Semantic Data for Humanities and Social Sciences (SDHSS)

A Methodology and Ecosystem of CIDOC CRM Extensions for Research Data Production and Reuse

Francesco Beretta
CNRS UMR5190 LARHRA - Université de Lyon

with the contribution of

Vincent Alamercery, Stephen Hart and Pierre Vernus

58th CIDOC CRM SIG Meeting — Paris, BNF, 20 March 2024
1.

The context:
production, publication and re-use of research data
in the Humanities and Social Sciences (HSS)
Findable
Accessible
Interoperable
Re-usable

«There is an urgent need to improve the infrastructure supporting the *reuse* of scholarly data »

The FAIR Data Principles

To be **Interoperable:**

I1. (meta)data use a *formal, accessible, shared, and broadly applicable* language for *knowledge representation*.

I2. (meta)data use *vocabularies that follow FAIR principles*.

I3. (meta)data include qualified references to other (meta)data.

To be **Re-usable:**

R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a *clear and accessible data usage license*.

R1.2. (meta)data are associated with their *provenance*.

R1.3. (meta)data meet *domain-relevant community standards*.
Linked Open Data

https://lod-cloud.net/
"An ontology is a formal explicit specification of a shared conceptualization of a domain of interest"

- « Formality – ... a knowledge representation language that is based on the grounds of **formal semantics**. »
- « Consensus – ... an agreement on a domain conceptualization among **people in a community**. »
- « Conceptuality – ... in terms of **conceptual symbols** that can be intuitively grasped by humans, as they **correspond** to the elements in their **mental models**. »
- « Domain Specificity – ... limited to knowledge about a particular **domain of interest**. »

What we observe:

- more and more HSS research projects are using the CIDOC CRM to produce research data
What we observe:

- more and more HSS research projects are using the CIDOC CRM to produce research data
- different dialects, interpretations and data production practices emerge
What we observe:

- more and more HSS research projects are using the CIDOC CRM to produce research data
- different dialects, interpretations and data production practices emerge
- new project-driven or business-driven CRM extensions are being developed
What we observe:

- more and more HSS research projects are using the CIDOC CRM to produce research data
- different dialects, interpretations and data production practices emerge
- new project-driven or business-driven CRM extensions are being developed
- they often concern the same objects and states of affairs, and conceptualise them in different ways
2. The challenge

Produce a multi-disciplinary, shared and formalised conceptualisation which allows to cover different research domains and fits the needs of different scientific disciplines
What are scientific disciplines?
What are scientific disciplines?

- Communities of knowledge that share a common domain of discourse and of scientific interest.
What are scientific disciplines?

- Communities of knowledge that share a common domain of discourse and of scientific interest.
- They define scientific objects and research questions that are considered as valid in a scientific domain.
What are scientific disciplines?

- Communities of knowledge that share a common domain of discourse and of scientific interest.
- They define scientific objects and research questions that are considered as valid in a scientific domain.
- They share a specific methodology which must be empirical and provide reproducible results.
What are scientific disciplines?

- Communities of knowledge that share a common domain of discourse and of scientific interest.
- They define scientific objects and research questions that are considered as valid in a scientific domain.
- They share a specific methodology which must be empirical and provide reproducible results.
- They produce new knowledge in their domain and this knowledge may have more or less impact on human societies in which they are embedded.
The domain of HSS
Material and biological world
Mental reality

Material and biological world
Mental reality

Social reality

Material and biological world
Social representations

Individual minds

Social representations

Individual minds
How are conceptualized observable mental and social phenomena in HSS disciplines?
How are conceptualized observable mental and social phenomena in HSS disciplines?

- Social representations (social sciences):

«social identity emerges based on social representations ... Identity is a way of organizing meanings, of being constructed as a social subject» Álvarez Bermúdez, Javier and Juana Juárez-Romero. “Identities, Memory and the Construction of Citizenship.” Papers on Social Representations 28, no. 2 (2019).
How are conceptualized observable mental and social phenomena in HSS disciplines?

- **Social representations (social sciences):**

  "social identity emerges based on social representations ... Identity is a way of organizing meanings, of being constructed as a social subject"

- **Collective intentionality (social philosophy):**

  "Consciousness and intentionality are caused by and realized in neurobiology. Collective intentionality is a type of intentionality, and society is created by collective intentionality. ... language enables the creation and continue dexistence of status functions that do not require any physical existence beyond the linguistic representations themselves. »
Reference literature (selection):

- The Stanford Encyclopedia of Philosophy (Fall 2021 Edition), Edward N. Zalta (ed.) (online) (especially entries: Collective Intentionality, Mental Representation, Consciousness and Intentionality, Social Norms, etc.)
Research Questions
Artefacts and their social significance:
a boundary stone that defines a boundary

Border stone from c. 1280 on the border of the lands of the Wrocław bishops.

In the background the contemporary road sign marking the border between Opole and Wałbrzych voivodships.

Photo: Władysław Łoś (CC By SA 3.0)

The physical object, the notion of border, the symbol of the episcopal crozier, the function of bishop.

The bio-physical world, the social world
Social roles: the rite that establishes the function
Different perspectives on the same events: interpreting a car race as fun or as a crime
Scientific Methodologies
FIGURE 1.1 A schematic representation of the epistemological cycle

Outline a research agenda

Refine the research questions

Collect documents and artefacts, define a population

Produce information (stored in digital format)

Reuse existing data

Critical method

Classification model

Publication existing data

Code, analyze and visualise the information

Research questions

Statistical model

Interpret the analysis results

Wrangle, clean and aggregate the data

Complex model

Publish the scientific results and document the analysis process

Refine the research questions

Publish your research data

Conceptual model

Collect documents and artefacts, define a population

Produce information (stored in digital format)
3.

Building an ecosystem
of CIDOC CRM extensions
in order to produce reusable research data
in the Humanities and Social Sciences

The methodology
Information as representation
Building an *information* graph (*knowledge graph*) representing the objects in the world (example from Wikidata)
Information as *representation* of the world:
Information as **representation** of the world:

- representation of the **objects** in the world

Social Representations (Collective Intentionality)
Information as **representation** of the world:

- representation of the **objects** in the world
- of their **properties** (qualities)
Information as **representation** of the world:

- representation of the **objects** in the world
- of their **properties** (qualities)
- of their **relationships**
Information is not knowledge!
These are significantly different epistemic levels

Francesco Beretta (CNRS/Université de Lyon), 7 July 2020 CC BY-NC-SA 4.0
Information is not knowledge! These are significantly different epistemic levels.

Francesco Beretta (CNRS/Université de Lyon), 7 July 2020 CC BY-NC-SA 4.0
Whose collective intentionality is to be modelled?
Scientific knowledge and social representations

- Researchers
- Science as collective intentionality
- Critical (methodological) distance
- Social representations
- Societies under consideration

Knowledge
Information
A methodology based on a layered approach
Research specific data model
- Foundational ontologies & modelling best practices
  - Generic, domain related core ontology
    - Domain related extensions
      - Research specific data model
- Research data
Foundational ontologies & modelling best practices

Generic, domain related core ontology

Domain related extensions

Research specific data model
« Foundational ontologies are not directly used for applications; rather, they provide conceptual handles to solve cases of misunderstandings due to the limitations of expressiveness of the application languages. [...] 

DOLCE has remained fixed over the years fulfilling the purpose of top-level ontologies to provide a solid and stable basis for modeling different domains, in this way ensuring interoperability of reference and domain ontologies that use DOLCE. »

OntoClean


Essence (as defined by rigid intensional properties)
– A ‘student’ : not a class but a time-indexed property of a person

Identity (criteria to clearly distinguish individuals)
– The identity of a building as individual does not depend on its changing use

Unity (parthood)
– A crowd of humans vs a group having a plan

Dependence (one instance implies the existence of another one)
– No human without a birth and no birth without a human

Modelling Best Practices
– Property inheritance, quantifiers, multiple statements on same properties
– Rich controlled vocabularies (researchers) but concise ontology (semantic engineers)
Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) – a foundational ontology designed in 2002 in the context of the WonderWeb EU project, developed by Nicola Guarino and his associates at the Laboratory for Applied Ontology (LOA) – WonderWeb Deliverable D18, 2003, p.14
∀ x ¬CF( x, 2 CT eacher, t2 ) ∧
CF( Potter, 2 CTeacher, t1 ) ∧
CF(Bumblebee, 2CTeacher, t3 )

CF( Mary, 2CStudent, t1 )

Foundational ontologies & modelling best practices

Generic, domain related core ontology

Domain related extensions

Research specific data model

DOLCE + Descriptions and Situations & object-oriented modelling principles

CIDOC CRM

Research data
Foundational ontologies & modelling best practices

Generic, domain related core ontology

Domain related extensions

Research specific data model

DOLCE + Descriptions and Situations & object-oriented modelling principles

CIDOC CRM

SDHSS

Research data
Foundational ontologies & modelling best practices

Generic, domain related core ontology

Domain related extensions

Research specific data model

DOLCE + Descriptions and Situations & object-oriented modelling principles

CIDOC CRM

SDHSS

---

CRM Archaeo

FRBRoo

CRMsoc

CRMgeo

Research data
Foundational ontologies & modelling best practices

Generic, domain related core ontology

Domain related extensions

Research specific data model

DOLCE + Descriptions and Situations & object-oriented modelling principles

CIDOC CRM

SDHSS

CRM Archaeo

FRBRoo

Society & Law (SDHSS)

Literary life (SDHSS)

CRMsoc

CRMgeo

Education & Universities (SDHSS)

Ships & navigation (SDHSS)

Research data
Research specific data model

Foundational ontologies & modelling best practices

DOLCE + Descriptions and Situations & object-oriented modelling principles

Generic, domain related core ontology

CIDOC CRM

SDHSS

Domain related extensions

CRM Archaeo

FRBRoo

Society & Law (SDHSS)

Literary life (SDHSS)

Research specific data model

CRMsoc

CRMgeo

Education & Universities (SDHSS)

Ships & navigation (SDHSS)

Projects’ research specific extensions

Research data
Foundational ontologies & modelling best practices

Generic, domain related core ontology

Domain related extensions

Research specific data model

DOLCE + Descriptions and Situations & object-oriented modelling principles

CIDOC CRM

SDHSS

Research agenda

Projects’ research specific extensions

Application profiles

Research agenda

Research data

CRMArchaeo

FRBRoo

Society & Law (SDHSS)

Literary life (SDHSS)

CRMsoc

CRMgeo

Education & Universities (SDHSS)

Ships & navigation (SDHSS)
Research specific data model

Foundational ontologies & modelling best practices

Generic, domain related core ontology

Domain related extensions

DOLCE + Descriptions and Situations & object-oriented modelling principles

CIDOC CRM

SDHSS

Research agenda

CRM Archaeo

FRBRoo

Society & Law (SDHSS)

Literary life (SDHSS)

CRM soc

CRMgeo

Education & Universities (SDHSS)

Ships & navigation (SDHSS)

Projects’ research specific extensions

Application profiles

Interoperable research data

Research data
Research Project creates Application Profile

Application Profile groups CRMbase CRMarcheo CRMgeo

subset of classes and properties

Domain specific subclasses and subproperties

Research Project uses

Research Project uses

uses produces

Community + Experts validate

Ontology Management Environment

ontome.net
4.

Building an ecosystem
of CIDOC CRM extensions
in order to produce reusable research data
in the Humanities and Social Sciences

The content: SDHSS
Semantic Data for Humanities and Social Sciences (SDHSS) CIDOC CRM Top-Level Extension

Description:

Published by Francesco Beretta (CNRS/Université de Lyon), 7 December 2020. Last revised on March 30 2021. (CC BY-SA 4.0)

The extension of CIDOC CRM for semantic data for humanities and social sciences (SDHSS) stems from the need to conceptualise the reality in the world, and more specifically factual information, from the point of view of historical research. The ontological commitment is therefore related to the domain of discourse of history but insofar as history, as a discipline that studies the life of humans and societies in the past, is interested in all the different aspects of social, economic, political, religious, literary and cultural life, the scope of this extension could be defined as the whole of social and human life, apprehended from the descriptive point of view, and global approach to reality, that characterises historical research.

This definition of the scope or domain modelled is based on the conviction that in a constructivist approach of scientific knowledge, a conceptualisation and data model can only be developed from the point of view of a specific discipline because scientific objects do not exist in the absolute but depend on the method and research agenda. They depend on the perspective or epistemic context researchers adopt in considering states of affairs: scientific objects, and semantic models modelling them, are not declared to be the only appropriate and exclusive representation of things in the pre-Kantian sense but defined as intentional objects constructed from the point of view of a discipline and methodological approach in relation to things in the world. Scientific objects are not the things in the world themselves, even if they must necessarily refer to them by way of observation or experimentation, if a scientific and therefore realistic approach is to be maintained. This corresponds to the notion of inter-objectivity in social sciences relying on the distinction between things in themselves and things as perceived, experienced and discussed by human subjects, in their shared intentionality and in relation to their social practices and context.
- **crm:E2 Temporal Entity**
  - **crm:E4 Period**
    - **crm:E5 Event**
      - **crm:E7 Activity**
  - **sdh:C2 Entity Quality**
    - Length of a bridge
    - Color of a bicycle
  - **sdh:C2 Epistemic Situation**
    - Weather in Paris in March 2024
    - Economic activity of France in 2023
How to conceptualize observable mental and social phenomena?

« In philosophy, intentionality is the power of minds and mental states to be about, to represent, or to stand for, things, properties and states of affairs. To say of an individual’s mental states that they have intentionality is to say that they are mental representations or that they have contents. »
https://plato.stanford.edu/entries/intentionality

« Collective intentionality is the power of minds to be jointly directed at objects, matters of fact, states of affairs, goals, or values. [...] Collective intentional attitudes permeate our everyday lives, for instance when two or more agents look after or raise a child, grieve the loss of a loved one, campaign for a political party, or cheer for a sports team. They are relevant for philosophers and social scientists because they play crucial roles in the constitution of the social world.
https://plato.stanford.edu/entries/collective-intentionality
Social representations

Individual minds

Social representations

Individual minds
Geovistory

Virtual Research Environment for Humanities and Social Sciences

Featured Projects

Tagebücher Anna Maria Preiswerk
Digitale Edition der Tagebücher der Anna Maria Preiswerk-Iselin (1758-1840).
Open →

ANR Processetti
Les Processetti : Migration et mariage à Venise au 16ème/17ème siècle.
Open →

Maritime History
Historical information about the Dutch East India Company, ready to explore and re-use at your hand.
Open →

Roma's deportation
Individual trajectories, and collective fates.
Open →

Geovistory : a multi-project, multi-disciplinary VRE that allows to produce LOD using SDHSS application profiles

Francesco Beretta, 20 March 2024  CC BY-NC-SA 4.0
Geovistory: a VRE for HSS research

Francesco Beretta, 20 March 2024  CC BY-NC-SA 4.0
Multi-project distributed information system architecture v. 0.2

ontome.net

FAIR

ontome.net

sdhss.org
Conclusion

• The SDHSS project is a way of promoting the use of CIDOC CRM to produce research data in HSS based on a robust methodology.
Conclusion

• The SDHSS project is a way of promoting the use of CIDOC CRM to produce research data in HSS based on a robust methodology.

• It aims to avoid the creation of new, more or less redundant, less interoperable CRM extensions for project or business needs.
Conclusion

• The SDHSS project is a way of promoting the use of CIDOC CRM to produce research data in HSS based on a robust methodology.

• It aims to avoid the creation of new, more or less redundant, less interoperable CRM extensions for project or business needs.

• It aims to map the CIDOC CRM+SDHSS ontology to existing relevant standards (LRM, RiC, Wikidata, etc.).
Conclusion

- The SDHSS project is a way of promoting the use of CIDOC CRM to produce research data in HSS based on a robust methodology.
- It aims to avoid the creation of new, more or less redundant, less interoperable CRM extensions for project or business needs.
- It aims to map the CIDOC CRM+SDHSS ontology to existing relevant standards (LRM, RiC, Wikidata, etc.).
- It aims to build a multi-disciplinary, community-driven, extensible ontology ecosystem suitable for the production of re-usable HSS research data.
Join the project, your contribution matters!

sdhss.org

dataforhumanities.org

geovistory.org

LOD4HSS
Linked Open Data for the Humanities & Social Sciences