LinkedMDR: un modèle sémantique de représentation de corpus de documents multimédia

Nathalie Charbel [1], Christian Sallaberry [2], Sébastien Laborie [1], Gilbert Tekli [4], Richard Chbeir [1]

1. UNIV PAU & PAYS ADOUR, LIUPPA, EA3000, 64600 ANGLET, FRANCE
{nathalie.charbel,sebastien.laborie}@univ-pau.fr,rchbeir@acm.org

2. UNIV PAU & PAYS ADOUR, LIUPPA, EA3000, 64000 PAU, FRANCE
christian.sallaberry@univ-pau.fr

3. UNIVERSITY OF BALAMAND (UOB), 100 TRIPOLI, LEBANON
gilbert.tekli@fty.balamand.edu.lb

RÉSUMÉ. *Dans le domaine du BTP, les projets de construction impliquent l'échange d'un volume important d'informations entre divers acteurs ayant des domaines d'expertises et des intérêts différents. La plupart des données échangées au sein de tels projets sont non ou semi-structurées, présentées dans des documents hétérogènes (souvent multimédia tels que des plans ou des rapports) et proviennent de sources variées. Bien évidemment, ces documents sont liés les uns aux autres par des liens explicites (p.ex., des références à tout ou partie de documents introduites par l'auteur) ou bien implicites (p.ex., selon les thèmes abordés dans les documents, tels que la plomberie, l'éléctricité ou l'isolation thermique du bâtiment). Identifier ce réseau de données liées entre documents tout au long de l'évolution d'un projet de construction, de l'indexation jusqu'à la recherche d'information, est aujourd'hui primordial pour faciliter la tâche d'un maître d'ouvrage ou d'un maître d'œuvre. Pour mener à bien cet objectif, dans cet article nous décrivons une nouvelle ontologie intitulée LinkedMDR (Linked Multimedia Document Representation). Cette ontologie est fondée sur l'intégration d'éléments issus de plusieurs standards de description de métadonnées multimédia dont Dublin Core (DC), Text Encoding Initiative (TEI) et Multimedia Content Description Interface (MPEG-7). Nous proposons de lier les standards de description les uns aux autres grâce à notre ontologie tout en fournissant de nouveaux concepts et relations non-pris en charge actuellement par ces standards. Cette représentation unifiée des documents nous permet donc de représenter sémantiquement un réseau de données liées sur un corpus documentaire. LinkedMDR est générique et offre une couche permettant de se spécialiser sur un domaine d'application métier (dans notre cas le BTP). Des expérimentations ont été menées afin de mesurer la qualité de notre proposition au regard d'autres solutions exploitant les standards de métadonnées multimédia actuels.*

ABSTRACT. *Projects, in the construction industry, involve the exchange of a large amount of information between several actors having different expertise and interests. Most of this information is unstructured, originated from different sources and dispersed across heterogeneous documents, thus producing implicit and explicit dependencies between them. This becomes very critical as it makes the annotation of the documents and the information retrieval more challenging at any stage of a building life cycle. In this work, we propose LinkedMDR: a novel ontology for Linked Multimedia Document Representation. Our ontology is based on the integration of the three standards addressing metadata and content representation: Dublin Core (DC), Text Encoding Initiative (TEI), and Moving Picture Experts Group (MPEG-7) together with the addition of new components offering more features especially in representing the collective knowledge of a document corpus. LinkedMDR is generic and offers, as well, a pluggable layer handling the particularities of a domain-specific knowledge. Experiments measure the efficiency and the effectiveness of our solution in comparison with the existing standards.*

MOTS-CLÉS : *Documents hétérogènes; Modèle de documents; Ontologies; Système de Recherche d'Information*

KEYWORDS: *Heterogeneous documents; Document Representation; Ontologies; Information Retrieval System*

1. Introduction

L'émergence des constructions durables, des architectures respectueuses de l'environnement économes en énergie ainsi que l'urbanisme moderne a conduit le domaine du BTP à suivre des approches communes de conception de projets immobiliers. Généralement, le processus de construction de tels projets est décomposé en trois phases essentielles : (i) la phase d'étude et de conception détaillée qui comprend la création, la préparation, l'analyse et la spécification des documents techniques et des plans d'exécution, (ii) la phase de construction et (iii) la phase opérationnelle qui couvre l'utilisation du bâtiment ainsi que la maintenance et le suivi du projet achevé (Klinger, Susong, 2006). Durant ce cycle de vie, de multiples acteurs (maître d'ouvrage, maître d'œuvre...) sont impliqués dans le processus de construction. Ils contribuent et échangent une grande variété de documents techniques et administratifs selon leurs expertises et leurs rôles au sein du projet. Par exemple, les contrats, les rapports techniques, les CCTP (Cahiers des Clauses Techniques Particulières), les CCAP (Cahiers des Clauses Administratives Particulières), les plans ainsi que les photos d'un projet sont généralement partagés durant les différentes phases de construction. Il apparaît très souvent que les documents échangés, provenant donc de sources différentes, n'ont pas de structure commune que ce soit entre les mêmes types de documents d'un projet (p.ex. des rapports techniques de type texte) ou que ce soit entre des documents similaires (p.ex., rapports thermiques) de projets de construction différents. Egalement, ils ont des versions, des formats d'encodage hétérogènes (p.ex., pdf, docx, xlsx, jpeg, etc.), des types de média différents (p.ex., images ou textes), évoquant des domaines métiers variés (p.ex., architecture, électricité, plomberie, mécanique, maçonnerie, etc.). De plus, ces documents peuvent comporter des références ainsi que des liens intra ou

inter-documentaires [1] de nature implicite ou explicite.
Dans la littérature, plusieurs travaux ont été engagés pour définir des métadonnées sur les documents et leur contenu. Ces modèles d'annotation peuvent être classés selon qu'ils traitent des contenus de type texte (e.g., TEI [2]), image (e.g., EXIF [3]) ou encore multimédia (e.g., (Arndt *et al*., 2007 ; Saathoff, Scherp, 2010 ; Garcia, Celma, 2005 ; Brut *et al*., 2009 ; Bloechle *et al*., 2006)). Néanmoins, aucun des travaux existants ne considère (i) un ensemble de documents multimédia hétérogènes, (ii) à la fois des annotations d'images et de textes qui portent sur leur contenu ainsi que leur structure, (iii) des spécificités liées à certains documents comme, par exemple, les légendes de plans, et (iv) différents liens inter et intra-documentaires. De plus, les standards actuels exploités dans le domaine du bâtiment, tel que l'IFC, se focalisent principalement sur des modèles de représentation d'objets 3D et ne prennent pas en considération des documents multimédia classiquement exploités durant un projet de construction.
Dans ce contexte, notre défi consiste à disposer d'une vue la plus claire et complète possible de ce réseau de données liées issu de documents multimédia relatifs à un projet immobilier. En effet, il est primordial de fournir aux différents acteurs un système d'information permettant de leur renvoyer des données souhaitées mais aussi et surtout de parcourir ce réseau de données en fonction de leur besoin et de leur expertise. Pour ce faire, nous devons au préalable définir un modèle qui permet de représenter ce réseau ainsi que la variété des données et des connexions qui le compose. Afin d'assurer l'interopérabilité des annotations, ce modèle supportera des concepts et des relations sémantiques exploitables lors de l'indexation, de la construction du réseau ou encore lors de la recherche d'information. En effet, la sémantique permettra aux différents acteurs de disposer d'informations pertinentes vis-à-vis de leurs expertises et de leurs préférences.
Dans cet article, nous proposons donc LinkedMDR : un modèle sémantique de représentation de réseau de données liées entre documents multimédia. LinkedMDR repose sur la combinaison d'éléments de standards de métadonnées existants tout en liant ces standards les uns aux autres, et en fournissant de nouveaux concepts et relations non-pris en charge actuellement par ces standards. Notre proposition est développée en collaboration avec la société Nobatek [4], une société française dans le secteur de la construction durable dont la mission consiste à assurer le transfert d'outils, de méthodes, de procédés et de produits innovants afin de contribuer à la performance énergétique et à la qualité environnementale. Dans ce cadre, des expérimentations ont été menées afin de mesurer la performance ainsi que l'efficacité de notre proposition au regard des autres solutions exploitant les standards de descriptions multimédia actuels. La présentation de notre contribution sera la suivante. La section 2 présentera plus en détail nos motivations au travers de situations réelles concrètes décrites par la société Nobatek.

1. Un lien inter-documentaire est un lien entre différents documents, tandis qu'un lien intra-documentaire est un lien entre deux éléments d'un même document.

2. Text Encoding Initiative, TEI P5 Guidelines for Electronic Text Encoding and Interchange, http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

3. Exchangeable image file format for digital still cameras, http://www.exif.org/Exif2-2.PDF.

4. http://www.nobatek.com

La section 3 fera état (i) des travaux existants dans le domaine de la représentation des métadonnées et des contenus, et (ii) des standards utilisés dans le domaine du bâtiment. Ces travaux ne satisfaisant pas des contextes métiers spécifiques, nous décrirons dans la section 4 notre ontologie nommée LinkedMDR et des résultats expérimentaux seront illustrés dans la section 5. La section 6 conclura cet article et développera de futures perspectives à envisager pour notre travail.

2. Motivations

Nous allons présenter un exemple de scénario de projets de construction qui nous a été fourni par la société Nobatek (§2.1). Ce scénario nous permet de dégager plusieurs défis qu'il conviendra de satisfaire par la suite (§2.2).

2.1. Le contexte

La société Nobatek est une société consultante qui assiste les maîtres d'ouvrages mais également les maîtres d'œuvres à développer leur projet de construction. Tout au long du cycle de vie d'un projet immobilier, cette société communique avec d'autres partenaires de construction, des bureaux d'études techniques, d'architecture, etc. Chaque acteur dispose donc de sa propre expertise que ce soit dans le domaine de la maçonnerie, de l'électricité, de la plomberie, etc. Dans ce contexte, de multiples documents (rapports techniques, plans...) sont élaborés faisant très souvent des références les uns aux autres. Par exemple, la figure 1 présente différents documents hétérogènes relatifs à un projet immobilier. Comme il est possible de le constater dans cette figure, il y a plusieurs rapports qui décrivent notamment les lots techniques du bâtiment (d_1 et d_5), ses propriétés thermiques (d_2) et acoustiques (d_3) ainsi qu'un extrait de plan d'étage du projet (d_4) et une photo (d_6). Actuellement, les ingénieurs de la société Nobatek doivent parcourir manuellement l'ensemble de ces documents. En effet, si ces derniers désirent vérifier la compatibilité des propriétés des façades extérieures avec les critères exigés par les normes environnementales, ils devront chercher d'eux-mêmes les informations dans ces documents rédigés par divers spécialistes (bureau d'étude thermique, bureau d'étude acoustique...). Cette recherche d'information est très fastidieuse et, de part le volume d'information important, peut conduire à ne pas consulter certains documents qui pourraient pourtant s'avérer utiles.

2.2. Les défis

– **Défi 1 : Représenter un réseau de données liées issu des documents** - Les ingénieurs doivent pouvoir rechercher et parcourir un réseau d'informations construit à partir d'un ensemble de documents. Il est évident de constater dans la figure 1 que les documents d_i ont de multiples relations les uns aux autres (p.ex., références, thématiques partagées, versions...). Ces relations peuvent être également de nature implicites ou explicites. Par exemple, certaines sections de texte entre d_1 et d_3 sont en relation implicite puisqu'elles décrivent la même thématique (p.ex., les façades extérieures). Le document d_5 quant à lui fait référence explicitement au plan technique

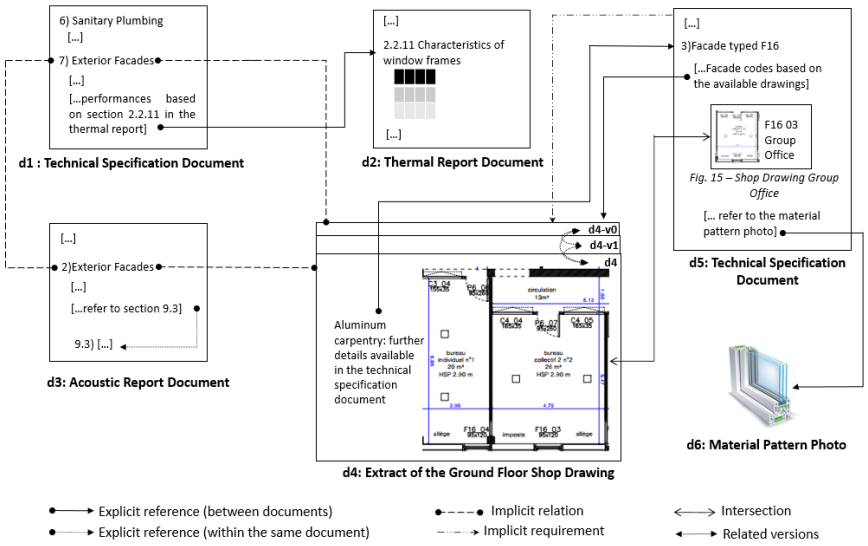

FIGURE *1. Exemple de documents multimédia hétérogènes relatifs à un projet de construction.*

d_4 (tout en sachant que de multiples versions de d_4 existent) et à la photo d_6. Par conséquent, il sera nécessaire d'indexer ces documents et d'en analyser les métadonnées afin d'établir ce réseau de relations.

– **Défi 2 : Exploiter la sémantique de l'information** - Les ingénieurs ont besoin de trouver avec différents niveaux de précision une information pertinente contenue au sein de documents évoquant de multiples thématiques. Par exemple, d_5 contient des données générales sur les façades extérieures, tandis que d_1 et d_2 contiennent des données plus détaillées sur ces façades extérieures avec d'autres éléments, tels que la plomberie et l'électricité. En effet, d_1 décrit les propriétés des façades comme pour la plomberie ou l'électricité, alors que d_2 se focalise particulièrement sur les propriétés thermiques des façades qu'elles soient intérieures ou extérieures. Outre la richesse des termes employés, il conviendra d'analyser les métadonnées d'ordre général ainsi que les structures des documents qui décrivent ces informations selon différents niveaux de granularité (p.ex., section, paragraphe ou phrase).

– **Défi 3 : Considérer la multimodalité des documents** - Les ingénieurs travaillent avec des documents hétérogènes ayant des types ainsi que des formats différents. Par exemple, d_1, d_2 et d_3 sont des documents Word, d_4 est un plan de construction, d_5 un document PDF et d_6 une photo au format JPG.

– **Défi 4 : Assurer l'extensibilité de l'information** - Les ingénieurs peuvent travailler sur d'autres types de construction avec d'autres types d'information et de média. Par exemple, dans un autre projet de construction que celui de la figure 1, ils peuvent

être amenés à analyser des documents sonores (de type audio) au sujet de bruits ambiants d'une pièce avant ou après la mise en place d'un revêtement spécifique sur les façades d'un bâtiment.

Dans ce qui suit, nous démontrons que les travaux de recherche actuels ne couvrent que partiellement ces défis au sein d'un même système d'information.

3. État de l'art

Nous allons présenter les standards et modèles existants de représentation de métadonnées, de structures et de contenus de documents multimédia (§3.1). De plus, nous décrirons un standard correspondant au domaine spécifique du bâtiment (§3.2). Nous conclurons cette partie par une comparaison ainsi qu'une discussion sur les limites de ces standards de description (§3.3).

3.1. Les standards et modèles existants de représentation de documents

Dublin-Core [5] - "Dublin Core Metadata Initiative" (DC) est un standard de métadonnées d'ordre général (p.ex., titre, date de création, format) décrivant une grande variété de documents. Il comprend 15 éléments ainsi que des composants, appelés "qualifiers", permettant le raffinement de ces éléments.

MPEG-7 [6] - "Multimedia Content Description Interface" est un standard décrivant différents types de contenu (p.ex., une image, une vidéo, un son). Il comprend trois composants principaux : les descripteurs (Ds) décrivant des éléments de base du contenu (p.ex., la couleur, la texture), les schémas de description (DSs) décrivant la structure et la sémantique des relations entre Ds et entre DSs, et le langage de définition de description qui est fondé sur les schémas XML (DDL).

Les modèles ontologiques - De nombreuses initiatives ont été menées pour spécifier des ontologies de description de documents multimédia. L'objectif principal de toutes ces approches consiste à combler le fossé entre les descripteurs de bas niveau, généralement extraits automatiquement par des indexeurs, et ceux de haut niveau exploités par les humains et décrivant la même information (Suarez-Figueroa *et al*., 2013). Comme indiqué dans les travaux de (Scherp *et al*., 2012), il est nécessaire de combiner plusieurs standards afin de disposer d'un système d'information multimédia le plus complet possible. Cette situation a donc ouvert la voie à la spécification d'ontologies dites multimédia. Par exemple, l'ontologie COMM (Core Ontology for MultiMedia) (Arndt *et al*., 2007) a été construite pour l'annotation de documents multimédia. Cette ontologie est fondée sur le standard MPEG-7, sur l'ontologie DOLCE ainsi que sur deux autres ontologies relatives aux design patterns. D'autres ontologies basées sur MPEG-7

5. Dublin Core Metadata Initiative, Metadata Basics, http://dublincore.org/documents/dcmi-terms/.

6. Multimedia content description interface, Technical report, Standard No. ISO/IEC n15938, 2001, http://mpeg.chiariglione.org/standards/mpeg-7/.

ont bien évidemment vu le jour, telles que MPEG-7 Rhizomik (Garcia, Celma, 2005) ou encore Multimedia Metadata Ontology (M3O) (Saathoff, Scherp, 2010). MPEG-7 Rhizomik fournit des correspondances directes entre le standard MPEG-7 et OWL, tandis que M3O propose un méta-modèle de représentation de documents multimédia qu'il est possible de spécialiser en fonction de son besoin. Le groupe de travail du W3C "Media Annotation Working Group" a également spécifié une ontologie intitulée "Media Resource Ontology" [7]. Cette dernière propose divers alignements avec les standards de métadonnées existants, tels que MPEG-7, Dublin Core et EXIF [8].

XCDF Format - XCDF est un format utilisé pour la représentation des résultats d'extraction et d'analyse des structures physiques de documents PDF (Bloechle *et al*., 2006). Ce format est basé sur le langage XML et sa DTD décrit un ensemble d'éléments permettant de représenter un document textuel : page, police, paragraphe, phrase, mot...

EXIF - Bien qu'il existe de multiples standards pour annoter des images, nous nous focaliserons sur le format EXIF (Exchangeable Image File Format) puisqu'il s'agit d'un format très complet permettant de décrire tout ou partie d'une image. En effet, ce langage comporte des éléments descripteurs de structure d'une image (hauteur, largeur, composition en terme de pixels), de version, de caractéristiques de l'image (couleurs, configuration, compression), d'informations sur son créateur (auteur, commentaires), sur le fichier (date de création, données GPS, droits...).

TEI - Il existe également de multiples standards de description de textes. Nous nous focaliserons sur le standard TEI (Text Encoding Initiative), basé sur XML. Le format TEI n'est pas seulement basé sur la structure du texte et ses annotations, il permet de faire référence à des concepts sémantiques qui facilitent la recherche d'information. Il est possible de classifier ces éléments selon les catégories suivantes : éléments sur la structure (p.ex., chapitres, sections, paragraphes, listes, tables), la mise en forme (polices de caractères), les annotations (titre, date, abréviations, signets, renvois), les figures et les graphiques.

3.2. Les standards de descriptions dans le domaine du bâtiment

Le standard IFC [9] (Industry Foundation Classes) est un des standards les plus utilisés pour l'échange de données BIM (Building Information Modeling) dans le domaine du bâtiment. Il contient toutes les informations utiles, comme les composants physiques d'un bâtiment, les espaces, les systèmes, les processus, les acteurs, ainsi que l'ensemble des relations entre ces éléments (Huovila, 2012). Les spécifications IFC peuvent être sérialisées en XML suivant un schéma XSD ou en EXPRESS, un autre langage de définition de données.

7. W3C, Ontology for media resource 1.0, http://www.w3.org/TR/mediaont-10/.

8. Exchangeable image file format for digital still cameras, http://www.exif.org/Exif2-2.PDF.

9. Industry Foundation Classes, IFC4 Add1 Release, http://www.buildingsmart-tech.org/specifications/ifc-releases/ifc4-add1-release.

3.3. Discussion

Nous avons étudié les standards et modèles existants de description de documents au regard du contexte ainsi que des défis que nous avons énumérés dans la section 2. Nos résultats sont synthétisés dans la Table 1.

TABLE *1. Synthèse des propriétés des standards et modèles existants au regard des défis dégagés.*

Défis	Standards et Modèles / Propriétés		Représentation de Métadonnées et de Contenus								
			Orientés Multimédia							Orientés Image	Orientés Texte
			Dublin Core	MPEG-7	COMM	M3O	MediaOnt	Mpeg-7 Rhizomik	Format Canonique XCDF	EXIF	TEI
Défi 1	Représentation d'un réseau sémantique de documents		x	x	x	x	x	x	x	x	x
	Description de relations	Intra-documentaire	Partielle	Partielle	Partielle	Partielle	Partielle	Partielle	x	x	Partielle
		Inter-documentaire	Partielle	Partielle	Partielle	Partielle	Partielle	Partielle	x	x	Partielle
Défi 2	Représentation de métadonnées descriptives		√	√	√	√	√	√	x	√	√
	Description de contenu		x	√	√	√	x	√	x	x	√
	Représentation de métadonnées de structure	Image	x	√	√	√	x	√	x	x	x
		Texte	x	Partielle	Partielle	Partielle	x	Partielle	Partielle	x	√
Défi 3	Multi-modalité		√	x	x	x	x	x	x	x	x
Défi 4	Extensibilité		√	√	√	√	√	√	√	x	√

Il est évident que les standards qui se focalisent exclusivement sur le texte ou bien les images sont limités puisqu'ils ne gèrent pas les différents types de documents multimédia (Défi 3). Néanmoins, ces standards, et notamment la TEI pour le texte, offrent des éléments de descriptions relativement complets et pertinents qu'il convient de réutiliser. Par exemple, la balise <*ref*> pour la TEI permet de faire des références mais celle-ci se limite au simple *confer*. Pour MPEG-7, il existe les éléments *Still Regions* et *Text Annotations* qui pourraient éventuellement servir à représenter globalement les différentes parties des plans techniques des bâtiments avec les légendes textuelles associées. Néanmoins, ces deux éléments ne permettent pas de représenter la structure complète des légendes avec différents niveaux de précision (Défis 2 et 4), ni même de faire des relations avec d'autres descripteurs provenant d'autres standards de représentation, comme la TEI ou DC (Défi 1).
Les langages et modèles de descriptions multimédia sont eux aussi limités. En effet, les ontologies et modèles multimédia agrègent simplement à l'aide de correspondances les descripteurs des standards existants mais n'apportent pas vraiment de plus-value (Défis 1 et 2).
En ce qui concerne les représentations dans le domaine du bâtiment, le standard IFC se focalise principalement sur la représentation 3D d'une construction. Il ne permet donc pas de relever tous les défis que nous avons exposé dans la section 2.2 (Défis 1 à 4).
D'autre part, il existe peu de travaux d'annotation de données multimédia respectant ces standards à des fins de recherche d'information (RI). Par exemple, le projet LINDO

(Large scale distributed INDexation of multimedia Objects) (Brut *et al*., 2009 ; 2011) a relevé le défi d'exploiter différents standards de métadonnées dans son système d'information distribué, tels que Dublin Core, EXIF et MPEG-7. Dans (Bates, 2011), l'auteur propose le langage CQL qui combine des recherches plein-texte et des recherches dans les métadonnées Dublin Core, à base de mots-clés. Enfin, dans (Feng *et al*., 2013), les auteurs présentent le potentiel du standard MPEG-7 à des fins d'annotation et de recherche d'information dans des corpus de documents multimédia.
De manière générale, à notre connaissance, il n'existe pas actuellement de représentation d'un réseau sémantique de données liées dans un corpus documentaire multimédia, ni de travaux de recherche d'information qui visent l'exploitation combinée de descripteurs de contenu thématique et de structure des documents, des relations intra et inter-documentaires, et des métadonnées plus générales. Notre proposition, illustrée dans la partie suivante, permettra de satisfaire ce besoin à travers l'ontologie LinkedMDR. Ce type d'approche est considéré comme le moyen le plus fiable et efficace pour supporter la recherche d'information sémantique à partir de données hétérogènes multimédias (Guo *et al*., 2017).

4. LinkedMDR : un modèle sémantique de représentation de corpus de documents multimédia

Nous proposons une ontologie, intituléé LinkedMDR, pour représenter un corpus de documents multimédia dans un seul modèle de données. LinkedMDR est basée sur (i) l'intégration des standards les plus pertinents en matière de représentation de méta-données et de contenus (notamment DC, TEI et MPEG-7) et (ii) l'ajout de nouveaux concepts et relations dépassant les limites de ces standards. Elle est constituée de trois couches principales : (i) la couche noyau servant de médiateur entre les différentes couches, (ii) la couche intégratrice de méta-données standards comprenant des éléments de standards existants et (iii) la couche spécifique au domaine, qui est adaptée à un domaine d'application particulier tel que le domaine de construction. L'idée de diviser l'ontologie en plusieurs couches et de centraliser les concepts et les relations les plus abstraits dans la couche noyau, assure sa généricité et son extensibilité (voir Figure 2).

4.1. La couche noyau

Cette couche comprend de nouveaux concepts et relations qui n'ont pas été adoptés par les standards existants, soient principalement : (i) les concepts qui modélisent la composition globale d'un document et les propriétés de méta-données qui lui sont associées (p.ex., *Document*, *Media*, *MediaComponent* et *DocumentProperty*), (ii) une entité *Object*, abstraction de *Document*, *Media* et *MediaComponent*, qui induit un riche ensemble de relations potentielles (relations *hasPart*, sémantique, temporelle et spatiale), (iii) des concepts qui généralisent ceux contenus dans la couche méta-donnée standard (*DescriptiveMetadata*, *AdministrativeMetadata* et *TextElement* généralisent des descripteurs de méta-donnée de DC et de structure de documents de TEI, respectivement) et (iv) de nouveaux concepts étendant le potentiel de description des standards MPEG-7 et TEI (*TextStillRegion* hérite de l'élément *TEI:Text* et l'élément

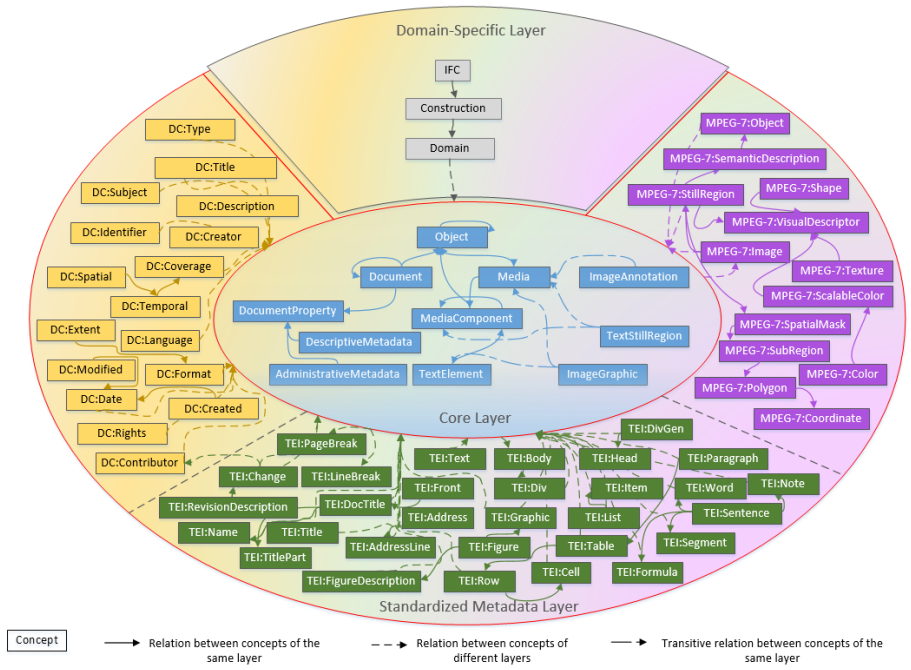

FIGURE 2. *LinkedMDR, le modèle sémantique de représentation de corpus de documents multimédia.*

MPEG-7:StillRegion pour représenter la structure d'une légende textuelle dans une image, *ImageGraphic* hérite de l'élément *TEI:Graphic* et l'élément *MPEG-7:Image* pour représenter les métadonnées d'une image insérée dans un texte et *ImageAnnotation* annote une image à l'aide d'une autre image ou d'un texte structuré).
Dans le reste du document, nous utilisons les triplets RDF [10] pour illustrer des instances de notre ontologie, car RDF est le modèle de données le plus largement utilisé pour représenter un graphe sémantique et extensible.
Ainsi, selon l'exemple de la figure 1, la section 3 (*div3*) du document d_5 contient une figure qui est un extrait de plan du document d_4. Cette relation se traduit par les triplets <*d5, hasPart, d5.div3.figure15*> et <*d4, contains, d5.div3.figure15*>. Cet exemple illustre une réponse apportée aux défis 1, 2 et 3 : il introduit une relation spatiale de contenance (*contains*) entre des documents de types différents. De plus, cet exemple illustre la sémantique attribuée à des niveaux de précision variés, ce qui permet d'inférer d'autres relations enrichissant le réseau de données (p.ex., si <*d5.div3.figure15, contains, dn.div1.figure1*>, on peut déduire que <*d4, contains, dn.div1.figure1*> par transitivité). Cela ne peut pas être réalisé avec les standards existants car ils ne permettent pas de représenter de telles relations à des niveaux de granularité différents.

10. Resource Description Framework, https://www.w3.org/RDF/.

4.2. La couche intégratrice de méta-données standards

Cette couche est constituée d'une sélection de méta-données définies par des standards : Dublin Core, TEI et MPEG-7. Par conséquent, elle se scinde en trois sous-couches, chacune dédiée à un standard. La première correspond à DC et comprend des méta-données d'ordre général relatives à un document. La deuxième présente les méta-données TEI décrivant la structure et le contenu d'un texte. Enfin, la troisième correspond aux méta-données décrivant une image avec ses différentes granularités, caractéristiques visuelles et descripteurs sémantiques suivant le standard MPEG-7. Il est à noter que, dans cet article, nous avons uniquement exploité les méta-données concernant les textes et les images. Cependant, à l'avenir, d'autres médias (comme audio et vidéo) pourront être considérés dans notre ontologie, en particulier dans cette couche.
Cette couche décrit également des relations entre ces différentes sous-couches. Par exemple, nous avons ajouté la relation *isRevisedBy* afin de relier le marqueur *TEI:Change* (décrivant l'ensemble de modifications apportées à un document) au marqueur *DC:Contributor* (celui qui a participé à ces modifications) correspondant. En outre, chaque sous-couche est également reliée à la couche noyau par l'intermédiaire de relations entre leurs concepts respectifs. Par exemple, nous pouvons citer <*TEI:Text, isA, Media*>, <*MPEG-7:StillRegion, isA, MediaComponent*>, <*DC:Title, isA, DescriptiveMetadata*>, <*TextElement, isOn, TEI:PageBreak*>.
Ainsi, selon l'exemple de la figure 1, le document d_4 a un ensemble de propriétés qui peuvent être décrites par la réutilisation des métadonnées de DC. À titre d'exemple, le titre de d_4 peut être traduit par les triplets <*d4, hasProperty, d4.title*> et <*d4.title, hasValue, "Shop Drawing"*>. De plus, d_4 contient des plans d'étage du bâtiment, chacun décrit sur une page du document. De ce fait, la réutilisation des méta-données du standard MPEG-7 peut servir à la description des différentes régions des plans techniques mais sans renseignement sur les pages correspondantes. De même, la réutilisation des méta-données de la TEI sert à la description de la répartition des plans sur les pages du document mais sans information relative au contenu de ces mêmes plans. Ainsi, dans le cadre du défi 1, la liaison entre les méta-données de ces différents standards n'est possible qu'à travers les concepts de la couche noyau comme le montrent les triplets suivants : <*d4, hasPart, d4.imagegraphic*>, <*d4.imagegraphic, isOn, d4.page1*>, <*d4.imagegraphic, hasPart, d4.stillregion*>.

4.3. La couche spécifique au domaine

Bien que notre proposition soit faite dans le contexte du domaine de la construction, nous visons à fournir une ontologie générique qui pourrait être utilisée dans n'importe quel domaine spécifique. Les couches mentionnées précédemment sont génériques et indépendantes du contexte dans lequel les documents multimédia sont utilisés. Toutefois, il est important de tenir compte d'éventuelles particularités relatives à un domaine spécifique. Par conséquent, nous présentons cette couche comme une illustration de la façon dont nous pouvons mettre en œuvre cette ontologie générique tout en l'adaptant à une utilisation ciblée, comme le domaine de la construction, par exemple.

Pour ce faire, nous présentons un nouveau concept intitulé *Domain* et nous le lions au concept *Object* de la couche noyau. De cette façon, des concepts spécifiques à un domaine donné peuvent être ajoutés sous *Domain* et par la suite seront en relation avec les sous-concepts de *Object* (c'est-à-dire, *Document*, *Media* et *MediaComponent*).
Dans cet article, nous présentons un exemple montrant comment nous pouvons rendre cette couche adaptable au domaine de la construction. Nous ajoutons le concept *Construction* comme sous-concept de *Domain*. Nous relions également ce dernier au concept *IFC* qui comprend les concepts de ifcOWL [11], la conversion du standard IFC en ontologie.
À titre d'exemple, la section 7 (*div7*) de d_1 (cf. Figure 1) décrit les façades extérieures. Il est maintenant possible de lier la section 7 avec l'objet IFC correspondant (p.ex., ifcwindow4): <*d1, isA, Document*>, <*d1, hasPart, d1.div7*>, <*d1.div7, isA, TEI : Div*>, <*ifcwindow4, isA, IFC: BuildingElement*> et <*ifcwindow4, isRelated, d1.div7*>. Cela répond particulièrement aux défis 1 et 4.
Pour plus de détails sur l'ontologie LinkedMDR, les différents concepts et relations appartenant à chacune des couches décrites dans cette partie sont disponibles en ligne avec la documentation correspondante : http://spider.sigappfr.org/linkedmdr/.

5. Expérimentation

Nous avons expérimenté l'annotation de documents hétérogènes, liés aux projets de construction dans l'entreprise Nobatek, afin d'évaluer deux critères : (i) la performance de notre modèle de données en terme de concision des annotations et (ii) son efficacité (qualité de l'annotation – rappel, précision, F_1-Mesure) par rapport aux standards existants en matière de représentation de méta-données et de contenu, plus particulièrement DC, TEI et MPEG-7. Nous décrivons tout d'abord le jeu de données de test puis nous commentons les résultats d'expérimentation.

5.1. Données de test

Nous avons sélectionné 6 documents relatifs à des projets de construction présentés dans le scénario de motivation initial (cf. Figure 1). Bien que ce nombre paraisse réduit, ces documents sont choisis à la main et peuvent représenter un scénario complet qui met en valeur tous les défis déjà dégagés (cf. Section 2.2). Nous avons ensuite effectué les cinq expérimentations suivantes :

– **Test#1: Annotation selon DC**

Nous avons utilisé Dublin Core Advanced Generator [12] afin de générer, pour chaque document, une représentation XML correspondant au standard DC. Ce test a permis de générer un fichier d'annotation XML pour chaque document.

11. http://ifcowl.openbimstandards.org/IFC4_ADD1.owl
12. Disponible en ligne sur http://www.dublincoregenerator.com/generator.html

TABLE *2. Évaluation de la concision des annotations dans les différents jeux de tests.*

Groupes de Tests	Nb. de Documents Annotés	Nb. Cumulé d'Annotations	Nb. de Fichiers XML Générés	Nb. de Redondances
Test#1	6	79	6	0
Test#2	4	646	4	0
Test#3	2	198	2	0
Test#4	6	923	12	91
Test#5	6	656	1	0

– **Test#2: Annotation selon TEI**
Nous avons utilisé l'outil OxGarage [13] afin de générer, pour chaque document, une représentation TEI P5 XML correspondant au standard TEI. Puis, nous avons ajouté à la main des références internes et externes (éléments *ptr* et *ref*) pour compléter les annotations. Ce test a permis de générer un fichier d'annotation XML pour chaque document textuel (c'est-à-dire d_1, d_2, d_3 et d_5).

– **Test#3: Annotation selon MPEG-7**
Nous avons utilisé l'outil Caliph V0.9.27 [14] afin de générer, pour chaque document, une représentation XML correspondant au standard MPEG-7. Nous avons ensuite apporté quelques modifications manuelles : nous avons ajouté des éléments qui n'étaient pas décrits par l'outil Caliph V0.9.27 (par exemple, *FreeTextAnnotation* associé à des éléments *StillRegion*). Ce test a permis de générer un premier fichier d'annotation XML pour l'image (d_6) et un second pour l'extrait de plan (d_4).

– **Test#4: Annotation selon DC, TEI et MPEG-7 (combinés)**
Nous avons utilisé les résultats des annotations issues de Test#1, Test#2 et Test#3.

– **Test#5: Annotation selon LinkedMDR**
Nous avons créé des instances de l'ontologie LinkedMDR [15] via Protégée pour construire un fichier RDF représentant tous les documents selon ce modèle.

5.2. Résultats d'expérimentation

5.2.1. Évaluation de la performance

Nous avons évalué la performance sur le plan de la concision des annotations générées par les tests, en considérant l'ensemble des six documents.
Ainsi, nous comparons les annotations fournies par les cinq tests en termes de (i) nombre cumulé d'annotations [16] ; (ii) nombre de documents source annotés ; (iii) nombre de fichiers XML générés ; (iv) nombre de redondances (chevauchement de méta-données). L'objectif est ici de mettre en valeur le scénario qui génère le nombre minimum d'an-

13. Disponible en ligne sur http://www.tei-c.org/oxgarage/. Cet outil ne traite pas les documents PDF, nous avons donc utilisé le service Web PDF to DOCX disponible à l'adresse http://pdf2docx.com/ pour convertir des documents PDF en DOCX.

14. Disponible sur http://www.semanticmetadata.net/

15. Disponible sur http://spider.sigappfr.org/download/1175/

16. Le nombre de balises XML dans les fichiers d'annotation ou le nombre de triplets dans l'ontologie.

notations (sans perte d'information), de fichiers XML et de redondances. Ces résultats sont présentés dans la table 2. Nous constatons que seuls Test#1 (DC), Test#4 (DC, TEI, MPEG-7) et Test#5 (LinkedMDR) ont été en mesure de générer des annotations pour chacun des six documents. DC montre un nombre faible d'éléments d'annotation mais ne couvre que les méta-données génériques sans considérer la structure et le contenu des documents. LinkedMDR, quant à elle, couvre l'ensemble du potentiel d'annotation attendu et, pour autant, génère un nombre réduit d'annotations.
En ce qui concerne les annotations résultant de Test#4 et de Test#5, la table 2 montre de bons résultats pour LinkedMDR puisque ce scénario de test a permis de représenter, dans un même fichier d'annotation, les six documents avec un nombre relativement réduit d'éléments d'annotation et sans redondance de méta-données. Test#4, quant à lui, génère un nombre important d'annotations, parce que TEI et MPEG-7 sont très verbeux, sans toutefois couvrir l'ensemble du potentiel d'annotation attendu.
Les résultats d'annotation seront ensuite exploités à des fins de recherche d'information (RI). Des éléments de méta-données, de structure et de contenus représentés de façon concise, sans redondance, dans un seul document de description, seront d'un intérêt majeur dans des scénarios de RI. Par exemple, pour la requête *« Quels contenus, issus de cahiers de clauses, traitent de façades et font référence à des plans d'étage ? »*, nous pouvons interroger les triplets RDF faisant référence à notre ontologie LinkedMDR. Ces triplets nous permettent de retrouver la section 3 (div 3) du document d_5 puisque d_5 est un CCTP (cahiers de clauses), traite de façades extérieures (façades) et inclut d_4 (plan). Ceci n'est pas possible avec les annotations du Test#4, l'information y est incomplète et répartie de façon indépendante dans plusieurs fichiers d'annotation, selon différents standards.

5.2.2. Évaluation de l'efficacité

Nous avons également évalué l'efficacité de ces modèles en calculant les scores de Précision, Rappel et F_1-Mesure relatifs aux annotations générées pour chacune des séries de tests. Puisque les éléments d'annotation varient entre les standards et LinkedMDR, nous avons fixé un ensemble de critères pertinents sur lesquels nous avons fondé nos calculs indépendamment du nombre d'annotation ou de leur nature (balises XML ou triplets RDF). Ces critères se scindent en plusieurs catégories : liens sémantiques inter ou intra-documentaires, liens topologiques inter-documentaires, méta-données générales ou spécifiques aux textes/images, etc. Nous mesurons des scores de précision [17] et de rappel [18] et de f_1-mesure [19] relatifs à chacun des groupes de test selon ces critères. Dans cette évaluation, nous avons également utilisé deux documents supplémentaires : un fichier audio et une vidéo, relatifs à des constructions, qui sont sémantiquement liées à des documents de notre jeu de tests. Les résultats des tests sont présentés dans la figure 3. De façon générale, l'annotation selon le standard TEI (Test#2) offre des résultats plus efficaces que ceux correspondants à DC (Test#1) ou

17. Nombre de critères pertinents couverts pour le test / Nombre total de critères annotés par le test.
18. Nombre de critères pertinents couverts pour le test / Nombre de critères pertinents attendus pour le test.
19. (2×P×R) /(P+R)

MPEG-7 (Test#3). Même lorsque les trois sont combinés, le score de F_1-Mesure, qui combine précision et rappel, ne change que légèrement de 0.59 (Test#2) à 0.61 (Test#4). Ceci en raison de la présence d'un grand nombre de méta-données textuelles annotées sur la base du standard TEI et d'un certain nombre de redondances avec les autres jeux d'annotations. En outre, nous ajoutons que la présence de relations structurelles entre composants textuels et documents a fortement contribué à la mise en valeur du Test#2 (TEI). LinkedMDR s'avère être le plus efficace avec le meilleur score de F_1-Mesure qui a atteint environ 0.94 (Test#5). Seuls les types de documents audio et vidéo ne sont pas représentés par LinkedMDR. Bien que ces types de document ne soient pas dans les objectifs d'annotation visés pour l'instant, ils peuvent l'être dans le futur (défi 4). Notre ontologie, qui est basée sur MPEG-7, est facilement extensible pour couvrir ultérieurement ces types de documents.

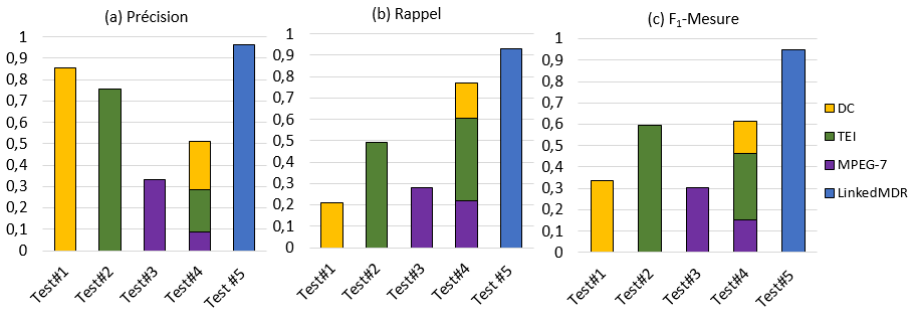

FIGURE *3. Évaluation de l'efficacité des modèles d'annotations.*

6. Conclusion

Cet article présente LinkedMDR, une nouvelle ontologie décrivant un corpus documentaire multimédia. LinkedMDR est adapté à tout corpus de documents hétérogènes spécifiques à un ou plusieurs domaines. Cette ontologie est basée sur l'intégration d'éléments des standards DC, TEI et MPEG-7 complétée par l'introduction de nouveaux concepts et relations dépassant leurs limites de représentation. LinkedMDR se scinde en plusieurs couches qui montrent explicitement, d'une part, sa généricité et, d'autre part, son potentiel de spécialisation, chacune mettant en exergue des possibilités d'extensions futures. Les expérimentations montrent de bons résultats de mesure de performance et d'efficacité d'annotations de méta-données et de contenus basées sur l'usage de l'ontologie LinkedMDR en comparaison aux annotations obtenues avec l'usage de standards existants.
Actuellement, nous développons une chaine de traitement automatique pour annoter un corpus de documents hétérogènes selon l'ontologie LinkedMDR et montrer que notre proposition peut être mise en place dans des scénarios réels. Pour ce faire, nous sommes en train d'exploiter les techniques avancées de collecte et d'extraction de métadonnées (Greenberg, 2004) ainsi que les techniques de traitement automatique du langage naturel comme dans (Maynard *et al.*, 2016). À court terme, nous envisageons également d'étendre nos expérimentations par des jeux de documents plus importants

et visons de nouveaux résultats confirmant une nouvelle fois la validation de notre modèle.

Bibliographie

Arndt R., Troncy R., Staab S., Hardman L., Vacura M. (2007). *COMM: designing a well-founded multimedia ontology for the web*. Springer.

Bates M. J. (2011). *Understanding information retrieval systems: management, types, and standards*. Auerbach Publications.

Bloechle J.-L., Rigamonti M., Hadjar K., Lalanne D., Ingold R. (2006). XCDF: a canonical and structured document format. In *International workshop on document analysis systems*, p. 141-152. Springer.

Brut M., Codreanu D., Manzat A.-M., Sèdes F. (2011). Distributed multimedia indexing and optimal resources utilization: An implementation based on metadata, context and usage. *JMPT*, vol. 2, n° 4, p. 197–225.

Brut M., Laborie S., Manzat A.-M., Sedes F. (2009). Integrating heterogeneous metadata into a distributed multimedia information system. *COGnitive systems with Interactive Sensors*.

Feng D., Siu W.-C., Zhang H. J. (2013). *Multimedia information retrieval and management: Technological fundamentals and applications*. Springer Science & Business Media.

Garcia R., Celma O. (2005). Semantic integration and retrieval of multimedia metadata. In *5th international workshop on knowledge markup and semantic annotation*, p. 69-80.

Greenberg J. (2004). Metadata extraction and harvesting: A comparison of two automatic metadata generation applications. *Journal of Internet Cataloging*, vol. 6, n° 4, p. 59–82.

Guo K., Liang Z., Tang Y., Chi T. (2017). Sor: An optimized semantic ontology retrieval algorithm for heterogeneous multimedia big data. *Journal of Computational Science*.

Huovila P. (2012). Linking IFCs and BIM to sustainability assessment of buildings. In *Proceedings of the cib w78 2012: 29th international conference.*

Klinger M., Susong M. (2006). The construction project: phases, people, terms, paperwork, processes. In, chap. Phases of the Contruction Project. American Bar Association.

Maynard D., Bontcheva K., Augenstein I. (2016). Natural language processing for the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 6, n° 2, p. 1–194.

Saathoff C., Scherp A. (2010). Unlocking the semantics of multimedia presentations in the web with the multimedia metadata ontology. In *Proceedings of the 19th international conference on world wide web*, p. 831-840. ACM.

Scherp A., Eissing D., Saathoff C. (2012). A method for integrating multimedia metadata standards and metadata formats with the multimedia metadata ontology. *International Journal of Semantic Computing*, vol. 6, n° 01, p. 25-49.

Suarez-Figueroa M. C., Atemezing G. A., Corcho O. (2013). The landscape of multimedia ontologies in the last decade. *Multimedia tools and applications*, vol. 62, n° 2, p. 377-399.