OMOS: Ontology for Western Saharan Manuscripts

Mohamed Lamine Diakité, Béatrice Bouchou Markhoff

UNIVERSITE FRANCOIS RABELAIS – TOURS
ECOLE POLYTECHNIQUE DE L'UNIVERSITE DE TOURS
Département Informatique

LABORATOIRE d'INFORMATIQUE
(UPRES EA n° 6300)

OMOS: Ontology for Western Saharan Manuscripts

Mohamed Lamine DIAKITE
Béatrice BOUCHOU MARKHOFF

Ecole Polytechnique de l'Université de Tours
Département Informatique
64 Av. Jean Portalis
37200 TOURS - FRANCE
Tél. : 02 47.36.14.14. Fax : 02 47.36.14.22.
e-mail : beatrice.bouchou@univ-tours.fr

Février 2015
Rapport Interne n°313x18p

OMOS:
Ontology for Western Saharan Manuscripts

Mohamed Lamine DIAKITE [a] and Béatrice BOUCHOU MARKHOFF [b]
[a] *DMI, Université des Sciences de Technologie et de Médecine, Nouakchott, Mauritanie*
[b] *LI, Université François Rabelais de Tours, France*

Abstract. As more efforts are performed to digitize Western Saharan manuscripts, for preserving the memory they represent, the need to be able to work on these digitized materials naturally grows. Beyond cataloguing, an ontology is the basis to provide to researchers new tools for retrieving and integrating these knowledge sources. In this paper, we present OMOS, an ontology describing West-Saharan manuscripts. We first precise the domain and the purposes, then we illustrate each step of the ontology's building, from experts and local resources to its alignment with well-established reference ontologies, including its automatic enrichment from existing thesaurus. This incremental process may be reused in other projects.

Keywords. Ancient Manuscript, Ontology Design, Semantic Web, Thesaurus

1. Introduction

Ancient manuscripts digitization campaigns are one important part of cultural heritage preservation policies led by UNESCO, or national archives and museum. For instance in Mauritania, most of written treasures are conserved in private libraries, i.e. family homes, and those manuscript's owners lack the resources to protect them from dust, bacteria, insects, etc. A straightforward way of saving the content of this heritage is to take numerical images of it. Moreover, even when original manuscripts are well protected, using them destroys them little by little. For this reason, experimented organizations, e.g. the large occidental and oriental libraries, also tend to build digitized collections of ancient manuscripts.

Beside the heritage's long-term preservation, and depending on the rights granted by the owners, digitized documents may be published, shared and disseminated for studies in humanities and social sciences, or for world citizens' education. This is the aim of the BIBLIMOS[1] project, led by CITERES[2], on which we are working. Roughly speaking, BIBLIMOS proposes to collect information, and facilitate thematic corpora constitution, from public and private archives pertaining to the history of the Western Saharan region. Its goal is to provide to local students and researchers the ability to study their history, through a remote access to original materials through their image, and also the ability to more easily collaborate with foreign teams, on these materials.

To this aim, in addition to digitizing manuscripts, it is crucial to store information about it. In general, basic information about the original manuscript is kept together

[1] BIBLIothèque digitale Multilingues des sources inédites de l'Ouest Saharien

[2] http://international.univ-tours.fr/centre-for-cities-territories-environment-and-societies-citeres--283347.kjsp?RH=INTER

with images, sometimes using the Dublin Core, or DCMI[3] vocabulary. But this useful generic information quickly shows its limits when it comes to deal with the knowledge contained in manuscripts. Even when Optical-Character-recognition (OCR) can be performed, which allows researchers to use natural language processing for some information retrieval or text mining, there is a need to associate more knowledge to each digitized document. The needed knowledge is then represented as an ontology, i.e. a shared conceptual representation of a domain, consisting of a set of concepts and their relationships. Moreover, due to their ability to explicitly represent data semantics, ontologies traditionally play an important role in data integration processes [1], which is also a key property when dealing with semantically heterogeneous sources such as Western Saharan manuscripts.

So, we started building an ontology for describing digitized manuscripts of Western Sahara, originally targeted to those held by the Mauritanian Institute of Scientific Research in Nouakchott (IRMS[4]). Our goal is that it serves for enriching the information associated with the images, as a first step, to then be able to offer researchers efficient ways of building their corpora. Of course, we plan to base the intended new tools on semantic web technologies. The paper is organized as follows: in Section 2 we detail the context of our work. In Section 3 we introduce OMOS, through the first building step that led to its current version, starting from the manuscript catalogue we get, and the expert interviews. In Section 4 we consider public reference resources on the semantic web, that allow us to enrich OMOS on the one hand, and on the other hand, to link it with standard models. We conclude in Section 5.

2. From Mauritanian Manuscripts to Semantic Web Technologies

The BIBLIMOS project aims to address manuscripts from Western Sahara, mainly from Mali and Mauritania. We first recall what these manuscripts represent for the world cultural heritage, focusing on Mauritania. Next, we briefly explain BIBLIMOS' goals, then we sum up a state-of-the-art for manuscript descriptions and we analyse some advantages of semantic web technologies with respect to our context.

2.1. Western Saharan manuscripts

"Mauritania is known […] for its enormously rich heritage of Arab manuscripts, many brought from the Arab East by pilgrims returning from Makkah, some recopied from those imported sources by students in the Qur'an schools […], and others composed by Mauritania's own jurists, poets and historians." [5] To illustrate this richness, the author of [2] evokes for instance the only existing copy of a work by Averroës, on grammar ("Al-Daruri fi Sina'at al-Nahw, or What Is Necessary in the Making of Grammar"). This copy is part of a family library, established since three hundred years, as many other existing families of scholars, writers and collectors. According to researchers, some Mauritanian manuscripts are even from the 10th-century, and their forms and subjects are very diverse. To have an access to this wonderful legacy, the first step is to build up a precise survey of all manuscripts repositories in existence in the territories of

[3] World widely used, simple and generic, digitized resources' description: http://dublincore.org/

[4] http://www.imrs.mr/spip.php?page=sommaire

[5] http://www.saudiaramcoworld.com/issue/200306/mauritania.s.manuscripts.htm

the Western Saharan region: this has been the goal of the West African Arabic Manuscripts Database Project, from the University of Illinois at Urbana-Champaign, since 1987. Their database, whose address is given at the first line of Table 1, is a catalogue that references more than two thousands of manuscripts, which may be queried on the web, with a combination of three criteria among the cataloguing information (collection, title, author, subject, etc.). Currently, it references eleven collections, which still is far from representing the actual reality of family libraries. Nevertheless, this is one of the web resources we plan to exploit in the next stages of the BIBLIMOS Project, in parallel of completing the repositories survey work (performed by our SSH colleagues). Several other web sites provide information on Mauritanian or, more generally, on Arabic manuscripts: we list the most representative ones in Table 1.

Table 1. Web sites about Mauritanian, or Arabic manuscripts.

Site	Description
http://www.westafricanmanuscripts.org/	**University of Illinois at Urbana-Champaign**. Online catalogue; references about 22500 manuscripts from eleven different collections, including Northwestern Univ.
http://digital.library.northwestern.edu/arbmss/index.html	**Northwestern University, Chicago**. Online catalogue; contains entries from four separate collections.
http://memory.loc.gov/intldl/malihtml/malihome.html	**Library of Congress**. Online catalogue, with access to images of 32 manuscripts from Timbuktu, Mali.
http://gallica.bnf.fr/	**French National Library**. Online access to 35 manuscripts from Timbuktu, Mali.
http://www.tombouctoumanuscripts.org	**University of Cape Town**. Tombouctou Manuscripts Project; access to primary sources upon registration.
http://omar.ub.uni-freiburg.de/	**Universities of Freiburg and Tübingen** (Germany). Online images of approx. 2.500 Arabic manuscripts (134.000 images) from Mauritania, with bibliographical metadata.
http://wamcp.bibalex.org/	**Bibliotheca Alexandrina** (Egypt): online resource of a collection of Arabic manuscripts related to classical medicine, around 1000 books and fragments.
http://www.qdl.qa/en	**Qatar Digital Library** (with the British Library): archives, maps, manuscripts, sound recordings, photographs with explanatory notes and links, in both English and Arabic.
http://www.archive.org/	Search on 'Arabic manuscripts' -> some digitized books
http://www.islamicmanuscript.org/extresources/manuscriptcatalogues.aspx	List of Islamic manuscripts catalogues.
http://openlibrary.org/	List of Islamic manuscripts catalogues.

Carrying out some scientific work by using the resources listed in Table 1 is still difficult, as there is no mean to perform cross-references, comparisons and to analyse the different points of view they provide, etc. One of the difficulties that can be easily noticed, is that the basic vocabulary, used to define the subjects of the manuscripts, changes from one resource provider to the other. Sometimes there is even no indication on the terms to use, and, in general, no precise details are given on the chosen vocabulary. Thus, the user has to rely on her own judgment for the choice of terms to use, which leads to some imprecision in the search results (or to no result at all). More generally, the used data models were asynchronously designed, making them difficult to compare. These are challenges that the semantic web can resolve, as explained in Section 2.3, and this is what motivates our work in the BIBLIMOS project.

2.2. BIBLIMOS

Today, manuscript sources concealed in the West Sahara are partly inventoried, and most European archive funds are available to the public. However, the query, the analysis, the combination and the intersection of these multiple funds, represent a major challenge for every interested person. As shown in Table 1, online digitized full-text manuscripts exist, duly indexed and cataloged, but in the field of automatically data-processing such sources, all has to be designed and built.

Led by a cross-disciplinary and international team of researchers in the humanities and in computer science, BIBLIMOS fits in with a view to renewing the knowledge and analysis of the West Sahara's societies, and the relationships they have maintained among themselves and with neighboring, and also with remote societies, through the centuries. The ambition is to make available to researchers from North and South an open and interactive tool for searching and comparing local archives funds, including the manuscripts of the desert, and European archives related to these regions. In due course, the BIBLIMOS' aim is to offer the cross-referencing of Arabic, Pulaar, Soninke, Wolof, French, Spanish, Portuguese, Italian, Dutch, German, English[6] sources relating to the political, military, economic, legal, social and religious history of the territories of the Western Saharan region, from the modern era to the end of the Cold War.

Concerning computer science, the BIBLIMOS programme aims to create an e-infrastructure for a network of information around the history of the West Sahara. This open tool will offer (i) an access to sets of archival sources and original manuscripts, (ii) a guide to navigate this knowledge network, (iii) an automatic registration of new sources and (iv) new tools for knowledge creation and visualizations. It will also be interfaced with various useful existing applications for research, such as electronic publishing platforms, collaborative editing tools, bibliography management tools, etc. To achieve this goal, two lines of work have been initiated. First, to instigate, assist and sustain the creation of quality digital resources from the original sources, and to develop partnerships with providers of already existing digital resources, and second, to incrementally build the target distributed e-infrastructure, including a web portal as mediator, relying on semantic web resources and technologies. The building of OMOS represents a preliminary step for both of these two work lines: for the locally digitized sources it is aimed to be *the basis of an annotation framework*, and for the e-infrastructure it will serve to *represent local sources* in the integration framework.

2.3. Related Works

Concerning manuscripts, many different descriptions may be stored in computer memories: (*i*) seeing the manuscript as an archeological *object*, i.e. starting from its external aspect, a set of features may be evaluated, for instance the material it is made with, the colour of ink, etc. This is called *codicology* and a well-established vocabulary for such a set of descriptors is provided by the IRHT[7]; (*ii*) a *numerical image* of the manuscript can be taken; (*iii*) a *transcription* of the manuscript's textual content can be created, either manually or automatically from its numerical image (with OCR tools); (*iv*) both the image and the transcription may be *annotated*, this is the case for many

[6] In future phases, the collections of BIBLIMOS could be increased, depending on the relevance of the sources, to Tamasheq, Songhay, Turkish and other languages.

[7] See http://codicologia.irht.cnrs.fr

european manuscripts[8], whose textual contents are encoded using the TEI standard[9]; (*v*) finally, the manuscript can be *cataloged*, i.e. classified and described by librarians or archivists, so it could be found again among collections, as simply as possible: this supposes to define and then identify descriptors, including the *location* of course, and some general information about the *content*.

For each of these points of view, active researches are conducted and, in some cases, they converged to well established standards. Concerning ancient Arabic manuscripts, for instance [3] presents the problem of cataloguing, stating the difficulties involved in identifying the metadata used by different schools (those dealing with specimen and those addressing whole volumes). The solution proposed for enhancing the interoperability is to rely on the DCMI vocabulary (an alignment is given). The TEI, aimed at helping libraries, publishers, museums and universities to encode texts in order to facilitate information retrieval from textual contents, is another important support for interoperability [4]. Nevertheless we cannot hope to use it in the short term because for now the only way to get transcriptions of Mauritanian manuscripts is to do it manually. Indeed, automatic character recognition algorithms hardly apply to these kinds of manuscripts, written with Arabic graphemes but very often actually in many other languages (e.g. Pulaar, Wolof, etc.). In [5], the author recall the existing difficulties for applying OCR to ancient Arabic manuscripts and, although recent advances are reported in [6] and [7], they need to be further developed. Manuscript's image analysis is not reduced to OCR: for instance, *word spotting* may be a useful alternative to character recognition. This is why several works propose to build ontological descriptions (or sets of metadata) of *graphical image features*, in order to index and retrieve manuscripts' digital images on this descriptive basis [8], [9]. But to the best of our knowledge, such proposals have never been applied to ancient Arabic manuscripts.

When it comes to ontological representation of ancient manuscripts, the work described in [10], about the SAWS[10] project (Sharing Ancient WisdomS), is clearly an example of what we target in the BIBLIMOS framework. It deals with collections of moral and social advice and/or philosophical ideas from Greek and Arab wisdom literatures. Many of the concerned manuscripts have been transcribed and annoted using TEI during the first stages of the project, and *an extension of the FRBRoo*[11] *ontology* [11] has been developed to describe the transmission of information (from one copyist to another and from one language to other ones). Then, the authors explain how they extract from the TEI annotations the relationships defined in the ontology, for generating a conceptual network expressed in RDF[12]. This network allows researchers to explore links between the different documents' contents. This is an example of how the semantic web technologies contribute to the building of new means of knowledge, by opening up and linking various sources for research, which would otherwise remain isolated and unused.

The semantic web is a network, or a graph of semantic representations of web-published information, that relies on the same technical principles as the web pages network. Programs can operate on data at this semantic level. Among main semantic

[8] For instance those of the BVH in Tours: http://www.bvh.univ-tours.fr/Rabelais/rabelais_en.asp

[9] Text Encoding Initiative: http://www.tei-c.org/index.xml

[10] http://www.ancientwisdoms.ac.uk/

[11] http://www.cidoc-crm.org/docs/frbr_oo//frbr_docs/FRBRoo_V2.0_draft_2013May.pdf

[12] Data model standard: http://www.w3.org/RDF/

web developments are (*i*) the web ontologies and (*ii*) the linked (open) data; they provide a global space of interoperability. Both are important features regarding BIBLIMOS' aims, hence for our work on OMOS. As we show in next sections, this has direct implications on the design process that we follow for OMOS.

We think that our design process is a generic one, enough to be followed in other ontology developments, currently needed in more and more SSH projects. The building of ontologies is studied from long time and several methodologies, languages and tools exist [13]. An ontology may be built either *manually* or *semi-automatically*, sometimes from a reference ontology, or from textual corpus, or by *integrating* existing ontologies, etc. In our design process, we start manually, and automatic enrichment are performed. But what is crucial to keep in mind is that a web ontology *can not be a closed item*, it must have input and output links. Local lightweight ontologies, that describe a given data set, must be aligned to more global reference ontologies, allowing mediator systems to integrate local data sets, for instance following the principles described in [14, 15]. Some well-established reference knowledge resources play the role of hubs in the linked data network[13], the most visible are fact resources, e.g. DBpedia, but at the conceptual level, *reference domain ontologies* such as CIDOC CRM[14] [16] for cultural heritage, with FRBRoo for libraries, AGROVOC for agriculture and environment, or SNOMED CT for medicine, also act as fundamental integration means. These reference domain ontologies are the product of a long, international collaborative work, reflecting a consensus among the domain experts on the representation of their knowledge. Notice that the *collaborative* dimension, together with the complete *distribution*, are two web's key features, which the semantic web naturally inherits. In the context of BIBLIMOS, this could be extremely powerful because these two features mirror the local structural organization of the Mauritanian family libraries, open to communities but distributed in the country rather than centralized in only one authoritative place. The semantic web resources also promote multilingualism, as evidenced by multilingual resources such as CIDOC CRM, VIAF[15] or RAMEAU, the French national library thesaurus now accessible on the semantic web and fully interlinked with the German (SWD) and the American (LCSH) thesaurus (thanks to the *Multilingual ACcess to Subjects* project). As shown in next sections, these resources are used in the design of OMOS.

3. OMOS: starting from local resources

To build up the OMOS ontology, we first proceeded with the identification of knowledge types we have to represent, and the sources that are actually available for this knowledge engineering task. Then, we designed a first version from the catalogue that the Mauritanian Institute of Scientific Research in Nouakchott (IRMS) provided us. This preliminary version allowed us to initiate several discussions with historians and other researchers participating in BIBLIMOS, giving rise to new descriptions, that we also modeled. This incremental process is shortly presented in what follows.

[13] http://lod-cloud.net/

[14] http://www.cidoc-crm.org/docs/cidoc_crm_version_5.1.2.pdf

[15] Virtual International name Authority File: http://viaf.org/

3.1. Descriptive Knowledge

We identify at least two different levels of knowledge that characterize the manuscripts. The first level corresponds to descriptive and situational knowledge, which applies to the manuscript's exploitation regardless of their content. The second level is related to knowledge coming directly from the manuscript's content. Clearly, this type of knowledge is both harder to get, far more diverse, and less bounded than the former description level. We started with the simplest level.

Knowledge external from manuscript's content, which we also call descriptive knowledge, is knowledge related to the manuscript's exploitation from a librarian point of view, i.e. what can be found in a catalogue. This is useful for finding the manuscript again later on, starting from its immediate apparent features' values. The only reference to content might be found in the subject related to the manuscript, when it is mentioned. In well-developed libraries, this field, the subject of the book or manuscript, is actually very important and the values it can take are carefully harvested in a shared knowledge support called a thesaurus, which in general contains some semantic relationships existing between the terms. Figure 1 sums up the sources that might be used to build an ontology for the descriptive knowledge concerning manuscripts.

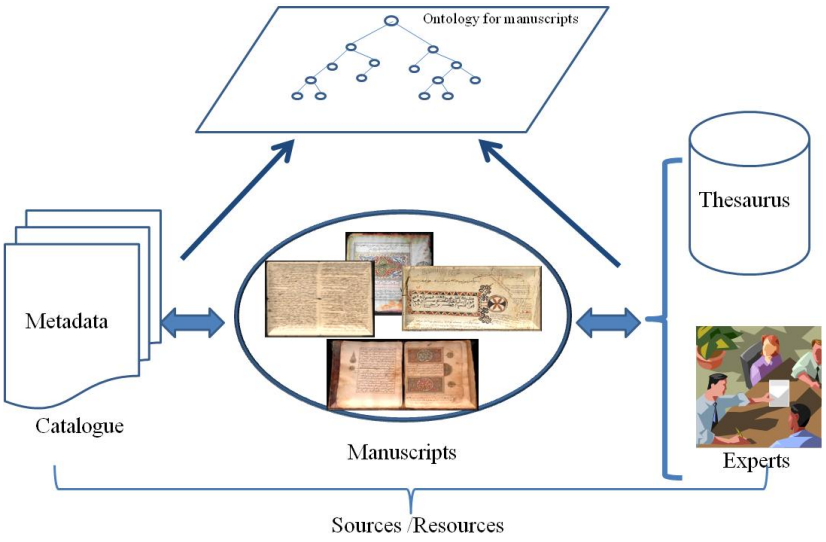

Figure 1. Sources for the knowledge engineering process.

In our context, the descriptive knowledge comes from the exploitation of existing metadata in the IMRS's library catalogue, and from discussions we had with experts knowing the corresponding manuscripts. We give in Table 2 the main elements taken from the IRMS's catalogue of manuscripts, enriched with suggested fields from the experts. In the first rows are the catalogue fields (in Arabic). They all describe a manuscript, either an original or a copy. Notice that a copy may be an exact copy but, more often, it is a copy with added comments. In OMOS, we represent most of these fields as attributes of the Manuscript concept, but it can be noticed in Table 2 that some of them are represented as concepts. This is the case for the author (or copyist) and for the subject. Obviously, the author may himself be described in more details, and such a

description is precisely considered very important by researchers, because it may be the basis of some prosopography studies. Concerning the subject field, it is less clear whether it might be a concept, considering that there is no use of thesaurus at the IMRS. But the aims of our social sciences colleagues imply that this field should be detailed. We give in last rows of Table 2 some examples of concepts also suggested by these experts. Most of them are represented as concepts in OMOS, modeling the persons, the places and the periods related to the manuscripts.

Table 2. Concepts from the IMRS catalogue (in English and Arabic), and some of those suggested by experts.

Information		About Resource	Type
Author	المؤلف	Manuscript	Concept
Title	العنوان	Manuscript/Copy/Copy with comments	Attribute
Subject	الموضوع	Manuscript/Copy/Copy with comments	Concept
Copyist	الناسخ	Copy	Concept
Write mode	الخط	Manuscript/Copy/Copy with comments	Attribute
Ink color	الحبر	Manuscript/Copy/Copy with comments	Attribute
Page surface	قياس الصفحة	Manuscript/Copy/Copy with comments	Attribute
Page number	الصفحات	Manuscript/Copy/Copy with comments	Attribute
Written surface	مساحة النص	Manuscript/Copy/Copy with comments	Attribute
Line number	الأسطر	Manuscript/Copy/Copy with comments	Attribute
Entire or not	النص تام ؟	Manuscript/Copy/Copy with comments	Attribute
Year[16] of publication	تاريخ النشر	Manuscript/Copy/Copy with comments	Attribute
Incipit[17]	البداية	Manuscript/Copy/Copy with comments	Attribute
Explicit[18]	النهاية	Manuscript/Copy/Copy with comments	Attribute
Exegetes		Copy with comments	Concept
Birth year / Death year		Author/Copyist/ Exegetes	Attribute
Birth Place		Author/Copyist/ Exegetes	Concept
Place of origin		Author/Copyist/ Exegetes	
Library		Manuscript/Copy/ Copy with comments	Concept
Location		Library / Private Library / Public Library	Concept
Language[19]		Manuscript/ Copy/Copy with comments	Concept
Field		Manuscript/ Copy/Copy with comments	Concept
Named time period		Manuscript/ Copy/Copy with comments	Concept

The modeling of that kind of knowledge, the descriptive one, allowed us to build a first version of OMOS, illustrated in Figure 2 and Figure 3. Figure 2 is an excerpt of the taxonomy of concepts, which allows programs to infer new facts at query time with the standard SPARQL entailment rules[20]. Figure 3 shows the relationships that are defined between concepts. For instance, a manuscript is written by an author, a library is located in a place, which belongs to a country, a person lived during some historical periods, etc. It is also possible to apply automatic inferences at this level, in temporal and in spatial dimensions. For example, we can imagine a situation where it could be possible to establish a manuscript's influence with respect to places and times of its copies, thanks to relations in between OMOS' concepts.

[16] Either Egira or Gregorian.

[17] Manuscript's first words.

[18] Manuscript's last words.

[19] Manuscripts are written with the Arabic script, but sometimes in local languages, for instance in Pulaar, in Soninke, in Wolof, in Tamasheq, etc.

[20] http://www.w3.org/TR/sparql11-entailment/

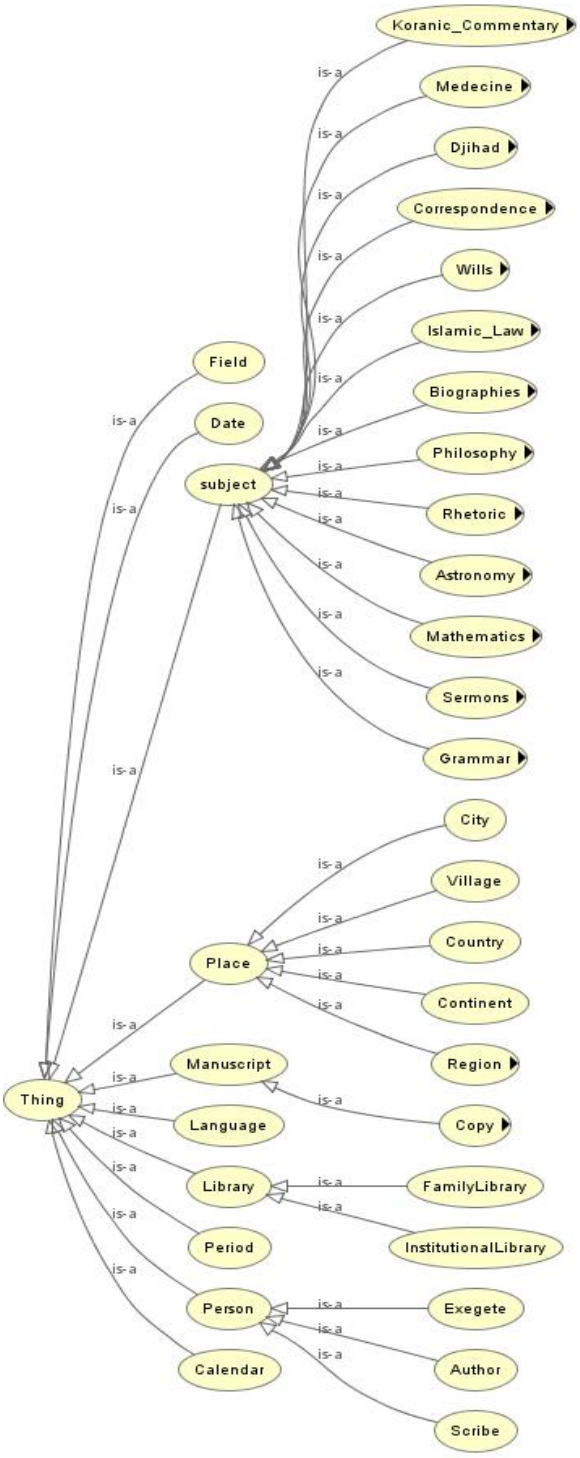

Figure 2. Taxonomy of concepts.

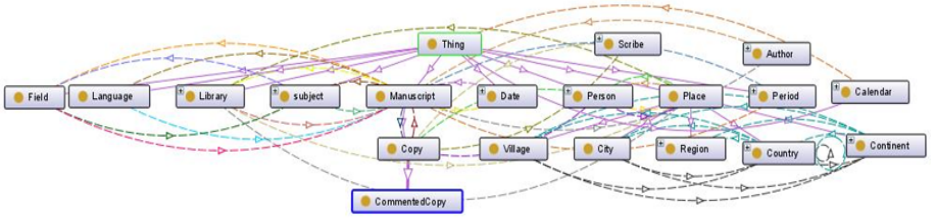

Figure 3. Concept's relationships.

3.2. Content Knowledge

From discussions with humanities and social sciences researchers, the OMOS' first version might be extended to model what the manuscript is talking about. This process highly depends on the expert's domain and particular interests; it can concern many different subjects, related to different domains. This is a kind of extension design, leading to a modular ontology, hinged on the generic descriptive kernel. For instance, we give in Figure 4 a small excerpt of a military conflict's description. The focus of this excerpt is on actors, but the model includes dates, locations, etc. Such a model may be built for a specific corpus, i.e. it may concern a subset of manuscripts, and a manuscript can be described using several different such models.

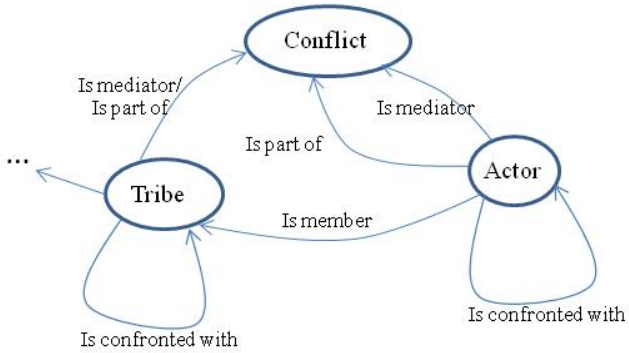

Figure 4. Excerpt of a conflict's conceptual description.

Each of these models is associated to the kernel version of OMOS, for instance as illustrated in Figure 5. It can be noticed that the concept of Conflict is defined as a possible subject (of a manuscript) and, as for persons, dates and places, the concepts needed for the content's descriptions are aligned with the generic ones, defined for the descriptive kernel.

The aim here is to incrementally enrich OMOS with content descriptions, in a modular and reusable way, as researchers build their corpora of manuscripts and conduct their works. In parallel of these manual extension buildings, we also performed experiments for automatically enriching the content-description part of OMOS, using some existing public resources built for this purpose, namely some reference thesauri. This is described in the following section.

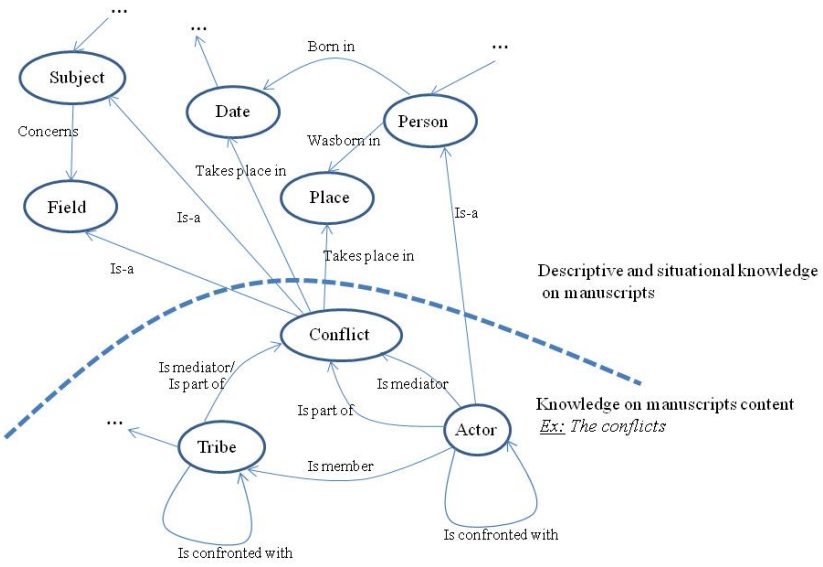

Figure 5. Linkage between a specialized part and the generic one.

4. OMOS and Semantic Web Resources

Even if potentially very useful for assisting researchers with intelligent tools, the building of sub-ontologies for content descriptions, conducted with experts as presented in Section 3.2, is a long and sensitive task. As the semantic web grows, more and more resources become available that may be automatically operated for enriching the manuscript's subject description. Moreover, reference ontologies must be used for making OMOS a semantic web resource, easily usable in integration systems. The best candidates as reference ontologies in the case of OMOS are CIDOC CRM and FRBRoo, its extension for libraries and archives.

4.1. Semi-automatic enrichment from thesaurus

A thesaurus is a networked collection of controlled vocabulary terms, i.e. well defined terms, organized into a hierarchical structure, and with associative relationships in addition to parent-child relationships. The expressiveness of the associative relationships varies from a thesaurus to the other. Large national libraries have specified their thesaurus since long time, in order to guide the subject indexing of documents they have to manage. Indexing means to associate terms to the document's record, in order to retrieve it by subject, using these terms. A thesaurus is the result of a long collaborative work lead by experts in information science, which may be re-used in all libraries using the same language. Sometimes, using the same language does not mean having the same culture or the same special interests: for this reason, a thesaurus can also be extended, depending on an editorial committee. This is the case for

RAMEAU[21], the French national library's thesaurus, also for the LCSH of the Library of Congress[22], or for the Dewey Decimal Classification (DDC) currently maintained by the Online Computer Library Center[23]. Many smaller and simpler thesauruses are also built in SSH, for listing specific features such as materials used for art works, or existing professions during the Middle Age, etc.

A thesaurus differs from an ontology on the purpose it is built for. An ontology is a kind of knowledge representation supposed to support *automatic reasoning services*. In particular, a semantic web ontology, i.e. an ontology expressed using RDFS or one of the OWL languages, follows the principles of Description Logics [19]. SPARQL implementations are intended to provide the corresponding reasoning service, called *entailment* regime, consisting in inferring all possible new facts before evaluating the query. In short, the knowledge of thesauruses is designed for human consumption while the knowledge of ontologies is computer-processable. Nevertheless, a W3C standard called SKOS[24] exists that allows librarians to re-express their thesaurus knowledge as a semantic web ontology. In this way, both RAMEAU and DDC can be queried in SPARQL (see the given footnotes for accessing their respective SPARQL Endpoint). We conducted an experiment in using RAMEAU for semi-automatically enriching the description of subjects in OMOS, by exploiting its public SPARQL endpoint.

RAMEAU comes from an adaptation of the Laval University's library's subject headings, which are themselves a translation-adaptation of the Library of Congress authorities (LCSH Library of Congress Subject Headings). The keywords are structured and selected for brevity, objectivity, specificity and consistency of document description. The set of keywords (in fact, descriptive records) is the national authority list, which is managed by the French National Library. It allows librarians indexing all types of materials (printed, audio-visual, graphic documents, ...) and hence provide a search by subject in catalogues. The thesaurus RAMEAU consists of a set of subject headings who have different types of relationships. The RAMEAU translation into SKOS was made according to the following principles: each subject heading in RAMEAU gives birth to a *skos:Concept*. Terminological data corresponding to the subject heading's labels (prefered and equivalent terms) are represented by the properties *skos:prefLabel* and *skos:altLabel*. Semantic links that correspond to the hierarchical and associative relationships are represented using the properties *skos:broader* (more generic), *skos:narrower* (more specific) and *skos:related* (associative) that refer to other concepts.

We use Jena[25] to develop an extraction program from the public SKOS version of RAMEAU (using its SPARQL Endpoint[26]). The aim is to take benefit of the knowledge encoded in RAMEAU, concerning the subjects first identified by our experts, illustrated in Figure 2. The idea is to use the existing relationships in RAMEAU, starting from each of these subjects. We focus on the *skos:narrower* relationship, searching for the first level of sub-subjects. Of course, the first matter is to find the subject headers in RAMEAU that are sufficiently close to the current OMOS subject.

[21] http://data.bnf.fr/en/semanticweb

[22] http://www.loc.gov/aba/

[23] http://www.oclc.org/dewey.en.html; http://dewey.info

[24] http://www.w3.org/TR/skos-reference/

[25] A Java system for managing RDF data: https://jena.apache.org/

[26] http://data.bnf.fr/sparql

Algorithm 1 summarizes the process for one given OMOS subject. Here is an example of a SPARQL query used in the developed Jena program that implements Algorithm 1:

```
Construct {?concept skos:prefLabel ?form; skos:narrower ?sub; skos:altLabel ?alt}
Where {?concept skos:prefLabel ?form; skos:narrower ?spec. ?spec skos:prefLabel
    ?sub;  skos:altLabel  ?alt.  ?concept  dcterms:isPartOf  ?rameau.
    FILTER  (REGEX  (?rameau,  "http://data.bnf.fr/vocabulary/scheme"))}
```

Notations used in Algorithm 1:

T_p : set of subject headings in RAMEAU.

C : subject (concept for which we search sub-concepts).

$T_{Alt}(C)$: set of subjects headings related to C with the *skos:altLabel* relationship.

S_C: the set of narrower concepts of all the subject headings linked to *C,* either directly or with the *skos:altLabel* relationship. This is the result, which is given to the user in order to allow him to choose the accurate sub-concepts of *C* for OMOS.

$t_1 R_{alt} t_2$: t_2 is linked to t_1 with the *skos:altLabel* relationship

$t_1 R_{spec} t_2$: t_2 is linked to t_1 with the *skos: narrower* relationship (t_2 is more specific)

Input: C, T_p
Output: S_C
begin
$S_C = \emptyset$, $T_{Alt}(C) = \emptyset$
if $(C \in T_p)$ then // *if C is a subject heading*
  while $(\exists t \,/\, C\ R_{spec}\ t)$
    $S_C = S_C \cup \{t\}$
  end while
else
  while $(\exists t \in T_p \,/\, t\ R_{alt}\ C)$ //*if C is a subject heading's equivalent term*
    $T_{Alt}(C) = T_{Alt}(C) \cup \{t\}$
  end while
end if
while $(\exists t \in T_{Alt}(C))$
  while $(\exists t_{spec} \,/\, t\ R_{spec}\ t_{spec})$
    $S_C = S_C \cup \{t_{spec}\}$
  end while
end while
end

Algorithm 1. Semi-automatic enrichment from a thesaurus such as RAMEAU

Our choice is to show to the user the result S_C, and to let him choose which item in this list is meaningful in the context of OMOS. This is why we talk about a semi-automatic enrichment from the thesaurus. This is always important in such ontology building process to provide to the experts a mean of validating, or correcting, the results obtained from programs. In Algorithm 1 this is absolutely mandatory when there are more than one subject heading linked to *C* with the *skos:altLabel* relationship.

But even when only one subject heading is identified with C, it may be useful to verify if all its sub-terms are accurate in the context of OMOS.

In Algorithm 1, it can be noticed that the result S_C may be empty when there is no subject heading linked to *C,* neither directly nor with the *skos:altLabel* relationship. In this case, the user might find in the thesaurus a subject heading semantically close to *C* whose label is lexically quite different from *C*. In our experiment, we started with 20 input concepts (subjects) and about 40% of them were in this case. For instance, the subject *Lettres* (*Letters*) that we found in the IMRS catalogue corresponds to the subject heading *Correspondance* (*Correspondence*) in RAMEAU. In the same way, the subject *Commentaires* (*Commentaries*), which is very general (thus ambiguous), matches to *Commentaires coraniques* (*Koranic commentaries*) in RAMEAU, taking into account Mauritanian manuscripts' context. Here again, only user-interaction can lead to the best choice, even if solutions developed for ontology mappings are more and more efficient to help in those cases [20].

Concerning our experiments, considering the semantic of subject allowed us to have 70% of the subjects in OMOS directly linked to a subject heading in RAMEAU, while it was only the case for 40% of the subjects using only syntactic matching. Among the other 30%, half of them were equivalent terms of subject headings in RAMEAU and the others did not match anything existing in RAMEAU. A last important lesson learned from these experiments concerns the cultural dimension of the used public resources: this is obvious when it comes to sub-concepts of *Sermons*, for instance, a part of which are the following in RAMEAU: *Jésus-Christ -- Passion – Sermons / Marie, Sainte Vierge – Sermons / Église catholique – Sermons*, which are not meaningful for describing the targeted West Saharan manuscripts that are dealing essentially with the Muslim religion.

We give in Figure 6 an excerpt of the sub-subjects finally extracted from RAMEAU.

4.2. Reference Ontologies

We set as a goal to produce an open, scalable and connected ontology, especially interoperable with other useful models for the whole BIBLIMOS project (e.g. which represent some particular archives of the Western Saharan area colonizing countries[27]) . We know that this can be done through its alignment with *reference ontologies* of the field covered by BIBLIMOS, since this is the principle of our proposals about semantic mediation [14], [15]. We are fortunate that the pivot reference ontologies needed for the domain of BIBLIMOS exist, which are CIDOC-CRM (Comité International de la DOCumentation - Modèle Conceptuel de Référence, ISO 21127) [11] and its extension, FRBRoo (Functional Requirements for Bibliographic Records - object oriented) [16][28]. Before presenting the alignment of OMOS on these reference ontologies, we briefly recall what they are.

[27] These archives also begin to join the opening process, for instance in France, being digitalized for, one day, may be also becoming linked data.

[28] Both are available in OWL DL 1.0 (Erlangen version): http://erlangen-crm.org/ and http://erlangen-crm.org/efrbroo/, and also in RDFS (see http://www.cidoc-crm.org/).

Figure 6. Excerpt of the resulting subjects taxonomy

As clearly explained in [11], "the CIDOC CRM transforms cultural heritage data from internal institutional inventories or catalogues into a highly valuable community resource because data accrues greater relevance and significance when harmonised to create densities of information, and also because the process of mapping data (the translation of source model to a target model) to the CRM returns both the meaning and context to the things represented in the data, essential for understanding". In other words, the main objective of CIDOC CRM is to provide a conceptual basis for information mediation amongst cultural heritage institutions such as museums, libraries and archives. The intention is to provide a common reference point with which differing and inconsistent sources of information can be compared, and their querying

finally harmonized (the *querying*, not the sources themselves, whose internal information system's model is certainly the best for their internal purposes). Concepts in CIDOC CRM (called Entities) are articulated around the interactions between *events* or *activities* (*temporal*, i.e. bounded by time) and *people*, *objects*, *ideas* or *concepts* (*persistent*, i.e. that survive over an indeterminate time). They are organized in a hierarchy and their relationships (called Properties) are also hierarchically linked.

FRBRoo [11] is one of the official extensions of the CIDOC CRM, built as a translation, or interpretation, of the FRBR reference for library cataloguing, in order to reach interoperability with other reference ontologies. While FRBR models the outcomes (work, expression...) of processes (such as creation), FRBRoo focuses on processes, or activities, in the same way as CIDOC CRM, which enables reasoning about the circumstances in which instances of works were designed. Conversely, FRBRoo adds to the CRM a basic model of intellectual conception and art work creation, which required the integration of the concept of work.

The alignment of OMOS with CIDOC CRM and FRBRoo ontologies is done by first checking, for each OMOS concept, the concept of CRM to which it will be mapped. As usual in alignment processes, the mapping is characterized with a hierarchical relationship (more specific or more generic), or with an equivalence relationship. We show in Table 3 the result of this first step of alignment, for a subset of concepts OMOS with CIDOC CRM and FRBRoo. More precisely, Columns A, B and C of Table 3 represent the mappings between OMOS concepts and the CIDOC CRM ones, while Columns D, E and F precise the relationship between the target CIDOC CRM concepts and FRBRoo (which indicates, transitively, the link between OMOS concepts and FRBRoo ones). Notice that the concept Manuscript is correctly related to the FRBRoo Entity *F4 Manifestation Singleton*, but the link between this one and the CRM's concept *E22 Man-Made Object* (ancestor of *E84 Information Carrier*) goes through *E24 Physical Man-Made Thing* (ancestor of *F4 Manifestation Singleton*).

Table 3. Main concepts of OMOS aligned with CIDOC CRM and FRBRoo

A	**B**	**C**	**D**	**E**	**F**
Most generic concepts	A w.r.t. C	CIDOC CRM	CIDOC Concepts linking C and E	FRBRoo	D w.r.t. E
Library	⊆	E40 Legal Body	E40 Legal Body	F11 Corporate Body	⊆
Calendar	⊆	E49 Time Appellation	E41 Appellation	F12 Nomen	≡
Date	≡	E50 Date	E49 Time Appellation	F12 Nomen	⊆
Domain	⊆	E28 Conceptual Object	E28 Conceptual Object	F6 Concept	≡
Language	≡	E56 Language	E55 Type	F6 Concept	⊆
Place	≡	E53 Place	E53 Place	F9 Place	≡
Manuscript	⊆	E84 Information Carrier	E22 Man-Made Object	F4 Manifestation Singleton	人
Period	⊆	E49 Time Appellation	E41 Appellation	F12 Nomen	≡
Person	≡	E21 Person	E21 Person	F10 Person	≡
Subject	⊆	E28 Conceptual Object	E28 Conceptual Object	F6 Concept	≡

The next step consists in considering the relationships between concepts to verify if there exist an equivalent in the reference ontologies, more precisely a possible ancestor property. This step is in progress. We are also studying how the CIDOC CRM proposes to attach specialized thesauruses to entities, and even to properties, in order to follow the same principles in OMOS.

5. Conclusion

To the best of our knowledge, OMOS is the first ontology dedicated to ancient Arabic manuscripts. Its construction carefully takes into account the state of the art on existing related works, the IMRS' catalogue and the samples provided by our SSH colleagues, together with their needs with respect to their analysis and corpus building tasks. Moreover, we deliberately conduct this construction incrementally, and one of our most important guiding principle is to make OMOS open and linked, i.e. to insert it into the semantic web.

Our design process follows several steps: in the first stage, completely manual, we used the IMRS catalogue and the state of the art for manuscripts' description. Then we initiated the design of extensions concerning special interest domains that our SSH colleagues wish to deal with (for instance the thematic of conflicts), and we defined how these extensions could be linked with the ontology's kernel. Next, we experimented algorithms for semi-automatically enriching the description of manuscript's content, by using public reference resources from large libraries which are now available on the semantic web (e.g. RAMEAU). Lastly, we studied how to effectively link OMOS with reference ontologies, existing on the semantic web, we identified CIDOC CRM and FRBRoo as the best candidates and we performed the alignment of OMOS with them (manually).

As a dynamic knowledge tool, OMOS is still under construction and continual improvement, but we plan to use the current version in annotation tasks that we are preparing for students, for evaluating its relevance. This is an experiment inspired from works on corpus annotation developed in Natural Language Processing. There are several methods for ontology evaluation, some of them based on ontology's use cases, in given applications, other based on comparing the ontology with new data sources, etc. An overview of methods and tools for ontology evaluation is given in [17, 18]. More concretely, for validating the first version of OMOS with respect to standard design principles, we used the online tool called OOPS! (OntOlogy Pitfall Scanner), which detects most commonly done errors.

Our immediate future works are several already identified improvements, for instance with new experiments on using public thesaurus, this time with the DDC for integrating its Arabic version because it seems to be largely used in many libraries in Arab language. In parallel, we plan to build an annotation tool around OMOS, based on semantic web technologies, and to devise an efficient annotation protocol, such as those existing in NLP for corpus annotation tasks, or those defined in libraries for document indexing. Finally, our long-term objective is the complete development of the BIBLIMOS programme.

Aknowledgment

The authors thank Sophie Caratini and Francesco Coreale, from the CITERES laboratory (CIties, TERritories, Environment and Societies), UMR 7324 (Combined Research Unit CNRS-University François Rabelais of Tours), for initiating the very stimulating BIBLIMOS project, and their explanations on manuscripts and their usage.

References

[1] H.Wache, T.Vögele, U.Visser, H.Stuckenschmidt, G.Schuster, H.Neumann, and S.Hübner, *Ontology-Based Integration of Information – A Survey of Existing Approaches. IJCAI Workshop on Ontologies and Informations Sharing,* pp 108–117, 2001.

[2] L. Werner, *Mauritania's Manuscripts*, Saudi Aramco World, Vol. 54, No. 6, pp 2–16, 2003.

[3] M. O. Soulah and M. Hassoun, *Which metadata for Ancient Arabic Manuscripts Cataloguing? Proc. International Conference on Dublin Core and Metadata Applications*, The Hague, Netherlands, 2011.

[4] Desmond Schmidt, *Towards an Interoperable Digital Scholarly Edition*, Journal of the Text Encoding Initiative [Online], Issue 7, URL : http://jtei.revues.org/979, 2014.

[5] Abdel Belaïd, Nazih Ouwayed. *Segmentation of ancient Arabic documents*. Volker Märgner and Haikal El Abed. Guide to OCR for Arabic Scripts, Springer, 2011.

[6] W. Boussellaa, A. Zahour, H. El Abed, A. Benabdelhafid and A. Alimi, *Unsupervised Block Covering Analysis for Text-Line Segmentation of Arabic Ancient Handwritten Document Images*, *20th International Conference on Pattern Recognition* (ICPR), pp 1929—1932, 2010.

[7] Khemiri, A.; Kacem, A.; Belaid, A., *Towards Arabic Handwritten Word Recognition via Probabilistic Graphical Models*, *Frontiers in Handwriting Recognition* (ICFHR), pp.678—683, 2014.

[8] M. Coustaty, R. Pareti, N. Vincent and J.M. Ogier, *Towards historical document indexing: extraction of drop cap letters*, IJDAR, Vol. 14, n°3, pp 243—254, 2011.

[9] B. Coüasnonet and J. Camillerapp and I. Leplumey, *Access by Content to Handwritten Archive Documents: Generic Document Recognition Method and Platform for Annotations*. International Journal on Document Analysis and Recognition, IJDAR, 9(2):223-242, 2007.

[10] A. Jordanous, K. F. Lawrence, M. Hedges, and C. Tupman. *Exploring Manuscripts: Sharing Ancient Wisdoms across the Semantic Web*. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS '12)*. ACM, New York, NY, USA, Article 44 , 12 p., 2012.

[11] M. Doerr and P. Le Boeuf. *Modelling intellectual processes: The FRBR - CRM harmonization.* In C. Thanos, F. Borri, and L. Candela, editors, Digital Libraries: Research and Development, volume 4877 of Lecture Notes in Computer Science, Springer, Berlin / Heidelberg, pp 114—123, 2007.

[12] International Working Group on FRBR and CIDOC CRM Harmonisation, FRBR object-oriented definition and mapping from FRBR ER, FRAD and FRSAD (version 2.0), may 2013.

[13] Oscar Corcho, Mariano Fernandez-Lopez and Asuncion Gomez-Perez, *Methodologies, tools and languages for building ontologies. Where is their meeting point?* Data and Knowledge Engineering, vol. 46, pp. 41—63, 2003.

[14] C. Niang, B. Bouchou, Y. Sam and M. Lo, *A Semi-Automatic Approach For Global-Schema Construction in Data Integration Systems*, IJARAS, Vol. 4(2), pp. 35—53, 2013.

[15] B. Bouchou and C. Niang, *Semantic Mediator Querying*, *18th International Database Engineering & Applications Symposium (IDEAS)*, pp. 29—38, 2014.

[16] D. Oldman, *The CIDOC Conceptual Reference Model (CIDOC-CRM): A Primer,* Version 1, July 2014, CIDOC-CRM official web site (http://www.cidoc-crm.org/docs/CRMPrimer_v1.1.pdf), 2014.

[17] A. Gangemi, C. Catenacci, M. Ciaramit and J. Lehmann, *Modelling Ontology Evaluation and Validation*, In *The Semantic Web: Research and Applications*, LNCS 4011, 2006.

[18] S. Tartir, I.B. Arpinar and A.P. Sheth, *Ontological Evaluation and Validation* In *Theory and Applications of Ontology: Computer Applications,* Springer Netherlands, pp. 115—130, 2010

[19] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, and P.F. Patel-Schneider, Ed. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[20] P. Shvaiko and J. Euzenat, *Ontology Matching: State of the Art and Future Challenges*, IEEE Transaction on Knowledge and Data Engineering, Vol. 25(1), pp. 158—176, 2013.