

GUIDELINES for A LODLAM PROJECT

Students are asked to document their work in a report (preferably a Web page) in which they must demonstrate the ability to do as follows:

- choose and study an observed domain (*study of the domain*);
- elaborate models to describe the domain (*knowledge organization: elaborate models*);
- produce new semantic data (*knowledge representation: create data*).

These guidelines have to be thought of as a workflow that moves from real data to the level of abstraction, from interpretation of the observation to the conceptualization of the theory, and from data modelling back to data representation.

Study of the domain

Idea

- The requirement is to **focus on an idea** which could be the subject of a Wikipedia page, such as a notable person, a significant place, a concept, a historical event or a special date. The idea has to be in some way **related to different kinds of objects**, e.g. archival documents, bibliographic entities, or artefacts, on the basis of a variety of relations (e.g. assuming the idea is a person: s/he can be the subject of a book, the creator of an artwork, or an actor in some event). It's recommended that the scenario description (including idea, objects, and relations) be written down, and that data of interest, external links, references, "see also" links, and authority control identifiers available in the Wikipedia page are selected. The advice is to group links and follow such links so as to understand what is interesting for describing the scenario (e.g. people associated with the idea, cultural institutions preserving related objects, other events which happened at the same time, significant historical places).

Items

- After the scenario description, the task is to **select 10 different types of items** related to the idea, by mixing objects: archival documents/files/series/founds; bibliographic records; artefacts (photographs, maps, sculptures, buildings, manuscripts, sigils, etc.). Some other useful details:
 - objects must be already described on the web as items;
 - descriptions must be provided by cultural institutions, e.g. bibliographic records included in a OPAC, cataloguing records part of a museum/archive/library catalogue, or included in relevant data collections or aggregators, such as Europeana, WDL, SIGECweb, SAN, CulturalItalia, Worldcat, Internet Culturale, and so on (but, in the case of aggregators, mining the description directly from the provider is advised)¹;
 - the description of each item must include information about different kinds of entities: people, places, dates, and subjects/concepts.

Knowledge organization: elaborate models

Conceptual map and E/R model

- The first step in modelling is to explain the idea/scenario through a **conceptual map**. This map needs to represent the real data as extracted from the chosen items, together with the explicit relationships.

¹Europeana, <<https://www.europeana.eu/>>; WDL - World Digital Library, <<https://www.wdl.org/en/>>; SIGECweb <<http://www.catalogo.beniculturali.it/>>; SAN - Sistema Archivistico Nazionale <<https://san.beniculturali.it/>>; CulturalItalia <<http://www.culturalitalia.it/>>; Worldcat, <<https://www.worldcat.org/>>; Internet Culturale, <<http://www.internetculturale.it/>>.

- It's then possible to translate the conceptual map into an **E/R model** (entities, relationships and attributes) or in any other formalism. The production of a graphical representation of the abstract entities that characterise the scenario (people, places, dates and subjects/concepts) and the relations among the latter is recommended.

Metadata analysis and alignment

- The analysis of metadata can be conducted by identifying which **metadata standards** (content/structure) are **used by the institutions** for describing chosen items. e.g. the archival document is described according to EAD/XML schema, the photograph is described by means of ICCD-F content standard, the painting is described according to CCO cataloguing rules, the bibliographic record is described according to ISBD, the codicological description of a manuscript is encoded in a TEI/XML document, and so on.
- Another important step is the **alignment**. What is needed, is to align those metadata elements and properties that are useful to address information related to people, places, dates, and subjects/concepts. DC and DCTerms could be used as a guide for defining the categories, and missing elements must be added in light of the descriptive features of the chosen items.

Theoretical model

- A key step is to move to modelling by adding a new fundamental level, the one of data interpretation, the latter being the result of observation of the items and their potential connections. The activity is to **produce a theoretical model** expressed in natural language, which is able to describe all of the selected items and related information that could be added, by searching on the Web for other details related to the chosen entities. Wikipedia, Wikidata and Dbpedia are good sources, but students are advised to use a wider range of resources.
- Specifically, the model has to **answer the following questions**:
 - Who – people. Which kind of information is related to people? e.g. the aforementioned roles related to the lifecycle of a cultural object. Which kind of properties could enrich the description of a person?;
 - Where – places. What kind of information should be represented about locations? What notion of place is to be conveyed?;
 - When – dates. What data types can represent the notion of time? Which dates are important for describing items?;
 - What – subjects/concepts. What is the scope or content of the object? Is there a special event that that should be emphasised? Which keywords and classification features are useful in enriching the description of the content of the items?
- At this stage, it's possible to **refine the original E/R model and to create a new graph**, on the basis of the personal theoretical model, by adding new entities and relationships able to describe the enhanced observation of the domain.

Conceptual model

- Having taken these steps, it's possible to then move from the theoretical model to the **formal representation**, i.e. the ontological approach. A conceptual model must be created by reusing existing schemas, vocabularies, and ontologies (RDF/RDFS/OWL/SKOS; DC, DCTerms, FRBR, EDM, RDA, OAD, EAC-CPF)². Terms of the theoretical model must be essentially refactored with terms taken from existing models.
- In order to express the conceptualization activity, a **graphical representation** of the formalization (e.g. with grafoo, the graphical framework for OWL ontologies) must be produced.³

Knowledge representation: create data

- Once the modelling activity has been completed, the process requires a return to **items** in order to **describe** them on the basis of the conceptual model designed.

²For a complete overview on LAM ontologies, see Marilena Daquino, *Mining Authoritativeness in Art Historical Photo Archives: Semantic Web Applications for Connoisseurship*. Amsterdam: IOS Press, 2019.

³Grafoo, <<https://essepuntato.it/graffoo/>>.

- One **table** is produced for each item (.csv files), where columns correspond to triples subject/predicate/object (natural language for subjects and objects; predicates in the form of properties as selected from the chosen ontologies) and rows are the statements.

RDF production

- The first step is **naming**. What is needed, is to mint URIs for representing all the entities of at least one of the items (even if instances are not referenced).
- The next stage is to define **prefixes for the chosen ontologies** and to write triples in RDF (in one of the possible serializations), trying to connect entities related to the different perspective: person, place, date, and subject/concept.
- **Links to other related items** or entities must also be included. Links to:
 - the same authorities (*owl:sameAs*) for names (e.g. people, places, subjects);
 - other significant resources, related because they share one of the features of the selected item. The requirement is to identify up to 5 associations by using different types of semantic association (e.g. hierarchical, partitive, associative for people/places; broader, narrower, related, synonyms for terms).

RDF visualization

- The very final activity is the **visualization of the realized knowledge graph** (e.g. with RDF grapher⁴) and its addition to the documentation as a final example of the workflow.

A possible scenario for the LODLAM process

In order to better understand these guidelines, an idea of a possible project in LODLAM will now be presented, together with the expected steps of the process. With Lease Morgan, when doing a LODLAM work, it has to be remembered that: «moving forward with linked data is a lot like touring to Rome. There are many ways to get there, and there are many things to do once you arrive, but the result will undoubtedly improve your ability to participate in the discussion of the human condition on a world wide scale»⁵.

IDEA

Creating the scenario: Dante's illuminated Florentine manuscripts of the XV century held in Italian libraries.

ITEMS

Searching for items: a selection of manuscripts (images with descriptions), some paintings of Dante, portraits of illuminators and copyists, a historical map of Florence, a letter as an archival document where we find the description of the production of a manuscript, a book on the culture of the XV century, etc.

CONCEPTUAL MAP

Working on data: the manuscript "Riccardiano 1035 of the Divina Commedia" *held in* "Biblioteca Riccardiana" *located in* Florence *has been realized in* 1450 and *has been copied by* a person named "Giovanni Boccaccio" *depicted in* "Giovanni Boccaccio, from the cycle Famous People" *realized by* "Andrea del Castagno", etc.

E/R MODEL

Abstracting the view: a Manuscript (with attributes, i.e. metadata) *is written by* a Person (with attributes, i.e. metadata) *depicted in* an Artwork (with attributes, i.e. metadata) *held in* a Museum (with attributes, i.e. metadata), etc.

METADATA ANALYSIS

Studying standards: the institutions that held items which are to be used TEI for the codicological description of the manuscript, CCO for describing the paintings, DC for setting the categories useful for the map, EAD for the letter as an archival document, ISBD for the catalographic description of the book, etc.

⁴RDF Grapher, <<http://www.ldf.fi/service/rdf-grapher/>>.

⁵ Eric Lease Morgan and LiAM, *Linked Archival Metadata: A Guidebook*, version 0.99. April 23 2014, <<http://sites.tufts.edu/liam/>>.

ALIGNMENT

Mapping elements: People (e.g. as agent): DC:creator = EAD:creator = ICCD-F:author = ISBD:statement of responsibility = TEI:author, TEI:person, TEI:editor, etc; Places (e.g. as place of creation): DC:coverage = ISBD:place = TEI:placename; Dates (distinguish between the dates at the different levels of observation): DC:date; Subjects: DC:subjects.

THEORETICAL MODEL

Enrichment design activity: about people (Dante, Boccaccio, Andrea del Castagno), it may be possible to focus on, e.g., prosopographical data, roles, relations with objects they are creator of, authorities, other related people, etc; about places (Florence), it's possible e.g. to identify them on a map, to geolocalize, to connect with people, etc.; about dates (1450) it's possible e.g. to work on other important events happened in the same period, to connect dates with an external timeline, etc; about concepts (copying, manuscript, artworks) it's possible e.g. to link an external lexical semantic network, to find the controlled form of the name in an authority, to link keywords to artefacts, etc.

E/R MODEL ENHANCED

Original model enrichment: new entities and relationships could be added: A Person was born in a Place; A Manuscript was produced by a Person; An Event took place in a Date; etc.

CONCEPTUAL MODEL

Ontology design: identification of the pertinent ontologies in order to express classes and predicates according to the theoretical model. eac-cpf for Person, crm for production, edm for event, pro for roles, prov-o for provenance, frbr for levels, etc.

URIs DEFINITION

Creation of personal unique identifiers, e.g. from w3id.org.:

<http://w3id.org/myproject/person/dante-alighieri>;

<http://w3id.org/myproject/event/manuscript-production/riccardiano1035>;

<http://w3id.org/myproject/time-span/1450>;

<http://w3id.org/myproject/object/portrait-boccaccio-del-castagno>

RDF PRODUCTION

Writing of some statements, with @prefix and triples (here in Turtle serialization), e.g:

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
```

```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
```

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
```

```
<http://w3id.org/myproject/person/giovanni-boccaccio > a crm:E39_Actor ;
```

```
  rdfs:label "Giovanni Boccaccio" ;
```

```
  owl:sameAs <http://viaf.org/viaf/64002165> . # URI for 'Giovanni Boccaccio' from VIAF
```