifies the one or the other. Generally, the fields that are defined as two different things make the mapping difficult, such as the "Project URL", which is defined as "URL of resource/project related to the submitted item". The examples studied imply that the "project url" actually means the resource url. This is information that comes from the data entry. This makes the harmonisation difficult. However, the most important problem is that CMDI focuses on the representation of a resource and not of a workflow.

SSHOCro is a workflow model which aims to describe the whole data life cycle in social sciences and humanities research including both the generations and the processing of the data. CMDI is not a workflow metadata standard; in that sense, it is difficult to identify a common pattern of workflow between a static model and a procedural model.

CMDI mainly documents tools/web-services, such as a computational lexicon (which is a tool) or an annotated corpus. However, their description is static and from the perspective of the archivist. To connect the data produced or manipulated to a broader workflow, presupposes close inspection of the data and supporting material (cited papers). The notion of a workflow remains implicit and can be hinted at by the keywords used, insofar as these correspond to immediately accessible methods linked to distinct stages of the workflow.

It represents static stages of a research workflow. The basic item of the documentation which is the resource, is documented as the research output of a project, without explicitly describing the project as part of a workflow or even identify the project as an activity that used assets and had parts other activities or stages that produced it.

A characteristic example that shows the problem is the one from an online bilingual valency lexicon, of which the metadata offered do not permit to either locate when the said resource can be used within a larger research workflow (it is documented as the research output of a project) or to trace the stages of the workflow that produced it. In fact, to arrive at these distinct stages –plus any iterations thereof –the reader needs to extensively study the supporting files (manual editing of annotations performed iteratively, the results of statistic tests indicating the agreement among annotations and the certainty of the results and the verbs explored at each iteration). Aside that, to understand the processes involved required a lot of background reading on dependency treebanks, semantic roles in various frameworks, annotated corpora, parallel corpora, etc.

Publications (of data, services and papers) are not described as participating in a procedure, as input or output. Instead, they are documented as final products, which are not linked to any activity that produced it or used it.

None of these metadata standards include the documentation of (i) the link of research questions to the data collected to prove/falsify/describe them, (ii) the methods observed in data collection process, (iii) any interim results/conclusions that stem from the collected data, (iv) any causes for revising interim conclusions –for instance, biases in the data collection or data analysis stage – (v) repurposing and reusing data to address other research questions

CMDI is based on keywords or data types elements and is not a semantically rich model.

6. References

Arofan G., Open Data Foundation (2011). The Data Documentation Initiative (DDI): An Introduction for National Statistical Institutes. Available at: http://odaf.org/papers/DDI_Intro_forNSIs.pdf

Bekiari, C., Doerr M., le Boeuf P., Riva P. (2014). FRBR object-oriented; Definition and Mapping from FRBRer, FRAD and FRSAD (version 2.4). Online publication, available at: http://www.cidoc-crm.org/frbroo/sites/default/files/FRBRoo_V2.4.pdf [28.09.2020]

Bekiari,C., Kritsotaki, A., & Tsouloucha, E. (2020). SSHOC D4.18 SSHOC Reference Ontology (beta version) (Version v1.0). Zenodo.

Bhattacherjee, Al. (2012). Social Science Research: Principles, Methods, and Practices. Textbooks Collection. 3.

Bosch, T., Cyganiak, R., Gregory, A., & Wackerow, J. (2013). DDI-RDF Discovery Vocabulary: A Metadata Vocabulary for Documenting Research and Survey Data. LDOW.

Bosch, T., & Mathiak, B. (2015). Use Cases Related to an Ontology of the Data Documentation Initiative.

Broeder D., Gaiffe B., Gavrilidou M., Hinrich E. Lemnitzer L., Van Uytvanck D., Witt A., Wittenburg P. (eds.) (2009). Metadata Infrastructure for Language Resources and Technology.

Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel,T., and Windhouwer, M. (2012). Standardizing a component metadata infrastructure. In Proceedings of LREC, Istanbul, Turkey, pp.1387-1390.

CESSDA (2016). CESSDA Service Providers' Metadata Practices. Standards, Controlled Vocabularies and requirements for the CESSDA Portfolio. CESSDA Metadata Management Project Combined Deliverable D1 & D2. June 2016. Retrieved from: https://www.cessda.eu/content/download/834/7776/file/CMM_ServiceProvidersMetadataPractices_2016 [10.06.2020]

CLARIN (n.d.). Component Metadata. Retrieved from https://www.clarin.eu/content/component-metadata [23.06.2020]

CLARIN-D AP 5 (2020). CLARIN-D User Guide v1.1 (2020-01-21). Retrieved from: https://media.dwds.de/clarin/userguide/userguide-1.1.pdf [20.6.2020]

Crofts, N. et al. (eds.) (2011) Definition of the CIDOC Conceptual Reference Model - cidoc_crm_version_5.0.4.pdf. [online]. Available from: http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf [14.05.2014].

Doerr, M. (2003) The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI magazine. 24 (3), 75.

Doerr, M., Bekiari, C., Kritsotaki, A., Hiebel, G., Theodoridou, M. (2014). Modelling Scientific Activities: Proposal for a global schema for integrating metadata about scientific observation, CIDOC - International Documentation Committee of ICOM: Content metadata - Administrative metadata - Technical metadata - Legal metadata -Conference 6th-11th Sept. 2014 in Dresden

Doerr, M., Kritsotaki, A., Rousakis, Y., Hiebel, G., Theodoridou M. (2018). Definition of the CRMsci; An Extension of CIDOC-CRM to support scientific observation (Version 1.2.6). [online]. Available from: http://www.cidoc-crm.org/crmsci/sites/default/files/CRMsci%20v.1.2.6.pdf [10.12.2020]

DDI Alliance (2014). DDI-Codebook 2.5 XML Schema Documentation. Retrieved from: https://ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html [10.6.2020]

DDI Alliance (2020). DDI-Lifecycle 3.3 XML Schema Documentation. Retrieved from: https://ddialliance.org/Specification/DDI-Lifecycle/3.3/XMLSchema/FieldLevelDocumentation/ [20.6.2020]

DDI-Codebook: DDI 2.5 Detailed Change Specification https://ddialliance.org/Specification/DDI-Codebook/2.5/detailed_changes_to_ddi-2.pdf; [25.11.2020].

Ďurčo, M., Lorenzini M., Sugimoto G. (2018). Something will be connected - Semantic mapping from CMDI to Parthenos Entities. CLARIN. Selected papers from the CLARIN Annual Conference 2017. Budapest: Linköping University Electronic Press

ISO 24622-1 (2015) Language resource management — Component Metadata Infrastructure (CMDI) — Part 1: The Component Metadata Model.

ISO 24622-2 (2019) Language resource management — Component metadata infrasctructure (CMDI) — Part 2: Component metadata specification language.

Martinez, L. (2008) The Data Documentation Initiative (DDI) and Institutional Repositories. DISC-UK, February 2008. Available at: http://www.disc-uk.org/docs/DDI_and_IRs.pdf

Vardigan, Mary (2013). The DDI Matures: 1997 to the Present and 'Timeline'. IASSIST Quarterly 37.

Windhouwer, M., Broeder, D., & Van Uytvanck, D. (2012). A CMD core model for CLARIN web services. In Proceedings of the workshop on Describing Language Resources with Metadata: Towards Flexibility and Interoperability in the Documentation of Language Resources at LREC 2012 (pp. 41-48).

Online resources consulted

ADP Catalogue: https://www.adp.fdv.uni-lj.si/opisi/; [30.04.2020]

CLARIN Virtual Language Observatory: https://vlo.clarin.eu; [11.02.2020]

CMDI Component Registry: https://catalog.clarin.eu/ds/ComponentRegistry/#/; [06.05.2020]

DataverseNO: https://dataverse.no/; [17.03.2020]

EMM Survey Registry: https://ethmigsurveydatahub.eu/emmregistry/; [30/05/2020]

FSD Data Catalogue: https://services.fsd.uta.fi/catalogue/index?lang=en&study_language=en; [10.03.2020]

LINDAT/CLARIAH-cz Repository: https://lindat.mff.cuni.cz/repository/; [06.05.2020]

NESSTAR/SoDaNet: http://nesstar-server.sodanet.gr/webview/; [28.04.2020]

SND research data catalogue: https://snd.gu.se/en/search/content; [08.10.2020]

Appendix

A. Implementation of the mappings of DDI and CMDI to SSHOCro

This section refers to the implementation of the mapping of the selected elements from DDI and CMDI to SSHOCro in 3M.

The resources necessary to implement the mappings in 3M (see section: X3ML Toolkit) are listed below:

1. the rdf definition of **CIDOC-CRM v.6.2.1**[32]
2. **crmpc v1.0** – an implementation of the properties of properties (the ".1" properties) of CIDOC CRM in rdf [33]
3. the rdf definition of **crm-sci v1.2.6 CRMsci**– a scientific observation model [34]
4. the **uri-generators policy file** attached with this report (see Linked Files, below)
5. the rdf definition of **SHOCCro v1.1.3** attached with this report (see Linked Files, below)

The mappings can be accessed through the 3M installation at FORTH.[35] To access the mappings, one is required to log on to the service using either his/her account or the guest account (username: guest, password: guest9802). Guests are not granted editing rights, but they can view and download the files mentioned above.

The X3ML output of the mapping from DDI Codebook to SSHOCro[36] (**Title**: MappingFromDDItoSSHOCro.xml) is attached with this report (see Linked Files, below).

[32] **CIDOC-CRM v.6.2.1** (rdf definition): http://www.cidoc-crm.org/sites/default/files/cidoc_crm_v6.2.1-2018April.rdfs; [10.09.2020]

[33] **CRMpc v1.0** (rdf implementation of .1 properties): https://www.google.com/url?q=http://www.cidoc-crm.org/sites/default/files/CRMpc_v1.0.rdfs&sa=D&ust=1607536868985000&usg=AOvVaw03YXrRh6mCaUwkNhkeXNUw; [10.09.2020]

[34] **CRMsci v1.2.6** (rdf definition): http://www.cidoc-crm.org/crmsci/sites/default/files/CRMsci_v1.2.6.rdfs [10.12.2020]

[35] **3M** access point: https://isl.ics.forth.gr/3M/

[36] **ID**: 2178, **Title**: DDI Codebook --> Mapping to SSHOCro; accessible through: https://isl.ics.forth.gr/3M/

The DDI instances used to test the mapping to SSHOCro are represented by two use cases from FSD:

(1) FSD3062 Finnish Attitudes to Immigration: Suomen Kuvalehti Survey 2015[37]

The record forms the description of a survey, commissioned by the newsmagazine Suomen Kuvalehti, which charted attitudes in Finland towards immigrants from different countries as well as beliefs about race.

(2) FSD3063 Finnish Attitudes to Immigration: Iltalehti Survey 2015[38]

The record forms the description of a survey, commissioned by the tabloid newspaper Iltalehti, which charted public opinion on Finland's refugee policy and asylum seekers arriving in the country.

The X3ML output of the mapping from CMDI to SSHOCro[39] (Title: MappingFromCMDItoSSHOCro.xml) is attached with this report (see Linked Files, below).

The CMDI instances used to test the mapping to SSHOCro are represented by two cases from LINDAT/CLARIAH-CZ Repository Home:

(1) The English Valency Lexicon[40]

EngVallex is the English counterpart of the PDT-Vallex valency lexicon, using the same view of valency, valency frames and the description of a surface form of verbal argument. EngVallex is available in an XML format in LINDAT/CLARIAH-CZ Repository.

(2) ParCorFull: A Parallel Corpus Annotated with Full Coreference[41]

ParCorFull is a parallel corpus annotated with full coreference chains that has been created to address an important problem, the translation of coreference across languages. It is also available in xml.

[37] FSD3062 Finnish Attitudes to Immigration: Suomen Kuvalehti Survey 2015, accessed at: https://services.fsd.tuni.fi/catalogue/FSD3062?tab=description&lang=en&study_language=en; [16.09.2020]

[38] FSD3063 Finnish Attitudes to Immigration: Iltalehti Survey 2015: https://services.fsd.tuni.fi/catalogue/FSD3063?tab=description&lang=en&study_language=en; [16.09.2020]

[39] **ID**: 2175, **Title**: CMDI --> Mapping to SSHOCro; accessible through: https://isl.ics.forth.gr/3M/

[40] The English Valency Lexicon, accessed at https://hdl.handle.net/11858/00-097C-0000-0023-4337-2 [01.09.2020]

[41] ParCorFull: A Parallel Corpus Annotated with Full Coreference, accessed at: https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2614 [01.09.2020]

B. Linked files

A supplementary zipped file is attached to this report. Its contents are:

1. **D4.19_MappingFromDDiToSSHOCro_ID.2178.zip**; it contains all the necessary files to implement the mapping –(i) the source DDI Codebook .xml file, (ii) the target schemas used for the mapping, i.e. (ii.a) the rdf definition of SSHOCro (v1.1.3), (ii.b) the rdf definition of CIDOC-CRM (v.6.2.1), (ii.c) the rdf definition of CRMsci (v1.2.6), and (ii.d) the rdf implementation of CRM ".1" properties – crmpc (v1.0) – , (iii) the .xml file of the uri-generators policy, and (iv) the output of the mappings –(iv.a) the mapping in .x3ml and (iv.b) the integrated data in .ttl.
2. **D4.19_MappingFromCMDItoSSHOCro_ID.2175.zip**; it contains all the necessary files to implement the mapping –(i) the source CMDI .xml file, (ii) the target schemas used for the mapping, i.e. (ii.a) the rdf definition of SSHOCro (v1.1.3), (ii.b) the rdf definition of CIDOC-CRM (v.6.2.1), (ii.c) the rdf definition of CRMsci (v1.2.6), and (ii.d) the rdf implementation of CRM ".1" properties – crmpc (v1.0) –, (iii) the .xml file of the uri-generators policy, and (iv) the output of the mappings –(iv.a) the mapping in .x3ml and (iv.b) the integrated data in .ttl.
3. **SSHOC Reference Ontology v.1.13-report.pdf**; the definition of SSHOCro as it stands at the time of the publication of D4.19
4. **SSHOCro_version1_1_3.rdf**; the rdf definition of the SSHOCro as it stands at the time of the publication of D4.19
5. **SSHOCroUriGeneratorPolicy_v1.5**; the .xml file of the uri-generators policy
6. **CRMpc_v1.0**; the rdf implementation of CRM ".1" properties
7. **MappingDDI_SSHOCro.x3ml**; the X3ML output of the mapping from DDI Codebook to SSHOCro
8. **MappingCMDI_SSHOCro.x3ml**; the X3ML output of the mapping from CMDI to SSHOCro
9. **fsd_3062.xml**; the DDI-xml record documenting the Survey "Finnish Attitudes to Immigration: Suomen Kuvalehti Survey 2015"
10. **fsd_3063.xml**; the DDI-xml record documenting the Survey "Finnish Attitudes to Immigration: Iltalehti Survey 2015"
11. **CIMDI_LINDAT_ParCorFull.xml**; the CMDI-xml record documenting the "ParCorFull" parallel corpus
12. **CIMDI_LINDAT_EngValLex.xml**; the CMDI-xml record documenting the "EngValLex" valency lexicon.

C. X3ML Toolkit

The X3ML Toolkit [42] is a set of small, open source, micro services that follow the Synergy Reference Model of data provision and aggregation,[43] developed by the Centre of Cultural Informatics of Information Systems Laboratory of Institute of Computer Science of FORTH.[44]

The "Synergy Reference Model of Data Provision and Aggregation" is relevant to the aggregation of all cultural datasets and provides the necessary infrastructure for sustainable provisioning of data from museums, archives and libraries, as well as many other specialist datasets. Currently it draws upon the work of the CIDOC CRM Special Interest Group (CRM SIG), a working group of CIDOC, the International Committee for Documentation of the International Council of Museums (ICOM).

The Synergy model is designed to create an environment that can achieve a multiplicity of benefits because it is based around a mapping schema accessible to different kinds of experts (curators, technologists, data managers, etc.) on both the provider and aggregator sides. The Synergy model describes three key aspects of data provisioning. These are:

1. The alignment of a provider's internal data model with an aggregator's target model.
2. The transfer of data to populate the aggregator's data repository.
3. The ongoing processes to maintain this alignment and the regular update of data.

The Synergy model supports large scale quality aggregation and facilitates collaboration and the reuse of knowledge and expertise. It is also the foundation for creating communities of people and tools and developing greater awareness of the potential of data aggregation. The Synergy system is different from many other aggregation systems because it encourages and supports a fuller representation of data than is generally undertaken. It supports the aligning of a full source model to a target model that supports meaning and context. Once models have been aligned and all data errors have been resolved the data can be transferred to the aggregator. Synergy recommends that the raw data also be transferred so that information can be recovered without any need for action by the provider. The model includes processes for validating identifiers and creating modification dates. The transferred data is then transformed into the aggregator's model using the X3ML instructions. Mapping, furthermore, only needs to be done once, and then can be used for a whole range of purposes or converted to any other cultural heritage standards.

The X3ML tool allows data experts to transform their internal structured data and other associated contextual knowledge to other schemas and, in particular, the CIDOC CRM (Conceptual Reference Model). Fields or elements from a source database (Source Nodes) are aligned with one or more entities described in the target

[42] For more information on the X3ML Toolkit, see: https://www.ics.forth.gr/isl/x3ml-toolkit; [03.12.2020]

[43] Oldman et. al.; The Synergy Reference model of data provision and aggregation. Accessible at: http://www.cidoc-crm.org/sites/default/files/SRM_v1.5.pdf [09.12.2020]

[44] Centre of Cultural Informatics of Information Systems Laboratory of Institute of Computer Science of FORTH: https://www.ics.forth.gr/isl/centre-cultural-informatics; [09.12.2020]

schema so that the data from an entire system can be transformed. The purpose of this is typically for publication on the Web and in particular meaningful integration with other data also transformed to the same target schema.

The X3ML framework supports data transformation and aggregation, by producing mappings, to relate equivalent concepts or relationships from the source schemata to the aggregation schema, i.e., the target schema. A **mapping** is defined as the (automated) transformation of each instance of schema 1 into an instance of schema 2 with the same meaning.

The output of the mapping process is a collection of mapping rules. The rules are given in X3ML, a declarative, human readable language that supports the cognitive process of a mapping.

The main steps of the **data provisioning workflow** are:

Schema matching: source and target schema experts (i.e., the domain experts) define a schema matching which is documented in a schema matching definition file. This file should be human and machine readable and it is the ultimate communication mean on the semantic correctness of the mapping.

Instance generation specification: in this step the URI generation and data type conversion policies are defined for each instance of a target schema class referred to in the matching. In this step only IT experts are involved

Terminology mapping: the terminology mappings between source and target data/terms are defined. Providers may use anything from intuitive lists of uncontrolled terms up to highly structured third-party thesauri.

Transformation: once the mapping definition has been finalized (and all syntax errors are resolved) the data needs to be transformed, producing a set of valid target records. The transformation process itself may run completely automatically. In the case where any issues arise, the aggregator can resolve them on a temporary or permanent basis but it is also possible that these records are sent back to the provider for further analysis and resolution.