A CIDOC-CRM based ontology: The SSHOCro

The SSHOCro ontology: a workflow model

Athina Kritsotaki, FORTH

SHOC Archaeological Case Study Workshop
25 May 2021
Zoom

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Project:

SSHOC
social sciences & humanities open cloud

Horizon 2020
European Union Funding
for Research & Innovation

Type of action & funding:
Research and Innovation action
(INFRAEOSC-04-2018)

Partners: 45
(20 beneficiaries + 25 LTPs)

SSH ESFRI Landmarks and Projects
& international SSH data infrastructures

Duration: 40 months
(January 2019 – 30 April 2022)

Project budget:
€ 14,455,594.08

Project website:
www.SSHopencloud.eu

Objectives:

- creating the social sciences and humanities **(SSH)** part of European Open Science Cloud **(EOSC)**
- maximising **re-use** through **Open Science** and **FAIR** principles (standards, common catalogue, access control, semantic techniques, training)
- interconnecting existing and new infrastructures (clustered cloud infrastructure)
- establishing appropriate **governance model** for SSH-EOSC

The SSHOC Reference Ontology(SSHOCro) : Modeling the SSHOC data life cycle

a common metalevel schema, to be used as a top-level ontology for organizing knowledge and information distributed across various primary sources of information in the Social Sciences and Humanities Open Cloud (SSHOC).

to provide a semantic interoperability framework for the description of the SSHOC **data life cycle** in the Social Sciences and the Humanities.

Achieving this goal goes through the following steps:

- Consultation with SSH data producers
- SSHOCro version (RDF/S)
- Mapping selected metadata standards to the SSHOCro

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

The SSHOC Reference Ontology(SSHOCro) : Modeling the SSHOC data life cycle

the basic empirical foundations for the formulation of the model was built on:

- representative research workflows used by partners from SSH community, identified in a workshop organized by FORTH
- research papers reporting the methods and results of experimental studies in a number of scientific domains from the social sciences and humanities.
- extensive literature review on metadata standards used by the SSHOC communities
- search on online resources/repositories for metadata records adhering to their respective metadata schemas - retrieving and analyzing records and data from dedicated SSH repositories, such as FSD Data Catalogue, DataverseNO, EMM Survey Registry and LINDAT/CLARIAH-cz Repository
- use of existing top level models:
  CIDOC CRM, CRM-sci, Parthenos, SO ontology

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Replication of

Affirmative Action Policies Promote Women and Do Not Harm Efficiency in the Laboratory

by Balafoutas, L. / Sutter, M. (2012)
in: Science, 335, pp. 579–582

Replication Authors:
Felix Holzmeister, Jürgen Huber, Michael Kirchler, and Julia Rose

---

In a set of controlled laboratory experiments, Balafoutas and Sutter (2012) study the effects of different affirmative action policy interventions to encourage women's choice to enter competitions. Four different interventions are investigated: quotas, where one of two winners of a competition must be female; two variants of preferential treatment, where a fixed increment is added to women's performance; and repetition of the competition, where a second competition takes place if no woman is among the winners. Compared with no intervention, all interventions encourage women to enter competitions more often and performance is at least as good both during and after the competition.

Hypothesis to replicate and bet on:

With preferential treatment of women — i.e., each woman's performance is automatically increased by one unit in the competition — more women will choose to compete (a comparison of the fraction of women who chose the tournament scheme rather than the piece rate scheme in the 'preferential treatment one (PT1)' versus the 'control treatment (CTR)'; $\chi^2(1) = 5.62$, $p = 0.018$, p. 580).

(This hypothesis was picked by lottery instead of comparing PT2 to CTR; $\chi^2(1) = 10.89$, $p = 0.001$, p. 580).

Power Analysis and Criteria for Replication: First Data Collection

The original sample size is 144 participants and the standardized effect size measured as the correlation coefficient (r) is 0.197. To have 90% power to detect 75% of the original effect size, a sample size of 485 is required. The criteria for replication are an effect in the same direction as the original study and a p-value < 0.05 (in a two-sided test).

Power Analysis and Criteria for Replication: Second Data Collection

If the original result is not replicated in the first data collection, a second data collection is carried out. To have 90% power to detect 50% of the original effect size in the pooled sample (first and second data collection), a sample size of 1099 is required, i.e., a sample size of 614 in the second data collection is required. The criteria for replication are an effect in the same direction as in the original study and a p-value < 0.05 (in a two-sided test) in the pooled data.

Sample

The sample in the first data collection consists of 485 students from the University of Innsbruck. If the original result is not replicated in the first data collection (two-sided p-value < 0.05 in the same direction as the original study), a second data collection consisting of 614 additional students from the University of Innsbruck will be carried out such that the pooled sample size is 1099. Subjects who participated in the experimental sessions of the original studies are excluded from recruiting.

Materials

We use the software of the original experiment programmed in z-Tree (Fischbacher, 2007) along with the original German instructions which have been made available by the authors.

Procedure

We follow the procedure of the original study, with only slight but unavoidable deviations as outlined below. The following summary of the experimental procedure is therefore based on the explanations of the experimental procedure in the article (pp. 579–80) and the section "Notes on the experimental procedure" (p. 3–4) of the Supplementary Information.

Subjects are randomly assigned into groups of three men and three women. All groups go through several stages. The experimental task in each of the stages 1 to 4 is to add as many sets of five two-digit numbers as possible within 3 minutes. Ties between participants are broken randomly in stages 2, 3, and 4. The task in stage 5 is a simple coordination game. At the beginning of the experiment, subjects are informed about the number of stages but not about what the tasks in each of the stages will be. The instructions for each of the task are provided just before every new stage.

In stage 1 (piece rate), each subject receives €0.50 for each correct calculation. In stage 2 (tournament), group members compete against each other. The two members who solve the most calculations correctly are paid €1.50 per calculation. The other four group members receive nothing. Subjects do not receive any feedback on the outcome of the competition in stage 2 until the end of the experiment to avoid that subjects condition their choices on previous outcomes of a competition. In stage 3 (choice), subjects choose whether they want to solve the calculations under a piece rate scheme or a tournament scheme. If the tournament is chosen, a subject's performance in stage 3 is compared with the other group members' performance in stage 2. In this stage, the competition rule across the two treatments PT1 and CTR are varied to examine the effects of the policy intervention: In the control treatment (CTR), the winners are the two group

treatment	order	session	group	subject	period	gender	choice
Preferential	1	161103_0820	1	4	1	Female	
Preferential	1	161103_0820	1	4	2	Female	
Preferential	1	161103_0820	1	4	3	Female	Tournament
Preferential	1	161103_0820	1	4	4	Female	Tournament
Preferential	1	161103_0820	1	4	5	Female	Tournament
Preferential	1	161103_0820	1	4	6	Female	Tournament

that subjects condition their choices on previous outcomes. At the end of stage 3, beliefs of all subjects regarding their relative performance and their ranks in stages 1 and 2 are elicited. For each stage, subjects have to indicate their expected rank within the group of six members and within their own gender only. Correct guesses are rewarded with €1.00 each, and the feedback is given also only at the end of the experiment. In stage 4 (tournament with policy intervention), all subjects

Documentation of

- sample,
- data collection,
- software and
- stages of data collection phase

in Holzmeister et al., that replicated Balafoutas & Sutter's (2012) experiment

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

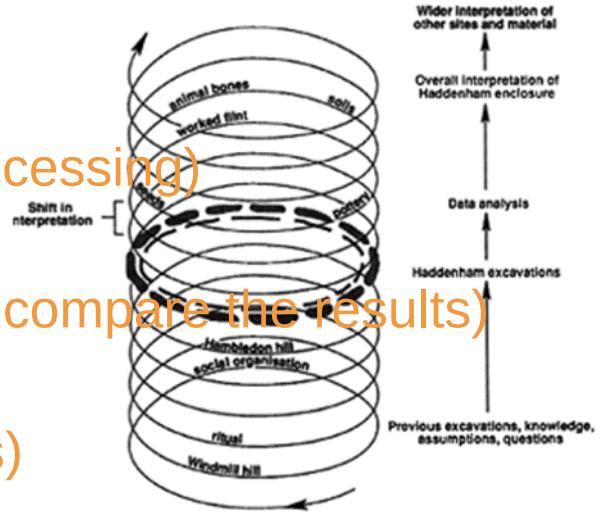
SSHOC
social sciences & humanities open cloud

SSHOCro practical use:

SSHOCro is a workflow model that aims to describe the full **data life cycle** in SSH research

- built on the ground of analytical methods used in various disciplines to inform a common **workflow**:

- **Form of a hypothesis to perform an observation**
- **Perform the observations**
- **Explain the observations made and the gathering of data** (processing)
- **Draw conclusions based upon this data**,
- **Deduce the implications** (test them through further observation, compare the results)
- **Confirm, deny, re-evaluate the original hypothesis**
- **Formulate valid theories** (allow others to repeat the observations)



- uses and extends the **CIDOC CRM (ISO21127),** an **event based** ontology

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

SSHOCro practical use:

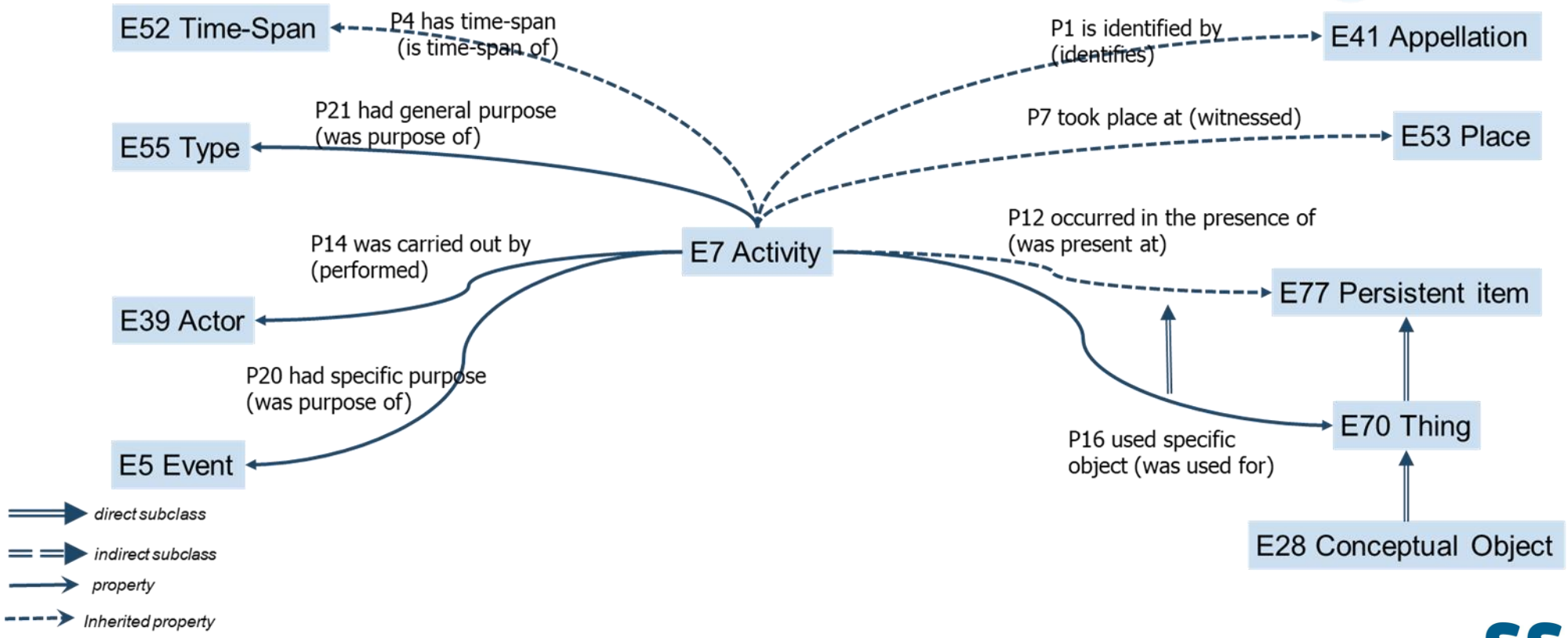
- It can be applied as a standard to devise and implement metadata capture schemes for tracking the data life-cycle in individual projects/institutions/disciplines.
- Common language between Social Science & Humanities researchers with IT specialists
- Encoded in a semantic data format (e.g. RDF), it can be of use for mapping, transforming and integrating existing data across projects/institutions/disciplines into interoperable pools (semantic repositories) of information for re-use and further exploitation.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

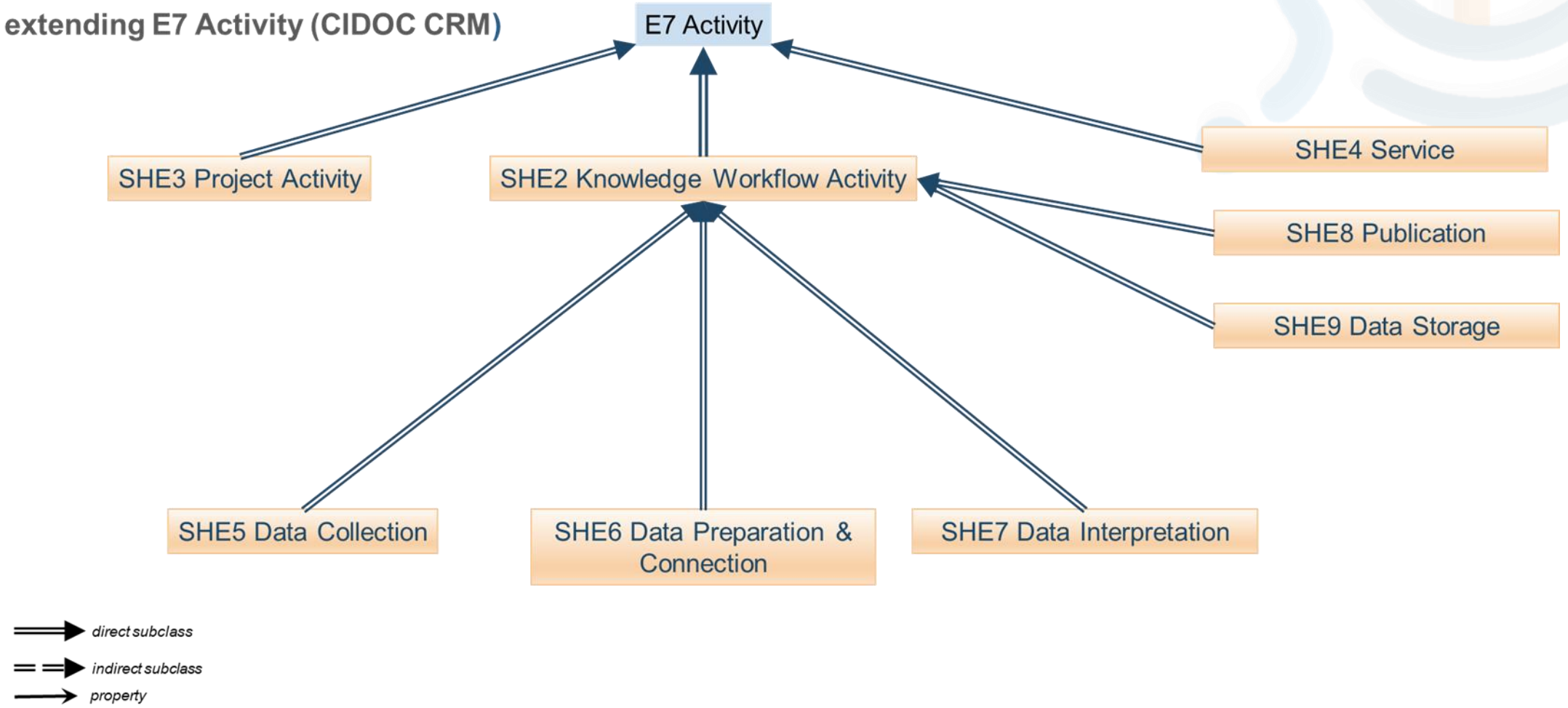
SSHOCro uses CIDOC CRM E7 Activity

The relations linking instances of E7 Activity to the entities necessary for describing them are inherited by the activities specifically defined for SSHOCro



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOCro (an extension of CIDOC CRM): overview



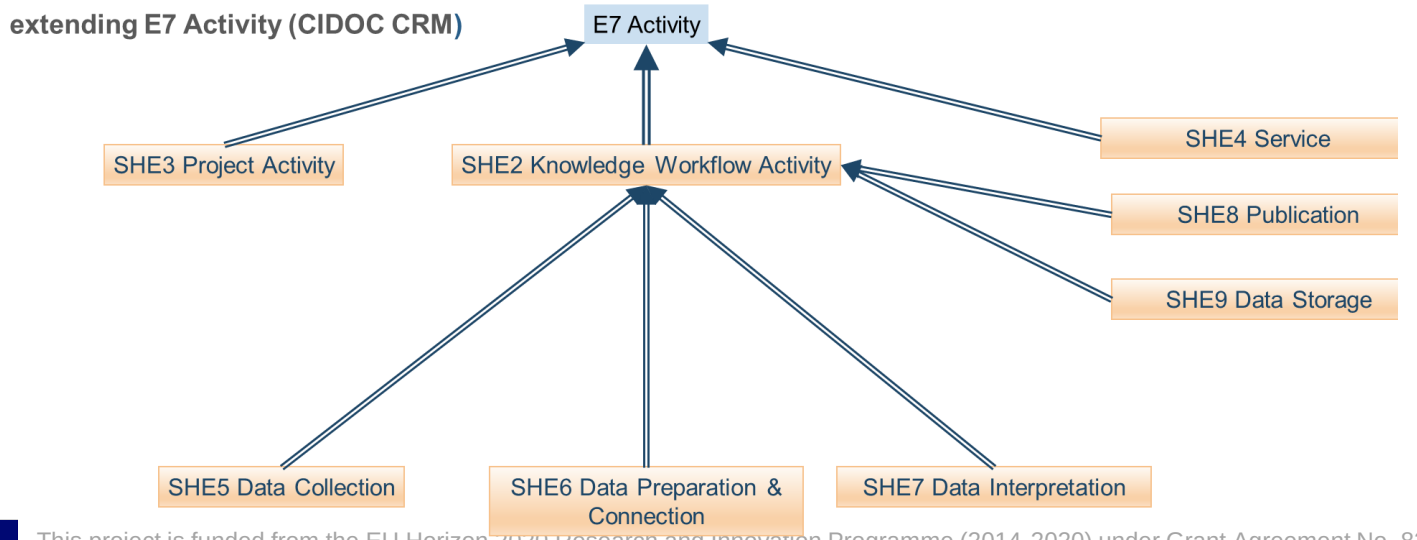
This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

The SSHOCro workflow research process

The main phases

- Data Collection phase
  the processes involved in collecting datasets
- Data preparation & connection phase
  how to treat missing values and outliers,
  or to of identify individuals across the datasets
- Data Interpretation phase
  examining or comparing to test theories
  offer plausible explanations regarding the examined phenomena



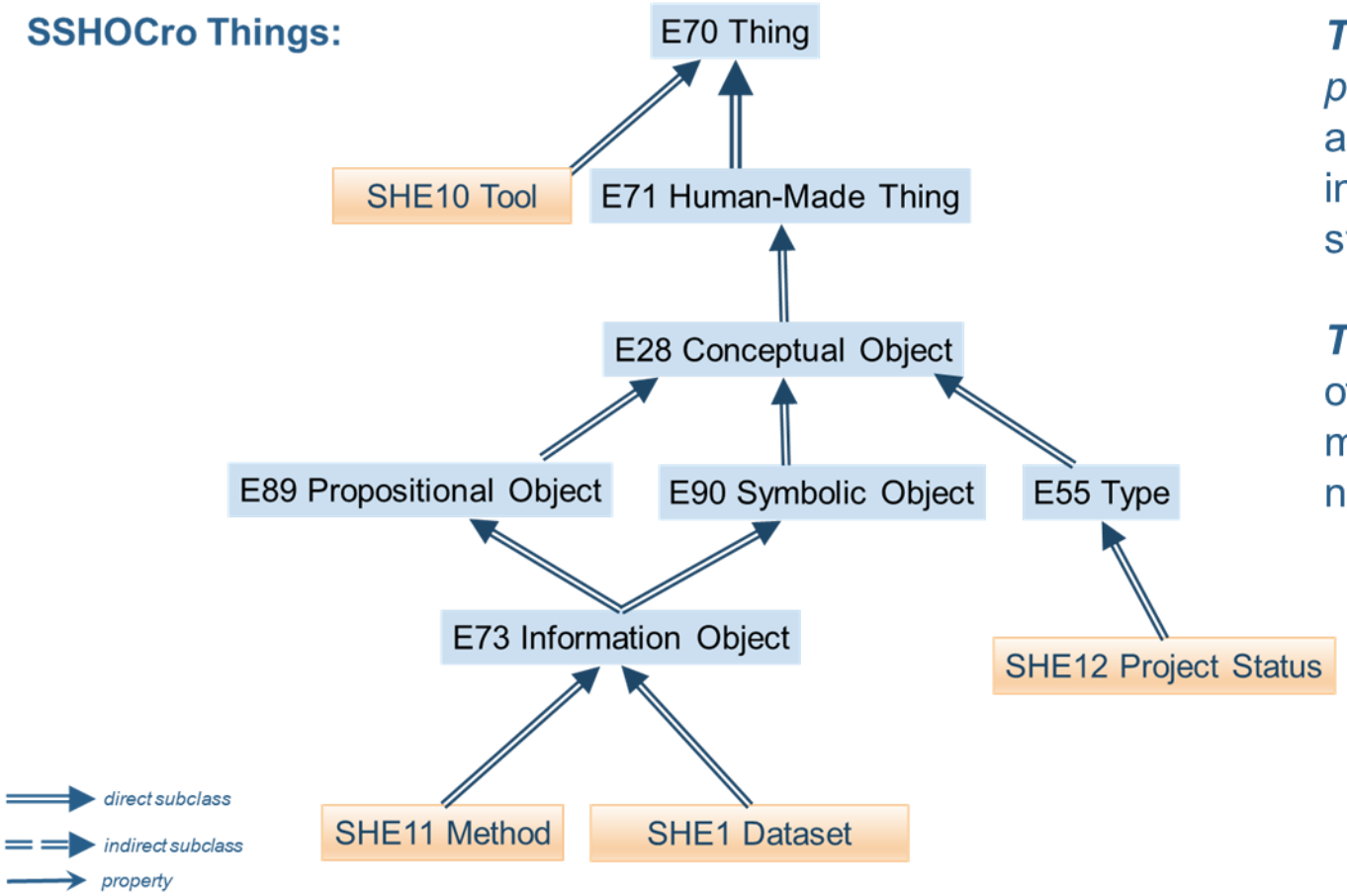
3 auxiliary services

- Persistent Storage
  - physical, protected storage spaces and object conservation
  - electronic media, digital preservation and physical media storage.
  - Curation and access methods.
- Publication and Presentation
  - medium is paper, digital file or active database
  - sites and collections to be visited
  - text, data, graphics, animation, VR
  - publication = announcing public availability
- Information Selection & access
  - finding, retrieving, inspecting, and selecting

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

SSHOCro Objects: SHE1 Dataset (input/output resources)



Things correspond to discrete, identifiable, *persistent items*. They can be material –like any sort of concrete object –or immaterial (f.i. images, texts, datasets, organizational structures etc).

Things are involved in activities in the context of which they are created, used/operated on, modified or destroyed –depending on the nature of the activity.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

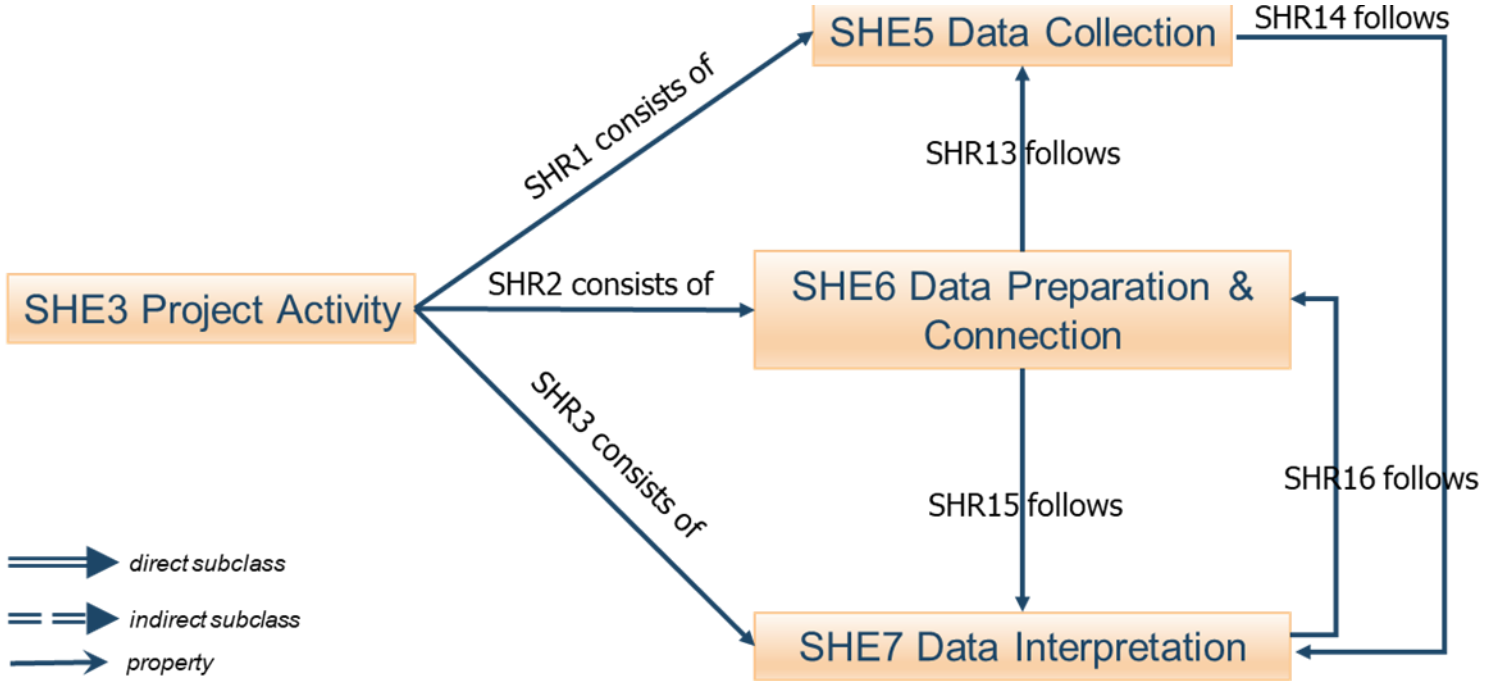
The SSHOCro workflow research process

- Among the greatest issues for empirical evidence oriented SSH feature:
  - verification/falsification of the final research results through the revision of primary data.
  - reuse and enhancement of scientific results by means of examining new empirical data.

- documentation of the provenance of knowledge in each phase of the workflow

- provenance documentation is necessary for achieving stepwise documentation

The SSHOCro non linear, iterative phases:



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

The Problem:

- The notion of a workflow remains implicit in all examined metadata standards; metadata instances primarily used to document research in SSH are static and adopt the perspective of the archivist, following the completion of a research outcome –typically some sort of publication (data/services/papers).
- Linking publications to their role within a research workflow requires additional effort than originally planned; it involves a close inspection of not only metadata instances, but also actual data and supporting publications.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Implicit workflow in DDI: the case of **<method>**

Methodology and Processing involved in a data collection. The elements embedded in this node:

- concern the process of data collection and data cleaning
  - sampling procedure followed (sampProc);
  - methods observed that are relevant for the data collection (timeMeth; sampProc; collMode; collSitu, resInstru)
  - data cleaning operations (cleanOps; actMin)
- refer to actions that precede & inform the actual data collection
  - sampling frame used for identifying the sampled population (sampleFrame);
  - desired sample size, given the population size (targetSampleSize)
- concern the data analysis
  - generalizations of the observed patterns across a population (free text descriptions, where any: DataAppr; EstSmpErr; respRate; dataProcessing)
  - Variable manipulation –presented as the output of the data analysis –does not fall under methods.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

Mappings to the SSHOCro

The conceptual mapping process: establishing correspondences among selected elements from indicative metadata standards (e.g DDI, CMDI) with elements/full paths of SSHOCro using the X3ML toolkit

Aims to:

- Facilitate the data integration from various sources of SSH repositories and databases
- To confirm that SSHOCro **approximates the reality** of the SSH research process
- Resolve any discrepancies between SSHOCro and the other metadata standards

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

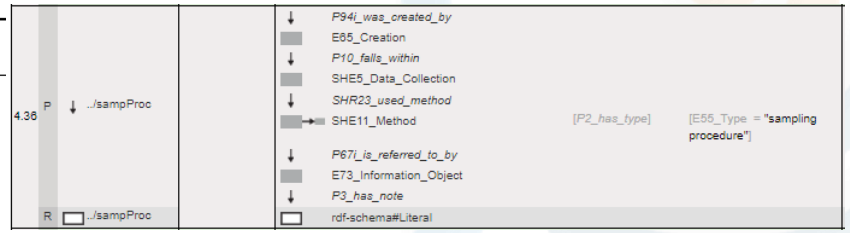
SSHOC
social sciences & humanities open cloud

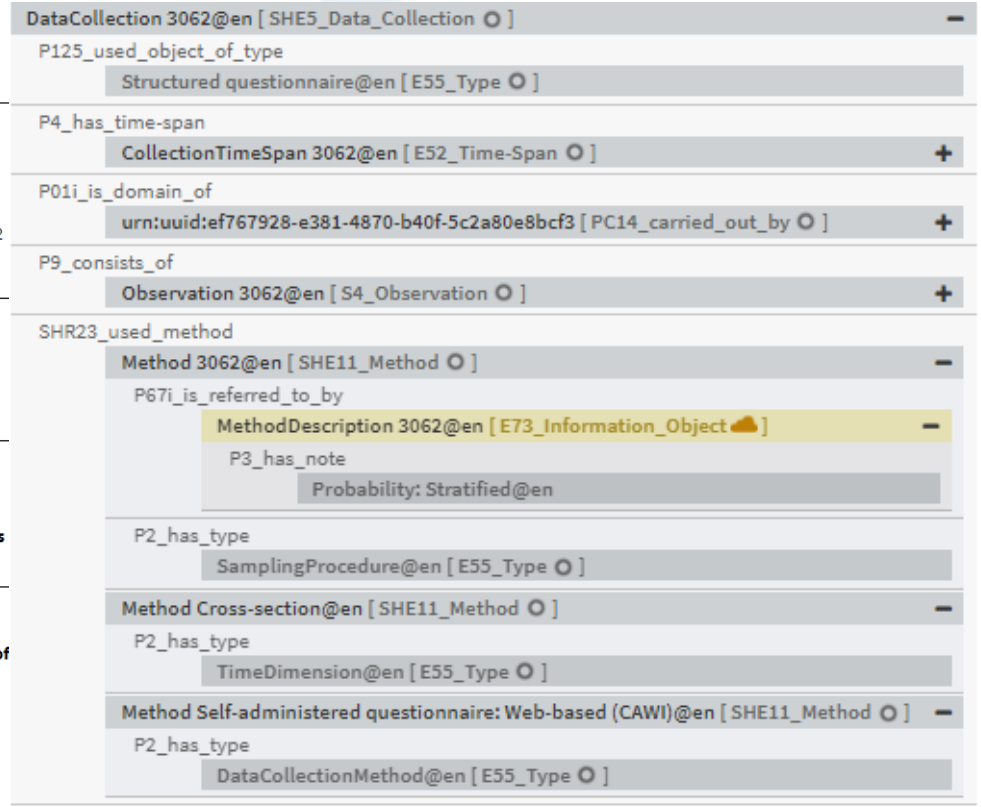
Mappings to the SSHOCro

```
<method>
 ▼<dataColl>
  ▼<timeMeth>
     Cross-section
     <concept>CrossSection</concept>
   </timeMeth>
   <dataCollector>Taloustutkimus</dataCollector>
  ▼<sampProc>
     <p>Probability: Stratified</p>
     <p>The sample was drawn from the Population Register of
     Finland through stratified random sampling.</p>
   </sampProc>
   <collMode>Telephone interview: Computer-assisted (CATI)
   </collMode>
   <resInstru>Structured questionnaire</resInstru>
   <weight>There is a weight variable bv13 which weighs the
   data to represent the adult population of Finland aged 15
   and over in terms of age, gender, region of residence and
   household size. The unweighted n-number indicates the
   number of interviewees in various population groups and the
   weighted n-number indicates the number of persons in the
   corresponding actual population group (per thousand
   persons).</weight>
  </dataColl>
  <stdyClas type="A">Detailed and specific data description in
  Finnish and English. Variable frequencies, variable and value
  labels, and missing values are checked. If necessary, the
  data are anonymised.</stdyClas>
</method>
```

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"}
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**





This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

Thank you for your attention!

Join our community

https://www.sshopencloud.eu

@SSHOpenCloud

info@sshopencloud.eu

/in/sshopencloud

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud