WoodsLOD documentation

The idea

We decided to realize an event-based LOD project about an event that is still nowadays seen as a moment of great changements are revolutions for the contemporary Western society: the Woodstock Festival of Music and Art held in New York State in 1969.
Due to the variety and complexity of the elements and concepts inolved in this mythic "3 days of Peace and Music" we though it could have been interesting to deepen its knowledge with the aim of retrieve and bring to light the network or relationships and interconnections that holds together heterogeneous and sometimes controversial world.

The 10 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 analize in order 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***, a track audio from MusicBrainz the Music Enciclopedy Catalogue
- ***Woodstock ticket***, an historical object form 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 Massachussets Amherst Libraries
- ***Woodstock Music from the original Soundtrack and More***, a vynil disk from the MusicBrain Catalogue
- ***D'oh-in the Wind***, a Simpson Tv Episode from IMDb
- ***Fender Stratocaster***, an historical object from Europeana.

All their description, although in many different ways and expressions, include information about people, places, dates and concepts involved in each specific item.
We decided to organize all these information in the entities involved in our project with related to

- People: **Eddie Kramer**, **Jimi Hendrix**, **Janis Joplin** and **Abbie Hoffman**
- Places: **Bethel**, **Woodstock**, **Bethel Woods Art Center**
- Time entities: the year **1969** and the time span 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 the scenario realized from all the object in their whole 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. In order to realize it we had to distinguish between our **items**, conceived as the real world objects related to our event and kept in the

different institutions; the **entities** involed in our project, defined as the people, dates, places and concepts that can be referred to our event and link together two or more items; and the **relationships** between them.
We used colors to distinguish items from entities and arrows to represent all the relationships.
Due to the complexity and variety of the system that had been forming, we had to **redefine** the E/R model **iteratively** until we found the best solution to represent compeltely and homogeneously all the relationships between every istances and the main event in consideration.

The identification of the metadata standards.

In our project we took in consideration mainly **descriptive metadata standards**: sets of rules, contraints and methodologies used by cultural institutions to describe the features of our items.
We analized the different behaviours and choices of the holding institutions involved and realized a table 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

Tab 1. Metadata standard identification

The standards involved in our project are 7:

- **SOMA =** acronym for 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-redeable cataloguing standards for the description of items bibliographic records. By the 70s they became the US national standards and then employed as international

ones. A MARC record is composed of three elements: the record structure, the content designation, and the data content of the record. MARC 21 Format for Bibliographic Data is designed to be a carrier for bibliographic information about printed and manuscript textual materials, computer files, maps, music, continuing resources, visual materials, and mixed materials.
Each record begins with a leader**,** which is a fixed field containing information for the processing of the record. Following the leader is the directory**,** which is an index to the location of the variable fields (control and data) within the record.
The standards present a generalized structure for records, but do not specify the content of the record and do not, in general, assign meaning to tags, indicators, or data element identifiers. Specification of these elements are provided by particular implementations of the standards.

- **CIDOC =** The International Committee for Documentation is a committee of the 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 catalog 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 invididuated all the metadata standards involved in our scenario, we proceded in finding **correspondaces** between them in order to align the different choices by the institutions in the description of the common features of our items.
Indeed, expecially 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 descriped. 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 commonly 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

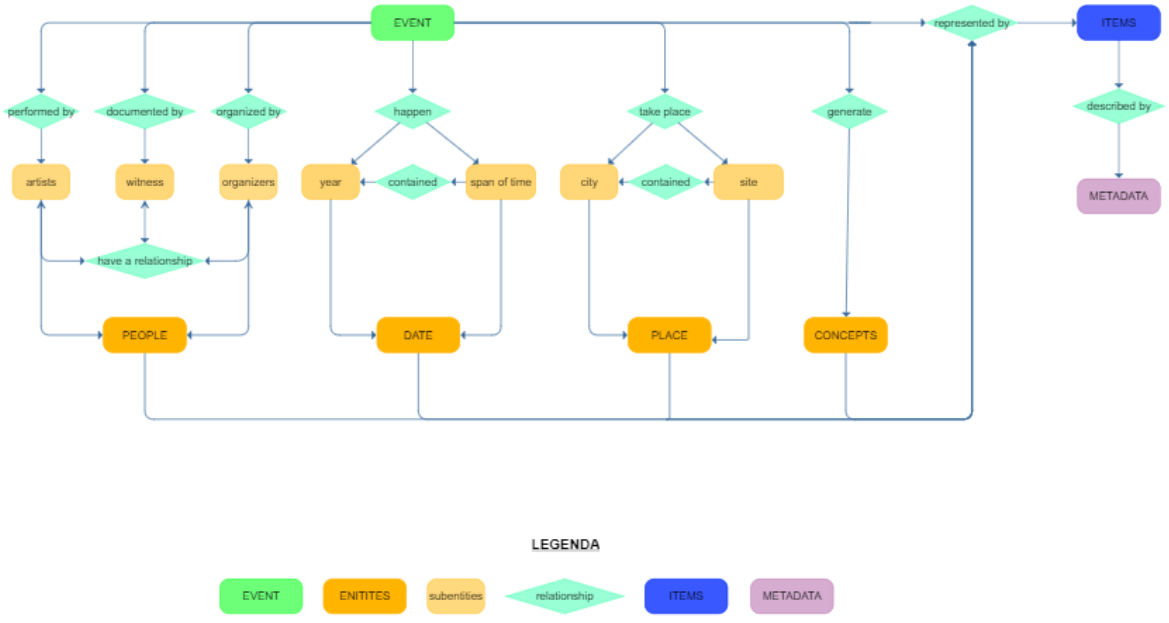
satisfiable solution will be realized only by the Semantic Web or Web 3.0 based on the explicit semantic interconnection of univoquely identifiable resources.
The metadata alignment step in our project wants hence to highlight and at the same time overcome the differences and uncompatbility of the different standards involved in the description of our items by finding the correspondances among all the descriptions.
Furthermore, in realizing it, we classified all the useful information provided about our items regarding to 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 analized the institution and most of all the metadata standards involved in our scenario, we could proceed to the real Knowledge Organization process. The firs 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 involed 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 represent a mid level of abstraction), linked together by relationships among them and with the main event, and represented by our chosen items.



Img 1.Theoretical model

Conceptual model

A futher level of abstraction was represented by the realization of a conceptual model able to answer to the same questions raised by the theoretical model in a more formal way.
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 thorugh a formal language, could allow us to draw abstract schemas of description of the main aspects involved in the representation of our main event.
In order 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 **rdf** was perfect for this prupouses – and then use more specific ontologies to describe the single features of each entity tipe 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 one that was used is the **Music Ontology** that provide 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 in order 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: **Bethe**l where it took place and **Woodstock** after which it was names, and the actual site where it was held now transformed in the B**ethel 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 explicitation 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 properites described are refered to it.

Data representation: the RDF statement

Finally, we managed to represent our data description through an **RDF statement** we decided to serliaze through **Turtle**.
The RDF statement allowed us to represent our data in 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 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 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 **the Vietnam war protest**.
As RDF identifies resources throug **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 stetment itself.
We managed to connect our entities with

- Term lists and **authority control** resources able to univoquely 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 identifiction 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 (member of).
- **Other resources identified by URL** already present in the web, mainly wikipedia resources.

Also, we gave a graphical representation of our data and their connection through a little **knowledge graph** able to visually represent the interconnection characterizing our information system.