Semantic Web for BIBLIMOS (position paper)

BÉATRICE BOUCHOU MARKHOFF[1], SOPHIE CARATINI[2], FRANCESCO COREALE[2], MOHAMED LAMINE DIAKITÉ[3] and ADEL GHAMNIA[1]

[1] Université François Rabelais Tours - Laboratoire d'Informatique LI (EA 6300)
beatrice.bouchou@univ-tours.fr, adel.ghamnia@univ-tours.fr
[2] Université François Rabelais Tours - Laboratoire CITERES (UMR 7324)- EMAM team
sophie.caratini@univ-tours.fr, francesco.coreale@univ-tours.fr
[3] Université des Sciences, de Technologie et de Médecine - DMI, Nouakchott, Mauritanie
diakite@ustm.mr

Abstract. We present the BIBLIMOS project, which aims to address the Western Saharan culture and history, by considering both local ancient Arabic manuscripts and European colonial archives. We describe the project's context and objectives before focusing on ancient Mauritanian manuscripts, the content of which covers many scientific fields. We assess the current state of such ancient manuscripts' digital processing and we analyse what the semantic web can bring for their use by scholars, from North and South: the ability for *applications* to operate jointly on several distributed and heterogeneous sources of digitized manuscripts and other kinds of archives, to support collaborative reflection.

Keywords: Ancient Arabic Manuscripts; Data Integration; Semantic Virtual Infrastructure; Western Saharan Cultural Heritage

1 Introduction

BIBLIMOS is a long standing programme, led by the CITERES laboratory[4], that proposes to collect information, and facilitate the constitution of thematic corpora, from public and private archives pertaining to the history of the Western Saharan region. Its first goal was to provide to local students and researchers the ability to study their history, through a digital remote access to original materials (through images and descriptions), and also the ability to collaborate more easily with foreign teams, on these materials. Moreover, in the long run, it is also planned to deal with both primary sources (original material created at the time under study) and secondary sources (material written by scholars). In parallel, it is intended to address colonial archives about this geographical area, from European countries (mainly France and Spain), in order to cross complementary points of views, and thus, to discover new knowledge.

Involving an international and cross-disciplinary team of researchers in the humanities and, more recently, in computer science, BIBLIMOS aims to renew the knowledge and analysis of Western Sahara's societies, by making available to researchers from the North and the South an open and interactive tool for searching and comparing local

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

archive funds, including the manuscripts of the desert, and European archives related to these regions. There is also an important multilingual challenge, as we plan to perform cross-referencing of Arabic, Pulaar, Soninke, Wolof, French, Spanish, Portuguese, Italian, Dutch, German, English sources relating to the political, military, economic, legal, social, scientific 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 is just getting started: it aims to create an e-infrastructure based on a network of information around the history of the Western 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, three lines of work have been initiated. First, to instigate, assist and sustain the creation of quality digital resources from the original sources, second, to develop partnerships with providers of already existing digital resources, and third, to incrementally build the target distributed e-infrastructure, including a web portal as mediator, relying on semantic web resources and technologies.

In the first line of work, BIBLIMOS stakeholders in Social Sciences and Humanities (SSH) are engaged in actions aimed at discovering new local sources and convincing their owners to join the programme. Concerning the second line of work, today manuscript sources concealed in the Western Sahara are already partly inventoried, and many European archive funds are now available to the public. As shown in Table 1, on the one hand, online digitized full-text manuscripts exist, duly indexed and catalogued, and on the other hand, institutions or associations offer to collaborate in order to index digitized materials from many sources (cf. last lines in Table 1). Clearly *the Web*, that provides information *exploitable by humans*, well supports all those very useful initiatives. However, the query, the analysis, the combination and the overlapping of these multiple funds, still represents a major challenge for every interested person. This paper is dedicated to the third line of work in the BIBLIMOS programme, which addresses the field of the automatic data-processing of such sources, in order to better assist humans in these tasks. This is a field in which almost everything has to be designed and built. *The Semantic Web*, i.e., *the web knowledge exploitable automatically by computers*, is the way to cope with these challenges, as we argue in Section 3, after having presented the state of the art of digital processing of Ancient Arabic Manuscripts in Section 2.

2 Digital processing of Mauritanian Ancient Arabic Manuscripts

2.1 Mauritanian Ancient Arabic Manuscripts

We focus on Mauritania's manuscripts because Sophie Caratini, the instigator of the BIBLIMOS programme, is an anthropologist specialist of Mauritania and she built strong collaborations with scholars in Nouakchott, in particular through the IMRS[5].

[5] Institut Mauritanien de la Recherche Scientifique, see `http://www.imrs.mr/spip.php?page=sommaire_fr`

Site	Description
http://www.westafricanmanuscripts.org/	**University of Illinois, 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, 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 (BnF)**. 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 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.
makrim.org	**IMRS** (Mauritanian Islamic Republic). Catalog of Mauritanian manuscripts, in both French and Arabic.
http://www.islamicmanuscript.org/extresources/manuscriptcatalogues.aspx	**The Islamic Manuscript Association** (Cambridge, stakeholders from 25 countries). List of Islamic manuscripts catalogues.
http://openlibrary.org/	**Open Library** (world wide **open access** project). List of resources on Arabic manuscripts (catalogues, books, etc.).
http://www.archive.org/	**The Internet Archive** (USA non profit association). A search on *Arabic manuscripts* gives some digitized books.

Table 1: Web sites about Western Saharan, or more generally, Arabic 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[6] [16]. According to researchers, some Mauritanian manuscripts were written as early as in the 10^{th} century, and their forms and subjects are very diverse, including law, science and religion. To have access to this legacy, the first step is to build up a precise survey of all manuscript repositories in existence in the territories of the Western Saharan region. This has been the goal of long term projects: for instance, the West African Arabic Manuscripts Database Project, from the University of Illinois at Urbana-Champaign, started in 1987, provides a catalogue (first line of Table 1) that references more than two thousand manuscripts. Currently, it references eleven collections, which still is far from representing the actual reality of family libraries. This is one of the web resources we plan to exploit in the BIBLIMOS programme, in parallel of completing the repositories survey work performed by the SSH teams. Several other websites provide information on Western Saharan or, more generally, on Arabic manuscripts: the list presented in Table 1 shows that there is already a lot of knowledge available on the web, but this knowledge still is exploitable only through human labour.

2.2 Digital Processing of Ancient Arabic Manuscripts

Concerning manuscripts, many different descriptions may be stored in computer memories: (i) seeing the manuscript as an archaeological 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 [4] 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 European manuscripts, whose textual contents are encoded using the TEI standard; (v) the manuscript can be catalogued, i.e. classified and described by librarians or archivists, so it could be found again among collections: this supposes to define and identify descriptors, including the location, and some general information about the content.

For each of these descriptions, active research is conducted and, in some cases, they converge to well established standards. Specifically for ancient Arabic manuscripts, in [15] the authors present the problem of cataloguing, assessing 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 interoperability is to rely on the DCMI[8] vocabulary. The TEI[9], aimed at helping libraries, publishers, museums and universities to encode texts in order to facilitate information retrieval from

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

[7] Institut de Recherche et d'Histoire de Textes, see http://codicologia.irht.cnrs.fr

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

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

textual contents, is another important medium for interoperability [14]. 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 manually enter the text. 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 [1], the authors recall the existing difficulties for applying OCR to ancient Arabic manuscripts and, although recent advances are reported in [3] and [11], they need to be further developed. Manuscript 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 [7, 6]. 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 annotated using TEI, and an extension of the FRBRoo ontology [9] has been developed to describe the transmission of information (from one copyist to another and from one language to another). The authors extract the relationships defined in the ontology from the TEI annotations, to generate a conceptual network expressed in RDF[11]. This network allows researchers to explore links between the different documents' contents. This is an example of how 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.

3 Semantic Web Architecture for BIBLIMOS

For humans, carrying out some scientific work by using the resources listed in Table 1 is still difficult, as there are no means to perform cross-references, comparisons, or to analyse the different points of view they provide, etc. Regarding BIBLIMOS' aims, other kinds of sources than manuscripts (e.g. European archives) should also be exploited, which increases again these difficulties. Fortunately, while *the web* allowed sources' owners (or depositaries) to publish their resources through websites, *the semantic web* now supports the development of softwares that help humans to cope with these difficulties. Indeed, the semantic web is a network of semantic representations of web-published information that relies on the same technical principles as the websites' network, but allows *programs* to operate on data at this semantic level. Main semantic web concepts are (i) web ontologies and (ii) linked (open) data; they provide *a global space of interoperability*, thus they are important components for BIBLIMOS' aims.

Figure 1 illustrates the intended general architecture for the BIBLIMOS programme. The novelties brought by the semantic web obviously start at the DATA level: to benefit

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

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

from these novelties, beyond all the work that has to be done to obtain results presented in the previous section, digital sources should also be *pushed up to the semantic level*. To this aim, the sources' concepts and their relationships must be specified, from the bottom-up (starting from the source contents), top-down (from already well defined consensual ontologies), or both. The source's content should be related to this conceptual level, which may be done by using tools called *Mapping Frameworks* in Figure 1. Some of those tools propose to export the source data into a set of RDF triples (the standard data warehouse approach in data integration systems), and some of them propose to access data through the conceptual level, based on the ontology-based data access (OBDA) principles [5] (the mediation approach, which is provided by, e.g., *ontop*[12]). Whatever the chosen approach, the source's content is then searchable at the semantic level, with SPARQL. Those contents may be combined using reference thesauri and ontologies.

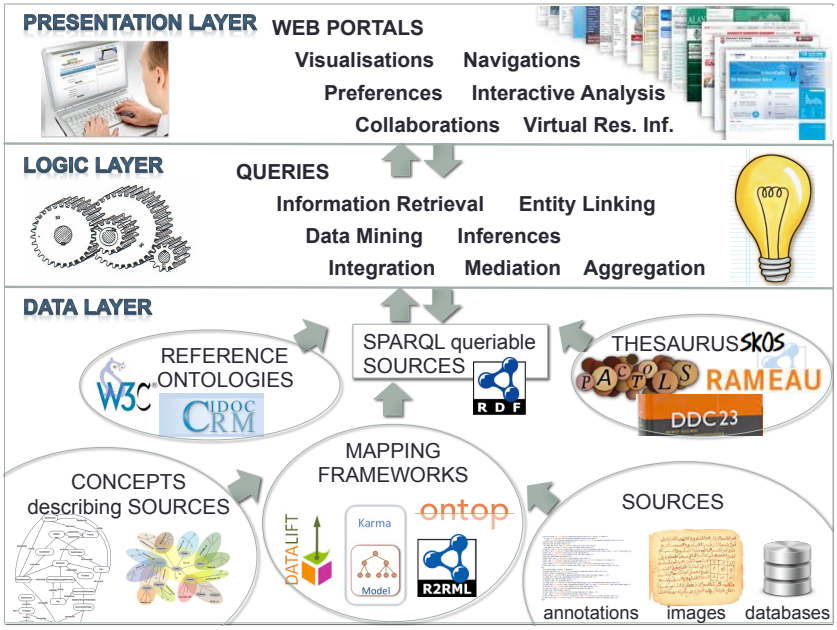

Fig. 1: Global BIBLIMOS' Virtual Infrastructure.

Querying the semantic web through its linked data sets is still in its infancy. Public well-established reference knowledge resources play the important role of hubs in this linked data network. The most visible are resources of facts, e.g. DBpedia, but at the conceptual level, reference domain ontologies also act as fundamental integration means. This is the case for CIDOC CRM [13] for cultural heritage, with its extension

[12] http://ontop.inf.unibz.it/

FRBRoo for libraries. These reference domain ontologies are the product of a long, international collaborative work, reflecting a consensus among the domain experts. These distributed and collaborative dimensions of the web are naturally inherited by the semantic web. In the context of BIBLIMOS, this is 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*, particularly in vocabulary resources such as *thesaurus*, as evidenced by multilingual ones, e.g. VIAF[13] or RAMEAU[14], the French national library thesaurus now accessible on the semantic web (in SKOS), which is fully interlinked with a German (SWD) and an American (LCSH) thesaurus (thanks to the *Multilingual ACcess to Subjects* project).

Above the DATA layer is the LOGIC layer, in which all the well-known successful inventions in the field of data operation (some of which are listed in Figure 1) may be revisited to take into account the semantic dimension of data. A corner stone for most of them is to access multiple sources conjointly, which supposes interoperability: one of the solutions provided by the semantic web is to align the local lightweight ontologies that describe the sources' content to the reference ontologies, allowing mediator systems to aggregate local data sets, for instance following the principles described in [12, 2]. Very active researches are conducted in the semantic web community to develop this LOGIC level, based on efforts to produce a strong semantic data layer. Lastly comes the PRESENTATION layer, whose innovative potential is also greatly boosted by the possibilities issued from the semantic web.

4 Conclusion

We first drew a state of the art concerning the ways ancient Arabic manuscripts are processed and made available to the public nowadays, considering that the picture is not so different in the area of European archives (except that OCR tools are more usable). Once digitized, sources must be pushed up to the semantic level, for the query, the analysis, the combination and the intersection of these multiple funds to be supported by automatic data-processing of sources. We presented the semantic-web Virtual Infrastructure designed to cope with these challenges within the BIBLIMOS programme.

We are aware that BIBLIMOS is a very ambitious programme - we are not aware of the existence of a similar enterprise anywhere else - as semantic web applications in this field are just beginning to emerge. For now, agreements are signed between our universities (Tours and Nouakchott), both in the computer science side and the social science side.AFD[15] currently funds a training campaign for librarians of the IMRS[16] on cataloguing documents, and the Mauritanian government is going to support all the needed local actions. Concerning the semantic web level, we are building an ontology

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

[14] `http://data.bnf.fr/en/semanticweb`

[15] Agence Française de Développement: `http://www.afd.fr/lang/en/home`

[16] Institut Mauritanien de recherches scientifiques: `http://www.imrs.mr/spip.php?page=sommaire_fr`

for the IMRS' manuscripts [8], a part of which is already digitized, and we plan to work on designing and building an annotation tool based on this ontology. In order to include the European side (archives on these countries), we are thinking about a MSC Action (deadline in january, 2016). The campaign of partnerships with already existing materials is still to be done, as we must first build the semantic web tools that we should propose to them.

References

1. Abdel Belaïd and Nazih Ouwayed. Segmentation of ancient arabic documents. *Guide to OCR for Arabic Scripts*, pages 2–16, 2011.
2. Beatrice Bouchou and Cheikh Niang. Semantic mediator querying. In *International Database Engineering and Applications Symposium (IDEAS)*, pages 29–38. ACM, 2014.
3. 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. In *20th International Conference on Pattern Recognition (ICPR)*, pages 1929–1932, 2010.
4. Stefanie Brinkmann and Beate Wiesmüller, editors. *From Codicology to Technology: Islamic Manuscripts and Their Place in Scholarship*. Frank and Timme GmbH, 2009.
5. Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Mariano Rodriguez-Muro, Riccardo Rosati, Marco Ruzzi, and Domenico Fabio Savo. The MASTRO system for ontology-based data access. *Semantic Web*, 2(1):43–53, 2011.
6. M. Coustaty, R. Pareti, N. Vincent, and J.M. Ogier. Towards historical document indexing: extraction of drop cap letters. *IJDAR*, 14(3):243–254, 2011.
7. B. Coüasnonet, J. Camillerapp, and I. Leplumey. Access by content to handwritten archive documents: Generic document recognition method and platform for annotations. *IJDAR*, 9(2):223–242, 2007.
8. Mohamed Lamine Diakité and Beatrice Bouchou Markhoff. OMOS: Ontology for Western Saharan Manuscripts. Technical Report 313, Université François Rabelais Tours, Laboratoire d'Informatique (available in HAL: https://hal.archives-ouvertes.fr/hal-01134010), 2015.
9. Martin Doerr and Patrick Le Boeuf. Modelling intellectual processes: The frbr - crm harmonization. In *Digital Libraries: Research and Development, volume 4877 of Lecture Notes in Computer Science*, pages 114–123. Springer, Berlin / Heidelberg, 2007.
10. A. Jordanous, K. F. Lawrence, M. Hedges, and C. Tupman. Exploring manuscripts: Sharing ancient wisdoms across the semantic web. In *2nd International Conference on Web Intelligence, Mining and Semantics (WIMS)*, pages 678–683. ACM, New York, 2012.
11. A. Khemiri, A. Kacem, and Belaid A. Towards arabic handwritten word recognition via probabilistic graphical models. In *Frontiers in Handwriting Recognition (ICFHR)*, pages 678–683, 2014.
12. Cheikh Niang, Béatrice Bouchou, Yacine Sam, and Moussa Lo. A Semi-Automatic approach For Global-Schema Construction in Data Integration Systems. *IJARAS*, 4(2):35–53, 2013.
13. Dominic Oldman. *The CIDOC Conceptual Reference Model (CIDOC-CRM): A Primer, Version 1*. CIDOC CRM (http://www.cidoc-crm.org/docs/CRMPrimer_v1.1.pdf), 2014.
14. Desmond Schmidt. Towards an interoperable digital scholarly edition. *Journal of the Text Encoding Initiative [http://jtei.revues.org/979]*, 7, 2014.
15. M. O. Soulah and M. Hassoun. Which metadata for ancient arabic manuscripts cataloguing? In *International Conference on Dublin Core and Metadata Applications, The Hague, Netherlands*, 2011.
16. L. Werner. Mauritania's manuscripts. *Saudi Aramco World*, 54(6):2–16, 2003.