Mapping non-workflow & non event-based models to CIDOC-CRM based, event-centric, workflow ontologies: The case of SSHOCro

Athina Kritsotaki, FORTH

49th CIDOC CRM & 42nd FRBR CRM sig meeting;
8 -11 March
Zoom

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:

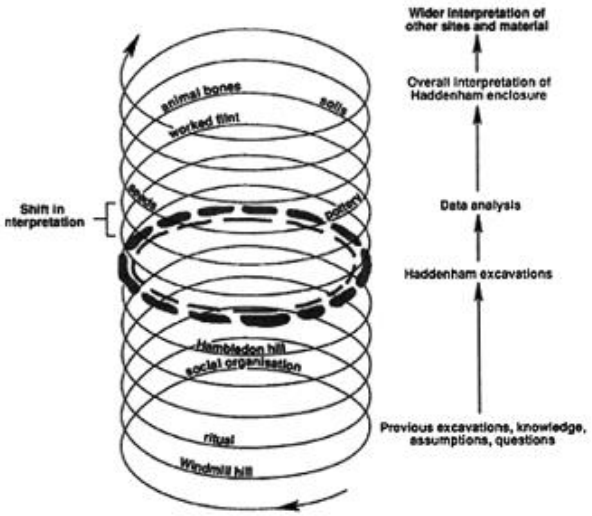
- Consultation with SSH data producers
- SSHOCro version (RDF/S)
- Mapping selected metadata standards to the SSHOCro

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** (processir
- **Draw conclusions based upon this data**,
- **Deduce the implications** (test them through further observation, comp
- **Confirm, deny, re-evaluate the original hypothesis**
- **Formulate valid theories** (allow others to repeat the observations)

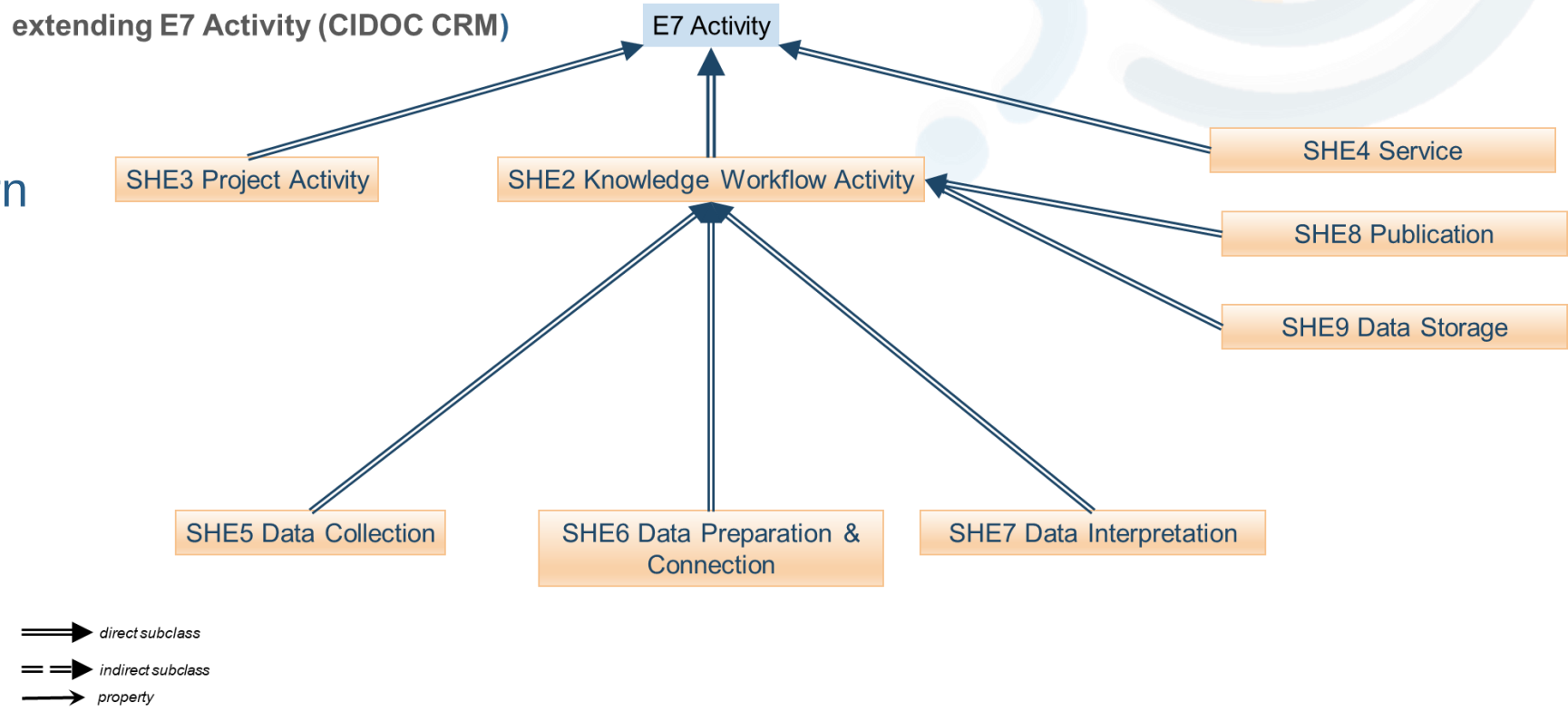


- uses and extends the CIDOC CRM (ISO21127), an **event based** ontology

SSHOCro (extends CIDOC CRM: E7)

The model captures:

- the dominating, iterative pattern found in SSH research:
  - Collection
  - Connection
  - Interpretation
- auxiliary actions that concern
  - persistent storage
  - publication
  - presentation
  - information selection

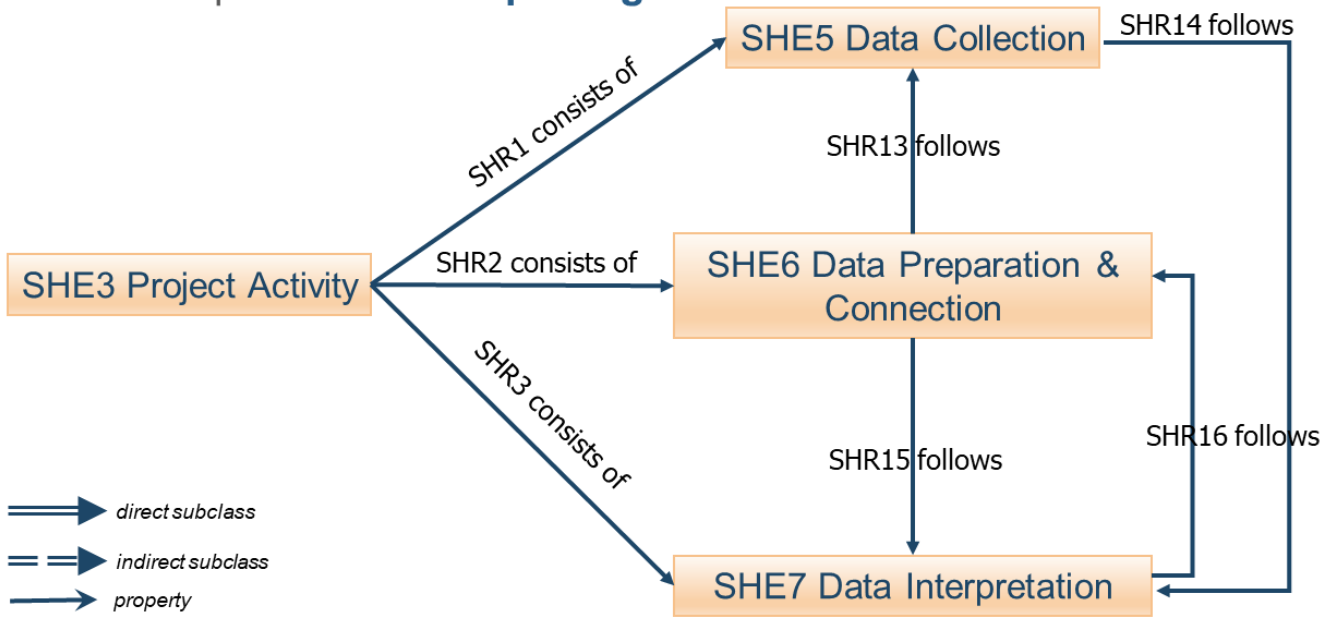


SSHOCro scientific workflow process

- 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 every step of the evaluation chain
- stepwise documentation proposed is only rendered possible to the extent that provenance documentation is integrated in the workflow.

SSHOCro divides the scientific workflow process into three nonlinear, iterated stages:

- ❑ the **data collection** phase - (qualitative and quantitative);
- ❑ the **preparation and/or the connection** of datasets .
- ❑ the process of **interpreting datasets**



Mapping selected standards to the SSHOCro

Information integration & harmonization tested by mapping selected* metadata from two indicative SSH metadata standards to the common SSHOCro schema.

- DDI Codebook:
  - International standard for describing surveys, questionnaires, statistical data files and social science studies.
- CMDI (LINDAT/CLARIAH-cz Repository):
  - Metadata schema describing language resources (datasets and tools/services)

*Decision informed by metadata standards used by the SSHOC communities

Mapping selected standards to the SSHOCro – The problem

- The metadata standards examined represent **non-actualized idealizations of research activities or static stages** thereof.
- They do not define separate stages in the research workflows used across the SSH, establish the order in which separate tasks typically appear and whether they are delimited or connected
- **The notion of a workflow remains implicit;** metadata instances used to document research in SSH are **static** and adopt the perspective of the archivist. The basic concept is the resource, **without 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 **(no provenance).**

Mapping selected standards to the SSHOCro – The DDI

DDI Codebook:

- **Workflow**: inferred through close inspection of **methods** listed.
  - emphasizes the methods and processes involved in **collecting census/survey data <u>BUT</u>**:
    i. Not all concepts under "**methodology**" involve detailed and context specific problem-solving procedures (person names etc.)
    ii. Methods **are not linked to a stage**
    iii. No reference to processes involved in **data manipulation**

Mapping selected standards to the SSHOCro – The DDI

- SSHOCro events are ascribed temporal properties –not the entities participating in them.
- These events (f.i. creation of a dataset) remain unnamed and implicit in the DDI.
- It is linked to one of the major stages of the workflow observed in SSHOCro (SHE5 Data Collection).
  - SHE1 Dataset -P94i was created by:

E65 Creation-P10 falls within: SHE5 Data Collection
-P4 has timespan: E52 Time-Span

```
<stdyInfo>
<subject>
<keyword vocab="ELSST">immigration</keyword>
</subject>
<abstract>
<p>The survey, commissioned by the newsmagazine Suome
Kuvalehti, charted attitudes in Finland towards
immigrants from different countries as well as belief
about race.</p>
</abstract>
<sumDscr>
<timePrd date="2015-00-00" event="single"/>
<collDate date="2015-08-12" event="start"/>
<collDate date="2015-08-13" event="end"/>
<nation abbr="FI">Finland</nation>
<geogCover>Finland</geogCover>
<anlyUnit>
Individual
<concept>Individual</concept>
</anlyUnit>
<universe clusion="I">People aged 15-79 living in Fin
</universe>
<dataKind>Quantitative</dataKind>
</sumDscr>
</stdyInfo>
```

Mapping selected standards to the SSHOCro – The CMDI

CMDI (LINDAT/CLARIAH-cz Repository):

- Emphasis on **objects and their properties** (not events and temporal dimension)**,**
  - basic item of documentation: **the resource** (the research output of a project)
  - the project not represented as consisting of a workflow/ being part of a broader workflow

- Emphasis on **categorical information rather** (not factual): based on keywords or data types elements (not a semantically rich model)
  - "relationType”: specification of the relation between resources and link to the related resource.

  The type of the relation depends on values from terminologies rather than particular semantics links

Mapping to SSHOCro : questions– issues addressed for discussion

- Can we create a successful data transfer mechanism **if** the mapping is not complete? (the source schemas are not event based)

  if we miss semantic equivalences? **If** a workflow cannot be represented?
- What kind of mappings we create if we can’t have CRM compatible propositions?
- Where do we stop the mapping?
- In many cases we have to introduce intermediate nodes implying activities that do not exist in the source schema in order to interpret a temporal aspect of the data – is this effective / a practical solution? What kind of interpretations should be produced?
- Are there other cases of similar mapping challenges/problems to be shared? If yes, what kind of approaches have been followed?

Thank you for your attention!