Construction semi-automatique d'une ontologie sur des manuscrits ouest sahariens

Mohamed Lamine Diakité, Béatrice Bouchou Markhoff

Construction semi-automatique d'une ontologie sur des manuscrits ouest sahariens

Mohamed Lamine DIAKITÉ[1], Béatrice BOUCHOU MARKHOFF[2]

[1]DMI, Université des Sciences de Technologie et de Médecine, Nouakchott, Mauritanie,
diakite@ustm.mr
[2] LI, Université François Rabelais de Tours, France,
beatrice.bouchou@univ-tours.fr

Résumé : Dans le cadre de la sauvegarde et de la valorisation des documents patrimoniaux, des campagnes de numérisation des manuscrits anciens ont été entreprises dans différents endroits notamment dans une partie de l'ouest africain. Ces campagnes de numérisation ont généré un nombre important des ressources numériques potentiellement riches en informations que les chercheurs en sciences humaines et sociales et le grand public désireraient exploiter. Dans cet article, nous proposons un moyen d'accès à toutes les informations sur les manuscrits qui soit plus riche que ceux disponibles dans les catalogues. Pour cela, nous avons construit de façon semi-automatique une ontologie regroupant les connaissances sur les manuscrits. Les différentes étapes suivies dans la construction de l'ontologie allant de l'acquisition des connaissances à partir d'un certain nombre de ressources jusqu'à son enrichissement semi-automatique à partir d'un thésaurus sont présentées. Nous avons par la suite procédé à son alignement avec certaines ontologies de référence.

Mots-clés : manuscrits arabes anciens, ontologie, thésaurus, CIDOC-CRM, FRBRoo.

1 Introduction

Dans le cadre de la sauvegarde et de la valorisation du patrimoine culturel, on a assisté ces dernières années à des campagnes de numérisation des documents manuscrits anciens qui constituent l'héritage culturel des nations. Cette campagne de numérisation a eu comme corollaire la génération d'un nombre important des ressources numérisées potentiellement riches en informations que les chercheurs en sciences humaines et sociales et le grand public désireraient exploiter. En effet, certaines institutions détentrices de ces manuscrits étaient caractérisées par une absence de moyens et de bonnes conditions de conservation ce qui avait comme conséquences la mise en danger de ces manuscrits à cause de leur exposition à la poussière, aux insectes, aux pillages, etc. La conservation pérenne de ces fonds patrimoniaux passait donc par leur numérisation.

C'est dans ce contexte que le projet BIBLIMOS (BIBLIothèque digitale Multilingues des sources inédites de l'Ouest Saharien) a vu le jour avec pour ambition la mise à la disposition des chercheurs mais aussi du grand public des corpus thématiques d'archives privées et publiques relatifs à l'ouest saharien.

Au delà de la seule conservation, l'idée est donc, une fois les manuscrits numérisés, de faciliter leur accès aux chercheurs et autres utilisateurs à travers une plateforme et selon une organisation thématique. Le problème qui se pose à ce niveau est qu'on se retrouve confronté à une grande masse d'informations caractérisées par une grande hétérogénéité dans les documents (mélange de textes, de graphiques, etc.). Ceci rend donc difficile toute exploitation automatique des manuscrits par le contenu. L'objectif que nous nous sommes fixé est d'offrir une description qui permettra aux utilisateurs d'exprimer plus précisément leurs recherches et ainsi permettre un meilleur accès au contenu. Pour cela nous avons proposé une modélisation

des connaissances sur les manuscrits à travers une ontologie. L'activité de modélisation s'est appuyée sur l'aide des experts du domaine, le catalogue contenant les métadonnées sur les manuscrits et sur l'exploitation d'un thésaurus. Ainsi l'accès au contenu de ces manuscrits se fera par le biais de l'ontologie.

2 L'ontologie comme moyen d'une meilleure exploitation du contenu des manuscrits

Le format image dans lequel les manuscrits sont sauvegardés et leur nombre élevé font qu'il est difficile de pouvoir exploiter directement leur contenu de façon satisfaisante. La plupart des bibliothèques numériques en ligne ont opté principalement pour une recherche par mots clés associée à une présentation des résultats sous la forme d'images de pages numérisées. Parmi ces bibliothèques numériques, on peut citer la base de données OMAR (Oriental MAnuscripts Resource) développée à l'université de Freiburg en Allemagne[1]. Cette base de données contient 2500 manuscrits mauritaniens en arabe sur différentes thématiques. Ces manuscrits peuvent être visualisés en ligne et téléchargés en format PDF.

Les résultats donnés par de tels systèmes sont souvent imprécis et parfois l'utilisateur n'a aucune information sur les termes à utiliser pour formuler ses requêtes.

Nous pensons que l'approche ontologique, en permettant d'exprimer un ensemble de connaissances sur les manuscrits, améliorera la qualité de l'accès au contenu en permettant une formulation plus riche pour les requêtes utilisateur et par conséquent des réponses plus précises aux requêtes.

Nous retrouvons dans la littérature des propositions orientées ontologies pour l'accès par le contenu, qui consistent à associer des informations ou des annotations, extraites manuellement ou automatiquement, aux images des documents. Quelques exemples sont décrits dans (COÜASNON & CAMILLERAPP, 2003) et (COUSTATY, 2011).

Pourtant le potentiel important des ontologies, et plus généralement du web sémantique, pour l'exploitation de manuscrits anciens est bien illustré par exemple dans (JORDANOUS et al. 2012), à propos du projet SAWS (Sharing Ancient WisdomS), dans lequel des informations sémantiques sont extraites de documents au format TEI. Les documents sont issus de collections relatives aux anciennes sagesses grecques et arabes et les informations, récupérées sous forme de triplets RDF, sont des relations entre les contenus, formant un réseau conceptuel interrogeable par les chercheurs.

3 Démarche de construction de l'ontologie

Pour construire l'ontologie nous avons identifié deux niveaux de connaissances pouvant caractériser les manuscrits. Le premier niveau correspond aux connaissances descriptives qui se rapportent à l'exploitation des manuscrits indépendamment de leur contenu et le deuxième niveau est relatif aux connaissances issues du contenu même des manuscrits et sont plus difficiles à acquérir que le premier type de connaissances.

3.1 Connaissances descriptives

Les connaissances externes au contenu des manuscrits, que nous avons appelées aussi connaissances descriptives, sont des connaissances se rapportant à l'exploitation des manuscrits indépendamment de leur contenu. La seule référence au contenu se trouve éventuellement dans le sujet associé aux manuscrits. Ces connaissances sont issues de l'exploitation des métadonnées existantes dans le catalogue de bibliothèque de l'IMRS[2] et des

[1] `http://omar.ub.uni-freiburg.de/`

[2] Institut Mauritanien de la Recherche Scientifique.

discussions que nous avons eues avec les experts sur les manuscrits. La modélisation de ces connaissances nous a permis de construire une première version de l'ontologie (figure 1).

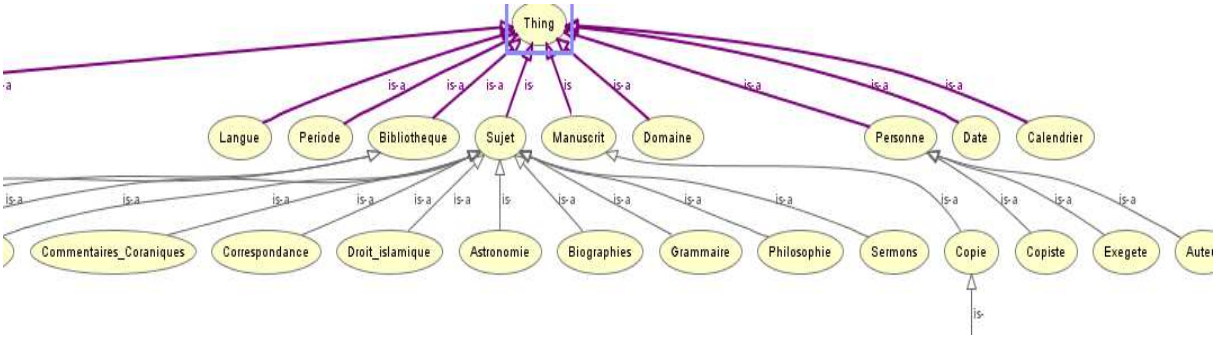

FIGURE 1 - Aperçu d'une partie de l'ontologie sur les manuscrits.

3.2 Connaissances sur le contenu des manuscrits

Ce type de connaissances associé au contenu des manuscrits est plus difficile à modéliser que le premier type de connaissances. Nous développons deux manières de les acquérir:

- soit avec des experts, proposer une modélisation du contenu des manuscrits,
- soit par l'exploitation des différents types de relations existantes dans un thésaurus.

3.2.1 Modélisation du contenu des manuscrits

Pour avoir accès aux connaissances contenues dans les manuscrits, l'apport d'un spécialiste maîtrisant le sujet de leur contenu est primordial. La difficulté d'acquisition de ces connaissances s'explique par le fait qu'elles sont de nature très diverse et sont associées à divers sujets et domaines. C'est donc une tâche qui demande l'intervention d'un grand nombre d'experts de domaines de compétence différents.

3.2.2 Enrichissement semi-automatique de l'ontologie

En complément de l'acquisition des connaissances issues des manuscrits avec l'aide des experts, nous avons développé une méthode d'enrichissement semi-automatique de l'ontologie par l'utilisation du thésaurus RAMEAU[3] mis à disposition en SKOS, donc au format RDF, par la BNF[4].

Il existe plusieurs relations entre les termes dans le thésaurus. Les relations que nous avons exploitées sont les relations hiérarchiques (terme générique et terme spécifique) et les relations d'équivalence (termes équivalents ou alternatifs). La récupération du thésaurus RAMEAU en format RDF s'est effectuée grâce à une requête SPARQL à partir du lien :

`http://data.bnf.fr/sparql`

Afin *d'enrichir la caractérisation des sujets* offerte par notre ontologie (figure 1), nous avons proposé un algorithme implémenté en java et les différentes relations (*skos:broader*, *skos:narrower* et *skos:altLabel*) ont pu être exploitées grâce à la librairie Jena.

L'idée de l'algorithme est de générer pour chaque concept correspondant à un sujet, tous les sous-concepts auxquels le concept est lié par la relation *skos:narrower*. S'il n'y a aucun sous-concept correspondant dans le thésaurus, les sous-concepts de tous les concepts ayant un libellé *skos:altLabel* en commun avec le concept initial sont générés et le choix est laissé à l'utilisateur (ou l'expert) de choisir parmi les sous-concepts générés, ceux qui lui semblent les mieux adaptés.

---

[3] Répertoire d'Autorité Matière Encyclopédique et Alphabétique Unifié.

[4] Bibliothèque Nationale de la France

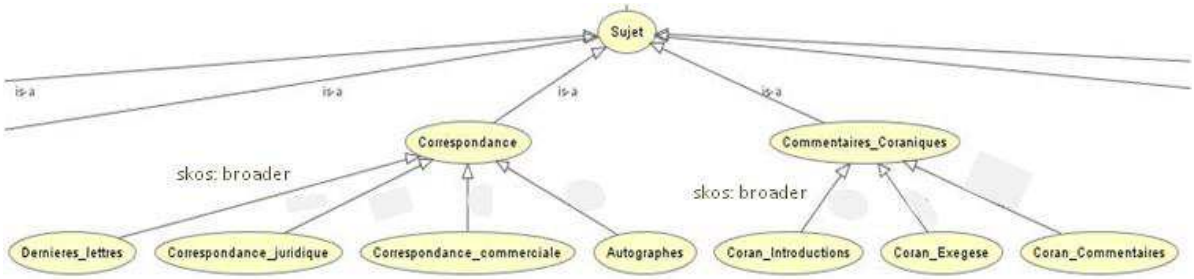

FIGURE 2 - Aperçu d'une partie de l'ontologie après son enrichissement.

La figure 2 illustre l'enrichissement de l'ontologie par des concepts SKOS décrivant les sujets des manuscrits.

4 Discussion sur les résultats obtenus par l'exploitation de RAMEAU

Lors de l'enrichissement de l'ontologie, la difficulté est de trouver pour chaque concept, une entrée correspondante qui est terme préférentiel dans RAMEAU. A défaut il faut chercher tous les termes préférentiels dans RAMEAU qui sont équivalents au label du concept.

Nous avons ainsi observé 4 cas de figure au moment de la comparaison du label d'un concept correspondant à un sujet avec les entrées RAMEAU :

- le cas où il existe une entrée qui concorde avec le label du concept correspondant à un sujet ;
- le cas où il n'existe aucune entrée qui concorde avec le label du concept correspondant à un sujet mais il existe au moins une entrée avec laquelle il existe une relation d'équivalence;
- le cas où il n'existe ni une entrée qui concorde avec le label du concept correspondant à un sujet, ni une entrée qui lui soit reliée par la relation d'équivalence, mais il existe une entrée qui a un sens très proche de celui porté par le label du concept correspondant à un sujet;
- le cas où il n'existe ni une entrée concordant avec le label du concept correspondant à un sujet, ni une entrée qui a un sens proche de celui porté par le label du concept correspondant à un sujet ni une entrée reliée au label du concept correspondant à un sujet par une relation d'équivalence.

Initialement nous avons constitué, à partir de l'exploitation du catalogue, une liste d'une vingtaine de termes correspondant aux labels des concepts correspondant à un sujet. A l'aide d'un programme que nous avons écrit, nous avons vérifié pour chacun des labels sa présence ou non dans le thésaurus. A l'issue de cette vérification, nous avons dû reprendre certains labels qui ne possèdent pas dans RAMEAU d'entrée correspondante, ni de termes qui leur sont reliés par la relation d'équivalence, mais pour lesquels il existe des termes dont la proximité sémantique était évidente pour nous.

Nous avons remplacé ces labels (40% du total des labels) par les termes présents dans RAMEAU qui leur étaient sémantiquement très proches. Par exemple, les concepts *Commentaires* et *Lettres* définis à partir du catalogue renvoient respectivement à *Commentaires coraniques* et *Correspondance* dans le contexte des manuscrits de l'ouest saharien, qui existent dans RAMEAU.

Finalement il y a 15% des labels des concepts correspondant à un sujet qui correspondent au deuxième cas et 70% correspondent au premier cas. Le fait de prendre en compte la proximité sémantique des termes a ainsi permis de ramener de 40 à 70% le nombre de labels qui possèdent une entrée correspondante dans RAMEAU. Dans les 3 premiers cas, l'algorithme permet de générer les sous-concepts du concept correspondant à un sujet. Dans le 4^{ième} cas (qui concernait 15% des labels), aucune génération de sous-concepts n'est possible.

Nous avons constaté que parmi les sous-concepts générés à partir de RAMEAU, certains se prêtent peu au contexte des manuscrits ouest sahariens, du fait que RAMEAU est construit dans un contexte culturel différent. Par exemple, pour le concept *Sermons,* on trouve :

Jésus-Christ -- Passion -- Sermons
Marie, Sainte Vierge -- Sermons
Église catholique -- Sermons

Nous considérons que les sous-concepts générés doivent être en relation avec les manuscrits de l'ouest saharien traitant des sujets portant plus sur des questions en relation avec la religion musulmane.

5 Alignement aux ontologies de référence CIDOC-CRM et FRBRoo

Nous nous sommes fixés comme objectif de rendre l'ontologie interopérable avec d'autres modèles. Nous avons donc procédé à son alignement sur les ontologies de référence CIDOC-CRM[5] et FRBRoo (*Functional Requirements for Bibliographic Records* - Spécifications fonctionnelles des notices bibliographiques).

L'alignement de l'ontologie que nous avons construite avec les ontologies CIDOC CRM et FRBRoo se fait en vérifiant pour chaque concept de notre ontologie, le concept de l'ontologie de référence avec lequel il sera mis en correspondance par une relation hiérarchique (plus spécifique ou plus générique) ou une relation d'équivalence. Nous l'avons fait pour CIDOC CRM d'une part et pour FRBRoo d'autre part, en partant du principe que selon les applications l'un ou l'autre sera utile.

6 Conclusion et perspectives

A notre connaissance, l'ontologie que nous avons construite est la première du genre sur les manuscrits patrimoniaux arabes de l'ouest saharien. Elle permet de formaliser des connaissances explicites et implicites sur les manuscrits. L'acquisition des connaissances suit un processus incrémental, menant à une ontologie modulaire. Ce processus passe à la fois par une interaction avec les experts et par un enrichissement semi-automatique de l'ontologie à partir du thésaurus RAMEAU.

Afin de faciliter l'interopérabilité de notre ontologie avec d'autres modèles, nous utilisons les langages et les modèles du web sémantique et nous alignons l'ontologie sur les ontologies de référence CIDOC-CRM et FRBRoo dédiées au patrimoine culturel.

Dans sa version actuelle, notre ontologie contient les informations qui se trouvaient dans le catalogue, enrichies par les modélisations issues des experts. Nous avons fait vérifier par les experts que toutes les informations ainsi explicitées sont couvertes par les concepts définis dans l'ontologie. Elle doit bientôt être confrontée aux manuscrits au cours de campagnes d'annotation, ce qui constituera une autre forme de validation. Dans un premier temps nous avons aussi utilisé un outil en ligne appelé OOPS![6] (OntOlogy Pitfall Scanner) pour y détecter des erreurs courantes dans la construction d'ontologie.

L'ontologie que nous avons développée dans ce travail est donc une première version d'une ressource destinée à croître, laquelle est elle-même une première étape servant de base à une série d'actions à mener pour réaliser les objectifs du programme BIBLIMOS, à savoir la création d'un portail web dynamique support d'un réseau d'informations autour de l'histoire de l'Ouest-saharien. La prochaine étape, en cours, est la construction à partir de l'ontologie d'un outil d'annotation des manuscrits s'appuyant sur les technologies du web sémantique. En complément de l'outil, nous étudions la spécification de protocoles d'annotation, inspirés de

[5] `http://www.cidoc-crm.org/`
[6] `http://oops.linkeddata.es/`

ceux utilisés dans le domaine du traitement automatique du langage naturel pour l'annotation des corpus de textes.

Références

BANU A., FATIMA S. S. ET KHAN K.R. (2013). Building OWL Ontology: LMSO-Library Management System Ontology. N. Meghanathan et al. (Eds.): Advances in Computing & Inf. Technology. AISC 178. pp. 521–530.

BELAÏD A. & OUWAYED N. (2011). Segmentation of ancient Arabic documents. Volker Märgner and Haikal El Abed. Guide to OCR for Arabic Scripts. Springer.

BIBLIOTHEQUE NATIONALE DE FRANCE (2012). Fonctionnalités requises des notices bibliographiques. Traduction française de *Functional Requirements for Bibliographic Records* : final report. 2ième édition française.

BOUSSELLAA W., ZAHOUR A., TACONET B., BENABDELHAFID A., ALIMI A. (2006). Segmentation texte /graphique : Application aux manuscrits Arabes Anciens. Colloque International Francophone sur l'Écrit et le Document CIFED'06.

CHRISMENT C., HERNANDEZ N., GENOVA F., MOTHE J. (2006). D'un thesaurus vers une ontologie de domaine pour l'exploration d'un corpus. AMETIST, INIST, Vol. 0 : p. 59–92.

COÜASNON B. & CAMILLERAPP J. (2003). Accès par le contenu aux documents manuscrits d'archives numérisés. Document numérique. Volume 7 – n° 3-4/2003, pages 61 à 84.

COUSTATY M. (2011). Contribution à l'analyse complexe de documents anciens Application aux lettrines. Rapport de mémoire de thèse de l'université de la Rochelle.

COUSTATY M., RAVEAUX R. et OGIER J.M. (2012). Historical document analysis: A review of french projetcs and open issues. 19th European Signal Processing Conference (EUSIPCO 2011). Barcelona.

FAROU B., HALLACI S. & AL. (2009) Système Neuro-Markovien pour la Reconnaissance de l'Écriture Manuscrite Arabe à Vocabulaire Limité. Conférence Internationale sur l'Information et ses Applications (CIIA'09). Saida, Algerie. 3-4 mai.

GRUBER T.R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition Academic Press Inc. 5(2).

INTERNATIONAL WORKING GROUP ON FRBR AND CIDOC CRM HARMONISATION (2013). FRBR object-oriented definition and mapping from FRBR ER , FRAD and FRSAD (version 2.0).

ISAAC A.& BOUCHET T. (2009). RAMEAU et SKOS. in *Arabesques*. 54, pp. 13-14.

JORDANOUS A., HEDGES M., LAWRENCE K.F., TUPMAN C. (2012). Exploring Manuscripts : Sharing Ancient Wisdoms across the Semantic Web. International Conference on Web Intelligence, Mining and Semantics (WIMS'12). Craiova, Romania.

JOURNET N. (2006). Analyse d'images de documents anciens : une approche texture. Rapport de mémoire de thèse de l'université de la Rochelle.

KERGOSIEN E. (2011). Point de vue ontologique de fonds documentaires territorialisés indexés. Rapport de thèse de doctorat en informatique de l'université de Pau et des Pays de l'Adour.

NF ISO 21127, Information et documentation (2007). Une ontologie de référence pour l'échange d'informations du patrimoine culturel.

NIANG C., BOUCHOU B., SAM Y. and LO M. (2013).A Semi-Automatic Approach For Global-Schema Construction in Data Integration Systems. IJARAS,Vol. 4(2). pp. 35—53.

PRADEL C., HERNANDEZ N., KAMEL M., ROTHENBURGER B. (2012). Une ontologie du Cinéma pour évaluer les applications du Web Sémantique. in IC 2012.

SHVAIKO P.& EUZENAT J. (2013). Ontology Matching: State of the Art and Future Challenges. IEEE Transaction on Knowledge and Data Engineering. Vol. 25(1), pp. 158—176.

SOULAH M.O. & HASSOUN M. (2011). Which metadata for Ancient Arabic Manuscripts Cataloguing? Proc. International Conference on Dublin Core and Metadata Applications, The Hague, The Netherlands.

WACHE H., VÖGELE T., VISSER U., STUCKENSCHMIDT H., SCHUSTER G., NEUMANN H. and HU□BNER S. (2001).Ontology-Based Integration of Information – A Survey of Existing Approaches. IJCAI Workshop on Ontologies and Informations Sharing.pp 108–117.

WERNER L. (2003). Mauritania's Manuscripts. Saudi Aramco World. Vol. 54, No. 6. pp 2–16.