WoodsLOD project documentation

The idea

We decided to realize an event based LOD project about an event that is nowadays seen as a moment of great changes are revolutions for the contemporary Western society: The Festival of Woodstock. Due to the variety and complexity of the elements and concepts involved in this mythic "3 days of Peace and Music", we thought it could have been interesting to deepen its knowledge with the aim of retrieve and bring to light the network of relationships and interconnections that holds together its participants, its long-lasting charm and generally its heterogeneous and sometimes controversial aspects.

The items

Firstly, we chose 10 items from different institutions in the LAM domain. Indeed, they cover a huge variety in the field of libraries, archives, music and movie catalogues and national museums. All of them were already described in the web by their holding cultural institutions or by more general providers that we listed and started to analyse to understand which kind of standards of descriptions they used, which typology of information they provide and generally what point of view they give about the record in consideration. The items involved in our project are:

- *Woodstock*, a documentary film from IMDb
- *1969*: the year that everything changed, a Robert Kirkpatrick's book from the Library of Congress Catalogue
- *And Babies*?, a poster from the Smithsonian American Art Museum
- *Abbie Hoffman Shouting*, track audio from MusicBrainz the Music Enciclopedy Catalogue
- *Woodstock ticket*, a historical object from the National Museum of American History
- *Audience near the stage at the Woodstock Festival*, a photograph from the Special Collections and University Archives at the University of Massachusetts Amherst Libraries
- *Woodstock Music from the Original Soundtrack and More*, a vinyl disk from the MusicBrainz Catalogue
- *D'oh-in the Wind*, a Simpson Tv Episode from IMDb
- *Fender Stratocaster*, a historical object from Europeana.

All their description, nonetheless in many different ways and expressions, including information about people, places, dates and concepts involved in each specific item.

We decided to organize all this information in the entities involved in our project concerning:

- People: Eddie Kramer, Jimi Hendrix, Janis Joplin, and Abbie Hoffman
- Places: Bethel, Woodstock, Bethel Woods Art Centre
- Time entities: the year 1969 and the period of the three (actually four) days of the festival
- Concepts: the hippie movement and the Vietnam war protest.

The E/R model

Then, we explained our scenario through the realization of an E/R model able to highlight all the relationships between our items and entities. An E/R model can be defined as a graphic representation of natural language definitions about the entities and the relationships among them related to our items. To realize it we had to distinguish between our items, conceived as the concrete objects related to our event kept in different institutions; the entities involved in our project, defined as the people, date, places and concepts that can be referred to our event and can link together two or more items; and the relationships among them. We used colours to distinguish items from entities and arrows to represent relationships among them.
Due to the complexity and variety of the items and entities involved in our project, we had to redefine the E/R model iteratively until we found the best solution to represent completely and homogeneously all the relationships among every instance and the main event in consideration.

The identification of the metadata standards.

Metadata are data that provide information about other data, allowing some particular strings of characters to become information about other ones – they are, in other words, specific data about data. We can have a different typology of them: descriptive metadata (about the identification and discovery of a resource), structural metadata (regarding the structure of a resource), administrative metadata (with information about the administration of the resource) and others. In the cultural heritage and LAM domain, metadata identify the controlled and normalized description of the features characterizing a CH object. A metadata standard can be defined as a requirement intended to provide a common understanding of the meaning of the data by specifying their metadata. Metadata standards are sets of rules and constraints aimed at providing a commonly agreeable and understandable way to intend a specific feature of an object. In our project, we took in consideration mainly descriptive standards: sets of rules, constraints and methodologies used by cultural institutions to describe the features of our items. We analysed the standards holding institutions for our items and

realized a table able to summarize which metadata standard was used by each of them in the description of the item:

#	Title	Object	Provider	Metadata
1	Woodstock	Film	IMDb	SOMA
2	1969: the year that everything changed	Book	Library of Congress	MARC-21
3	"And Babies?"	Poster	Smithsonian American Art Museum	CIDOC-CRM
4	Abbie Hoffman Shouting	Track audio	MusicBrainz	MMD XML Schema
5	Woodstock Tickets	Ticket	National Museum of American History	CCO
6	Audience near the stage at the Woodstock Festival	Photograph	Special Collections and Univeristy Archives, University of Massachussets Amherst Libraries	MODS
7	Woodstock: Music from the original Soundtrack an More	Vinyl	MusicBrainz	MMD XML Schema
8	Woodstock: Music from the original Soundtrack an More	Vinyl	MusicBrainz	MMD XML Schema
9	D'oh-in' the Wind	TV Episode	IMDb	SOMA
10	Fender Stratocaster	Guitar	Europeana	EDM

The standards involved in our project are 8:

- SOMA = Shared Online Media Archive, SOMA, is a draft metadata standard for the exchange of metadata for multimedia files, based on Dublin Core 1.1 and EBU Tech 3273 (Colorimetric Performance). SOMA is a collaboration between several NGOs to create an online media archive for use by community media centres.
- MARC-21 = It's the most used among a family of standards designed in the 60s by the Library of Congress as machine-readable cataloguing standards for the description of items catalogued by libraries. By the 70s they became the US national standards and then employed as international ones. MARC-21 describes three main aspects of the record: field designation, record structure and content.

- CIDOC = The International Committee for Documentation is a committee of the International Council of Museums. The CIDOC Conceptual Reference Model(CRM) provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation.
- MMD XML = The MusicBrainz XML Metadata Format (MMD) is an XML based document format to represent music metadata. It has been designed to be easy to read, powerful and extensible. MMD is the official successor of the old RDF-based metadata format, which was popular among semantic web enthusiasts, but didn't have much acceptance otherwise because of its perceived complexity.
- CCO = Published by the American Library Association (ALA) in June 2006, CCO provides guidelines for selecting, ordering, and formatting data used to populate catalogue records based on core categories in CDWA and VRA Core. CCO is a set of rules surrounding various elements from CDWA (which contains elements and rules) and VRA Core (which contains elements); it is more directly analogous to AACR and DACS.
- MODS = The Metadata Object Description Schema is a schema for a bibliographic element set that may be used to carry selected data from a subset of the MARC 21 records as well as to enable the creation of original resource description records.
- EDM = The Europeana Data Model (EDM) is a new approach towards structuring and representing data delivered to Europeana by the various contributing cultural heritage institutions. The model aims at greater expressivity and flexibility in comparison to the current Europeana Semantic Elements (ESE). The design principles underlying the EDM are based on the core principles and best practices of the Semantic Web and Linked Data efforts to which Europeana wants to contribute. The model itself builds upon established standards like RDF(S), OAI-ORE, SKOS, and Dublin Core.

Metadata alignment

Once individuated all the metadata standards involved in our scenario, we proceeded to find correspondences between them to align the different choices by the institutions in the description of the common features of our items. Indeed, especially in the LAM domain, there is a strong relationship between standards and institutions. Each institution – at least from an abstract point of view – hold a specific kind of object, with specific features that need to be described. Hence, more or less each institution develops a particular way to describe it due to the particular aspect of the record. Nonetheless, often happens that an institution has to deal with different objects from

the ones it is expected to (for example, a library may have necessity to catalogue a photograph, for which the bibliographic description standard is not perfectly suitable). Thus, it's quite common to meet missing aspects or overlapping points in the metadata description of each cultural institutions. The reality of cultural institutions situation in dealing with record descriptions is thus an extremely heterogeneous and sometimes even confused environment that lack of interoperability. A possibility to overcome this situation is given by some shared standard and methodologies (such as DC), but a fully satisfiable solution will be realized only by the Semantic Web or Web 3.0 based on the explicit semantic interconnection of uniquely identifiable resources. The metadata alignment step in our project wants indeed to highlight and at the same time overcome the differences and incompatibility of the different standards involved in the description of our items by finding the correspondences among all the descriptions. Furthermore, in realizing it, we classified all the useful information provided about our items regarding their relationships with people (who), places (where), date (when) and concepts (what) we already underlined in our E/R model.

Theoretical model

Once we had compared and analysed the institution and most of all the metadata standards involved in our scenario, we could proceed to the real Knowledge Organization process. The first step was to elaborate more abstract models of our data and their relationships. We needed to abstract our data and their relationships: we built a theoretical model able to describe all the selected items and related information in a more general way than the initial E/R one. Our theoretical model is still defined by the natural language, even if based on more general terms able to include and describe all the data involved in our project and the information about them. In particular, it is focused on the representation of the information about our data related to:

- Who: the people = how people contribute or are involved in the lifecycle of our items? How are they linked to the main event?
- When: the temporal entities = which are the temporal entities involved in our main event? What data type do they include?
- Where: the places = which kind of locations are involved in our main event? What data type? How can we describe them?
- What: the concepts = are there concepts involved in our main event? How are they connected with our items?

We chose to represent our theoretical model using Yed. For a more clear distinction of the entities involved in our main event, we decided to distinguish between entities and sub-entities (sub-classes of the first one which represents a mid-level of abstraction),

linked together by relationships among them and with the main event, and represented by our chosen items.

Conceptual model

A further level of abstraction was represented by the realization of a conceptual model able to answer the same questions raised by the theoretical model more formally. We intended the conceptual model as a way to represent all the main features related to our entities through the formal language of the ontologies. Indeed ontologies, as conceptualizations of data through a formal language, could allow us to draw abstract schemas of a description of the main aspects involved in the representation of our main event. To allow as much interoperability as possible to our descriptions we chose not to create a new ontology but to re-use already existent ones available on the web. We decided to use one only ontology to describe the typology of all the entities involved: … and then use more specific ontologies to describe the single features of each entity type we need to describe – in particular, people, dates, places and concepts. Sometimes they are very general ones, some others they are specifically conceived to describe a particular domain. In our case, one of the most important and specific ones that were used is the Music Ontology that provides a rich as well as flexible vocabulary for the formalization of concepts related to the music world.

Data description

After having realized quite wide models for the formal description of the entities involved in our project, we applied them in the representation of our data through the description of the data features. This step required a selection of the classes and attributes defined in our conceptual model to realize a description of the entity in consideration able to be abstract as well as complete and precise. We realized the description of all the entities involved in our project based on the conceptual model:

- People: Janis Joplin, Jimi Hendrix, Abbie Hoffman and Eddie Kramer;
- Temporal entities: the year 1969 and the three/four-day span of the Festival duration
- Places: the two towns related to the Festival: Bethel where it took place and Woodstock after which it was named, and the actual site where it was held now transformed in the Bethel Woods Center for Arts
- Concepts: the main concepts generated by the festival or which the festival represent the most known icon of the hippie movement and the Vietnam War protests.

In representing the descriptions we realized three-column tables representing the explication of the predicate, the formalisation of the predicate through the ontology

language and the object. The subject of each triple is represented above the whole table since all the properties described are referred to it.

Data representation: the RDF statement

Finally, we managed to represent our data description through an RDF statement we decided to serialize through Turtle. The RDF statement allowed us to represent our data in the form of triples composed by subject-predicate-object, each of which formally defined by an ontology vocabulary. Also, the triple structure enabled us to realize the connections between the different data involved in our project.We decided to describe two entities involved in our project which represent a very peculiar and maybe not very deepened aspect of the Festival of Woodstock: the political activism and the anti-Vietnam War protests that took place in the "3 days of peace and music" event. From this point of view, the two entities involved are the person Abbie Hoffman and the concept of the Vietnam war protest. As RDF identifies resources through URIs, we wrote URIs for our two entities as if they belong to a real LOD website and then proceeded in the realization of the statement itself. We managed to connect our entities with:

- Term lists and authority control resources can uniquely identify in the correct form the entities involved. In doing this we mainly used the property defined by OWL owl:sameAs and the authority control by VIAF; wherever it was impossible to use this form we directly write the connected resource with its controlled form, as it happened for the location identification through the use of GeoNames.
- Other entities or items involved in our project as well as
- Other resources already present in other online repositories, mainly DBpedia. We defined semantic associations related to relations (related items or concepts), people and places (know, took place in), hierarchical connections (broader of), partitive connection (a member of).
- Other resources identified by URL already present on the web, mainly Wikipedia resources.

Also, we gave a graphical representation of our data and their connection through a little knowledge graph.

Knowledge Organization

Knowledge Organization is a discipline about activities such as documentation, indexing, description and classification of cultural resources, once performed analogically by librarians, archivists and experts in the domain that nowadays needs to

deal with the improvements and at the same time challenges provided by new technologies. KO concerns the nature and the quality of Knowledge Organization Process (KOP) and the Knowledge Organization Systems (KOS) they aim to provide. Based on the DIKW pyramid (that allows us to define knowledge in terms of contextualized information and information in terms of contextualized data to reach wisdom), its main goal is to realize an available, reliable and commonly acceptable classification of data conceived as facts – "really existing piece of something" that always refer to something or someone in the real world –to make connections among them and retrieve information that can be transformed in knowledge. Also, it's important to stress that KO is focused on the process of the classification, and only in a secondary moment on the product of this process. To perform data classification, KO deals with value vocabularies which are classified into three main categories:

a. Term lists → authority control: aimed at providing a controlled form of the name of something and at uniquely identify something in the real word – usually providing a controlled textual form or an ID.
   (example: VIAF, Library of Congress Authorities, Getty Vocabulary, GeoNames, Pleiades…).
b. Classifications and categories → both for specific and general purposes, aim at uniquely classify resources due to their content or role.
   (example of specific purpose classification: Subject heading by the Library of Congress LCSH)
   (example of general-purpose classification: DDC Dewey Decimal Classification; UDC Universal Decimal Classification; CC colon classification or faceted approach by Ranganathan).
c. Relationship lists → aimed at providing a taxonomic approach for the classification of resources, organizing them in hierarchical structures based on their relationships
   (example: thesauri, the most common form, such as AAT; semantics networks such as WordNet; ontologies).

For our purposes, ontologies are the most important tools, followed by term lists.

Ontologies can be defined as conceptual modelling of a domain expressed in a formal language. They are the conceptualization or formalization of the concept of an item inside a domain by giving a particular point of view on that domain. They allow transforming an item in a conceptual model whose description is more expressive and suitable for letting machines make automatic inferences. They take the form of ontology vocabularies which are collections of controlled forms in a formal language (readable by the machine) for defining statements about resources on the web. Technically speaking, an ontology is a description written in OWL (Ontology Web

Language). But conceptually speaking they are the tool that enables to produce linked open data by the definition of classes and predicates about resources which can be linked together in triples allowing to describe an RDF statement. We have ontology thought for specific as well as general purposes.

Semantic Web

Ontologies are a fundamental tool in the realization of the Semantic Web. **Semantic Web or Web 3.0** is the next evolution of the Web as we know it, ideated by Tim Berners Lee since 2001. It is based on the possibility to substitute the present href links between resources with **explicit semantics connections between resources that will be uniquely identified on the web**. To realize the Semantic web we need basic as well as more complex tools that can be defined through the **Semantic Web stack**: the organization of all the technologies necessary to transform the existent web in a semantic one. The Semantic stack involves:

- **Unicode**: a unique way to express character in the web
- **URI (Unique Resource Identifier)**: a standard form of the name that allows to uniquely identify everything on the web. Every resource existent on the web can be wherever identified by its URI, differently from the URL (Unique Resource Locator) that individuated instead its location on the web (URI: fiscal code = URL: postal address).
- **XML + NS + XML schema**: system-independent and device-independent ways to describe resources on the web, thank the portability of XML documents combined with Name Spaces (set of specified and controlled forms homogenously used in more than one XML document) and XML schemas (language for the description of an XML document, a sort of grammar that controls which elements are allowed, how they are connected and how they have to be indicated in an XML document).
- **RDF (Resource Description Framework) + RDF schema**: it's a tool for the identification and interchange of structured metadata on the web aimed at providing **semantic interoperability** between applications on the web. It allows describing web data in form of **triples** composed by a **subject** (what it is described/what it is being said about), a **predicate** (a property that expresses the relationship of the subject with the object) and an **object** (what is related to the subject through the predicate). An RDF schema is a way to define some of the properties we can manage to describe aspects of the triples (e.g. classes, properties, ranges). RDF enables to represent data as triples of SPO. A set of this triples is called an RDF graph, composed by nodes and directed arc. In this environment, each triple takes the form of a node-arc-node link, each

represented by a URI. RDF uses URI to identify both the nodes and the arcs, and also to realize connections among those URIs (connection URI-another URI or connection URI-Class), identifying individuals as well as kind of things (e.g. a person) as well as values of the properties (e.g. mail), authorities. In this sense, the Semantic Web can be seen as a GGG, a Giant Global Graph. There are different serializations of RDF: different ways to encode the syntax of an RDF document. The most used are RDF/XML, Turtle, RDFa, N-Triples.

- Ontology vocabularies: controlled forms for the conceptualization of the statements we can produce through RDF. To express an RDF statement we need to base this statement on an ontology able to provide an abstract as well as a formal way to define its elements: we'll need an ontology for the subject, one for the predicate and one for the object.

Semantic Web is strictly related to KO because it aims to provide to the machine or to the user the possibility to make inferences about the information given, and thus realizing knowledge. Usually, its products take the form of a knowledge graph, a particular data structure able to explicit the semantic links between the different resources. Semantic Web it's based on logic (the possibility to make inferences based on the Aristotelic approach), provides proof (the possibility to prove the validity of the statements made) and aims at realizing trust (the possibility to believe in a statement because it's perceived as derived from authoritative information).

LOD

Linked Open Data are fundamental in the realization of the Semantic Web. To realize the SW we need open-licensed, structured, non-proprietary format, Uri-identified, linked data. When talking about LOD we can refer to three main different objects:

- Datasets: sets of triples that can be queried. They correspond to the triples we have created.
- Value vocabularies: controlled lists of allowed values for an element in a triple.
- Metadata element sets of controlled-form elements used to define classes and attributes used to described entities of interest.
  An ontology is a metadata element set since it provides a vocabulary to formally define properties and classes about resources taken into consideration.

Furthermore, we have to define the "linked" aspect of LOD: we have to find a way to define the relationship between the things on the web through:

- Relationship links → link resources to other data resources
- Identity links → link a resource to another resource described in other locations with which it is related by an identity bound (e.g. owl:sameAs).

It's fundamental in providing different points of view about the same resource and also to reconnect the same things.

- Vocabulary links → link the data to the vocabulary terms used in its description.

LOD workflow

We can follow a particular methodology in realizing a LOD project.

1. Specification
   Identification and analysis of data sourcing in a specific domain, URIs design and open license definition.
2. Modelling
   Abstraction and modelling of the data involved, the individuation of the most suitable ontologies needed to describe them (→the strong notion of reuse!).
3. Generation
   Transformation of the data through the formalization of their description to create an RDF based on the ontologies we decided to use.
   In this step is important to link our data to other existing repositories and to create connections among them.
4. Publication
   Publication of the project in the LOD data cloud and realization of its effective discovery by other users.
5. Exploitation
   The exploitation of the expressivity of the LOD resources to allow the final user to make inferences (nowadays still missing).