CIDOC CRM Training Material – 56th CIDOC CRM SIG Meeting

GB gave an introductory [presentation](https://docs.google.com/presentation/d/14gTGtofXtaAPJqNVRVGX6NWm_TAewko3qnKR8wcSv6w/edit?usp=sharing) concerning the overall purpose of the day’s activities. The presentation sums up the [document](https://docs.google.com/document/d/12tk5Gi6nAAYDQBJmOcVQvBJGaP4LqFSlGkYX_3RfbXE/edit?usp=sharing) that was circulated through the listserv and aims at redesigning the educational material around CIDOC CRM taking into account the learning needs of different kinds of audiences, and follows the modular structure proposed therein.

Following the presentation an open-ended discussion concerning each proposed module took place.

A summary of the discussions and relevant decisions immediately follows.

## Concerning the educational material target audience(s)

Enhance the target audiences taking into consideration the needs of

* data publishers vs.
* data consumers
* trainers
* universities (educators/lecturers vs. students)
  + differentiation in learning needs for specific groups of students: MA students, PhD candidates, etc.

Make sure to outline the specific/desired learning outcome of completing each training session.

Design material according to the time constraints typically observed for each identified group (f.i., PhD candidates can allocate a lot more time to studying something compared to MA students).

* Create a set of instructions on integrating an educational material into a learning context/scenario. Different sets of instructions depending on whether a seminar is taught in a week, a month, a semester, etc.
* Create use cases and associate them to user roles and learning objectives:
  + pathways relevant for expressing specific modeling constructs,
  + data transformations,
  + data analysis /data description using a formal ontology

Identify relevant material and contextualize it according to learning objectives and target-audiences.

## Concerning the language of the educational material

In principle there is no constraint with respect to the language that the training material is offered in. We expect the majority to be in English, but the CIDOC community is so linguistically diverse that we also expect there to exist training material in many other languages as well –especially the ones that have active translation initiatives such as French, Spanish, Chinese, Persian.

* Training material available in English should be available for translation in other languages. The translation of the training material can be informed from the process of translating the CIDOC CRM, which called for splitting the specification document into smaller, cohesive bits, in accordance with their overall purpose and content.
  + Training documents should be marked-up in a way that supports their translation.
* Documentation and bibliography relevant for the French translation has been gathered in a wiki and a Zotero group.
* **Things to consider:** the role of the SIG with respect to the training material translation WGs needs to be determined: does the SIG
  + offer validation for the translated training material?
  + provide the original training material (to be translated in other languages of interest)?
  + collect and display the material that has been independently drafted in the context of each translation initiative?
* Depending on the institution, you can have very little regulation wrt goals etc., except for material that sits at the very core of the CRM.

## Promoting the educational material through the CIDOC CRM website

Aggregating all the relevant material calls for an extensive reorganization of the website and a redesign of the currently dubbed “Use & Learn” section.

## Segmentation – training modules

### Original proposal:

1. Formal Ontology Skills Appropriation
2. Conceptual Modeling
3. Semantic Data Mapping and Transformation
4. Platform Implementation
5. Semantic Data Query
6. Ontological Engineering

### Revised proposal [[NAMES OPEN TO RELABELING]]

1. **Data literacy**An *Introduction to data modeling* module to sit on top of the list. Provide answers to questions:
   * why is formalizing one’s data useful?”
   * what is data integration? what standard can one use for data integration?
   * compare and contrast with AI approaches
   * Knowledge organization systems, distinguish thesauri and controlled vocabularies from ontologies etc.
     + **HW**: CEO in charge of providing suchlike material
2. **Formal Ontology** 
   * Basics of RDF section:
     + It is about community practices, but does not logically fall under the Formal Ontology. It’s an implementation mechanism, should not be bound to its semantics when one is explaining what a formal ontology is.
     + Previous training experiences show that using rdf to teach ontologies to audiences with little to no computing skills has proved to be extremely difficult (TA, MR).
     + Ariadne+ has offered seminars & summer courses but has consistently kept the rdf expressions apart from the conceptual modeling
     + Introducing rfd early on may work well with some individuals, it doesn’t have to be like that with every trainee. There are alternatives available to making the fundamental concepts concrete (CRM card game, using spreadsheets, rdf, description logic) that should be used for non-technical audiences. But redesigning the training material should allow and encourage audiences experiment with data (in whatever format they’re comfortable using).
     + The data examples should be useable across modules, and should be offered in different forms (text, cards, spreadsheets, rdf). They should also cover multiple fields (architecture, archaeology, art, history, conservation, whatnot). The problems to be addressed with the help of the data should cross disciplinary boundaries (from the point of the data).
   * **Decisions**:
     + The section will include the following subsections (Basics of Ontology, CRMbase & Extensions at a high level)
     + The proposed “Basics of Serialization” section is going to be made a subsection of the formal ontology section.
       - The rdf implementation (and other serializations) should appear in the section relevant for Mappings.
3. **Data Representation** ~~[Conceptual Modelling]~~
   * **Overall comment:** maybe one doesn’t have to go through all the modules one-by-one
     + An idea is to start at the top level of the CRM indicate which constructs are necessary to move forward, and then introduce audiences to an application profile that they would need to use in a certain training context.
   * Basics of Modelling| **Basics of Data Representation**
     + Start the section with entity-relation modelling and then move up to object-oriented.
     + The material should contain examples that reference the vocabulary used in CIDOC CRM ((multiple) inheritance, multiple instantiation, property quantification, etc.) in a way that ensures that audiences fully understand it. The “terminology” section of the specification document is useful in this respect.
       - Seems a better fit for the first module (Data Literacy, or whatever it ends up being called).
       - All examples should cross every training module to the extent that’s possible. Each module comes with a set of training objectives and learning purposes, that make use of different methodologies. Having a consistent set of training material across modules highlights different perspectives through which you can explore data.
       - **GH** has adhered to this approach in implementing a set of training materials for an rdf creation pipeline (very basic ontological modelling making use of a very small set of classes/properties that are absolutely necessary to describe “articles, authors, titles, subject: a place”).
   * **Learning and using existing/documented patterns**
   * **Decisions**:
     + Implement this module. It involves teaching basic data modelling and then introducing audiences to application profiles/semantic data models.
     + Relabel it to “Data Representation”.
4. **Semantic Data Mapping and Transformation**
   * PF has started working on a complete use case for maritime history, which he shared with the SIG. It can be extended to all modules. The workflow example for mappings can be found [here](https://docs.google.com/document/d/1JqzqtVXZ5EliSPXafVet7mwlalpp2yT-Uzy1G5ipx-E/edit?usp=sharing).
     + The example does not sit well in this module, it requires too much background knowledge and implicitly talks about creating an extension for the CIDOC CRM.
     + The example maps to all identified modules and showcases how mappings are parts of a broader project involving a lot of feedback and back btw different groups. Changes in data entail changes in the mappings. In that sense, managing mappings probably falls under a separate module for Semantic Data Management (revive Synergy as well).
   * For this particular module, the training material could also reference
     + the Herbarium examples that MD has mapped
     + the “Data transformation (using 3M/X3ML)” part of the example that PF has shared
   * GH has shared a [tutorial](https://drive.google.com/open?id=1mlHc3Uhu8qFKhB2OyyUK4Y4bXaMSZFgo&usp=drive_fs) he’s been working on, for the SIG to determine if it sits well with the material will be drafted/pointed to. It follows an approach similar to that of the workflow proposed by PF. Can be split to chunks and used for different modules.
   * MK has shared a set of guides and good practices has shared with the SIG as well (see [here](https://archaeologydataservice.ac.uk/help-guidance/guides-to-good-practice/)). It is an ASD document describing case studies from the subject field of archaeology (uses different implementations carried out by different groups etc.).
   * Mappings can be seen as meaning negotiations, in the sense that they involve different levels of expressivity, granularity, etc. Mapping decisions involve a number of methodological issues that need to have some theoretical grounding as well. AG is working on a document trying to formalize this concept.
     + Technical problems encountered in mappings but irrespective of the tool one’s using to do the actual mapping explained/discussed at an abstract level.
     + It ties closely to the question of who is responsible for creating mappings.
   * **Decisions**:
     + Determine the content and position in training pipeline of the module for Semantic Data Management (it’s more than just data transformations).
       - Illustrative use-cases in increasing complexity (start from very simple toy examples before fully fledged scientific research scenario are presented to audiences).
         * **PF** will be working on that,
         * Other use cases involve:

The mappings by MD for Herbarium,

the rdf creation pipeline by GH,

the ASD document describing case studies from the subject field of archaeology shared by MK.

* + - Technical mapping considerations: introduce a separate subsection in this module.
    - The rdf implementation (and other serializations) should appear in the current training module.

1. **Semantic Data Query**
   * Making use of the analysis of scientific research questions as a resource for translating end-user questions into complex queries, assessing data quality etc.
     + Rearrange the order for now.
2. **Selecting and Implementing Semantic Information Systems**
   * Drafting a requirements document for implementing a KB (evoking the training material of CIDOC) to allow specific communities to make informed decisions regarding existing platforms
     + Different systems can be used to complement one another, it’s not about using one piece of software exclusively
       - **HW**: SdS
   * Classifying platform types is also required.
     + Data access platforms
     + Triplestore / QuadStore / Graph Database
     + Semantic Data Management Platform
     + Research Databases / Infrastructure System
     + Aggregation Systems
3. **Making and Extending Ontologies** ~~[Ontological Engineering]~~
   * What kinds of statements about the real world one can make using this method of knowledge representation that an ontology stands for.
   * The redrafting of the *Principles* document (including the checklist for proposing new class/properties and extensions)
     + Annotations of classes and correspondence to principles they conform to.
   * Exploring the relation of extensions to CIDOC CRM (i.e., the fact that they stand for specializations of the CRM in a particular subject field) and developing new extensions falls under the “Ontology Engineering” module, not Data Representation/Conceptual Modelling.
     + History of CRM family models evolution can be handy (wrt failed modeling decisions etc.)
       - 10-year discussions for “Archaeological Finds”
       - Coreference model (that was ultimately abandoned)
         * MD can provide material, also consider deprecated classes from CRMbase and family models.
     + Consider relabeling “Ontology Engineering”. Alternative labels (tbd):
       - Extending the Ontology
       - Making and Extending Ontologies,
       - Ontology Specializations
4. Semantic Data Management (Synergy Model)
   * Not yet specified –about the data integration lifecycle in general

### Discussing data examples to be used for training purposes across training modules

* MD & ETz have been in contact with Prof. Sarantos Kapidakis of the West Attica University. They have experimented on data examples by one of his PhD candidates on History of Art, where the diagrams that she used were automatically transformed to rdf. They do come with errors, but they can be easily edited. The dataset could be published, it would make a good example.
* PF to provide examples from SeaLiT (they have been working on that already) and it represents a complete research scenario, spanning across training modules.
* Building blocks for the examples should consist of the following:
  + Data (f.i. from SeaLiT or some other use case of an extension model)
  + Modelling using CIDOC CRM (possibly some extension too)
  + Data Transformation (schema mapping, URI policies)
  + Data Access
* Characteristics Required for a Base Example Dataset:
  + Integration aspect (at least two files)
  + Information needs (questions)
  + Contextual Explanation of Data (the narrative)
  + Data Dictionary of the fields in the dataset
  + Provenance of the dataset (ie who makes it, why they make it)
  + Paradata - digital provenance
* Known example datasets:
  + SeaLiT
    - **Disciplines**: maritime history, trade, archival data
  + Census
    - **Disciplines**: History of Art & Architecture (Classics & Renaissance)
  + OpenArchaeo
    - **Discipline**: excavation archaeology
  + Funerary Archaeology – MASA Consortium
  + Linked Conservation Data
    - **Discipline**: conservation science
  + Notre Dame
    - **Disciplines**: architecture, archaeology, interdisciplinary scientific data integration
  + Ricontrans
    - **Disciplines**: art, history, political science
  + Doremus
    - **Discipline**: music
  + LRMH/C2RMF mappings
    - **Discipline**: conservation of buildings & artworks
* Dataset examples from disciplines that could be added to the overall set:
  + Natural heritage science
  + Biodiversity (FORTH)
  + CNBR & sampling strategies (FORTH)

## Overall planning and assigning HW & tasks

Work on the training modules piecemeal, not at once/in parallel. It’s better to make enough progress in one or two module and use them as templates. Bring them to the next SIG meeting to be discussed.

Start with modules **(3)** and possibly **(4)** below.

1. **Data Literacy** –contributors: Mélanie Roche
2. **Formal Ontology** –contributors: Olivier Marlet, George Bruseker
3. **Data Representation** –contributors: George Bruseker, Anaïs Guillem, Markos Katsianis, Olivier Marlet, Christian-Emil Ore
4. **Semantic Data Mapping and Transformation** –contributors: Gerald Hiebel, Athina Kritsotaki, Martin Doerr, Anaïs Guillem, Markos Katsianis, Olivier Marlet
5. **Semantic Data Querying** –contributors: Pavlos Fafalios, Martin Doerr
6. **Selecting and Implementing Semantic Information Systems** –contributors: Stephen Stead
7. **Making and Extending Ontologies** –contributors: Thanasis Velios, Martin Doerr
8. **Semantic Data Management –**contributors: tbd

**HW**: GB to contact key-people individually.