LinkedMDR: A Collective Knowledge Representation of a Heterogeneous Document Corpus

Nathalie Charbel, Christian Sallaberry, Sébastien Laborie, Gilbert Tekli, Richard Chbeir

LinkedMDR: A Collective Knowledge Representation of a Heterogeneous Document Corpus

Nathalie Charbel[1], Christian Sallaberry[2], Sebastien Laborie[1], Gilbert Tekli[3], and 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

Abstract. The ever increasing need for extracting knowledge from heterogeneous data has become a major concern. This is particularly observed in many application domains where several actors, with different expertise, exchange a great amount of information at any stage of a large-scale project. In this paper, we propose LinkedMDR: a novel ontology for Linked Multimedia Document Representation that describes the knowledge of a heterogeneous document corpus in a semantic data network. LinkedMDR combines existing standards and introduces new components that handle the connections between these standards and augment their capabilities. It is generic and offers a pluggable layer that makes it adaptable to different domain-specific knowledge. Experiments conducted on construction projects show that LinkedMDR is applicable in real-world scenarios.

Keywords: Heterogeneous Documents · Document Representation · Ontologies · Information Retrieval

1 Introduction

Modern large-scale projects adopt a systematic approach to project planning, where several actors of different expertise are involved. Throughout the project, they contribute and exchange a wide variety of technical and administrative knowledge depending on their background and role in the project. As an example, in the construction industry, actors (i.e., owners, consultants and contractors) exchange contracts, technical specifications, administrative forms, technical drawings and on-site photos throughout the different stages of a construction process [13]. The interchanged documents, originated from different sources, do not usually have a common standard structure. Also, they show heterogeneity in their formats (e.g., pdf, docx, xlsx, jpeg, etc.), contents (e.g., architecture, electrical, mechanical, structure, etc.), media types (e.g., image, text, etc.) and

versions. In addition, the documents may have implicit and explicit inter and intra-document links[4] and references.

In the literature, several works have been undertaken to define metadata on documents and their contents. These annotation models can be classified according to whether they deal with text-based content (e.g., TEI [22]), image-based content (e.g., EXIF [7]) or multimedia-based content [1–3, 8, 16]. Nevertheless, none of the existing works considers (i) the various inter and intra-document links, (ii) semantic annotations of both texts and images on a content and structural levels, (iii) documents multimodality, and (iv) particularities of domain-specific documents (such as technical drawings).

In this context, our aim is to provide a clear and a complete picture of the collection of heterogeneous documents exchanged in multidisciplinary and large projects. It is of a key importance to have the needed data easily available and adapted to the specific actor at any point in time based on his domain and requirements. To do so, there is a need to define a common and generic data model representing all the documents and their possible connections by means of semantic concepts and relations. Such linked data network would provide a collective knowledge that allows better indexing and semantic information processing and retrieval. This minimizes actors workload, cognitive efforts and human errors.

In this paper, we present LinkedMDR: a novel ontology for Linked Multimedia Document Representation. LinkedMDR combines existing standards that address metadata, structure and content representation (i.e., DC [6], TEI [22] and MPEG-7 [21]). In addition, it introduces new components augmenting the capabilities of these standards and allowing adaptation to different domain-specific knowledge. Our proposal is evaluated in the context of various construction related documents. Also, on-going evaluations in other application domains (such as the energy management domain) are still in progress.

The remainder of this paper is organized as follows. Sect. 2 illustrates the motivations through a real-world scenario encountered in the construction industry. Sect. 3 reviews related works in metadata, structure and content representation. Sect. 4 describes our LinkedMDR data model. Finally, Sect. 5 presents experimental results, while Sect. 6 concludes the paper and shapes some future works.

2 Motivations

To motivate our work, we investigate a representative scenario in the context of construction projects, provided by Nobatek[5], a French technological resource center involved in the sustainable construction domain.

[4] Inter-document links are relations between documents. Intra-document links are relations between elements of the same document.

[5] http://www.nobatek.com

2.1 Context

Nobatek is a consultancy office that assists investors in managing services related to sustainable construction projects. Its main role is to ensure the compliance of a project with the environmental standards and quality performance. Throughout the different construction stages, Nobatek communicates with other consultancy offices, contractors and specialized subcontractors. Across all of the involved actors, there exist members with different interests and profiles. Fig. 1 illustrates some of the various heterogeneous documents[6] that are involved in such projects. For instance, there are several documents that describe technical specifications (d_1 and d_5), thermal properties (d_2) and acoustic characteristics (d_3) of the building. Also, there is an excerpt of a technical drawing related to the ground floor (d_4) and an image (d_6). Currently, the engineers at Nobatek have to manually search for information in all these documents. In order to ensure the compliance of the building's exterior facades with the environmental standards, one needs to search for the related information contained in these documents produced by several specialized consultants (e.g., offices for acoustic and thermal analysis services). Given the large volume of information, the search becomes tedious and misleading as one can miss useful information.

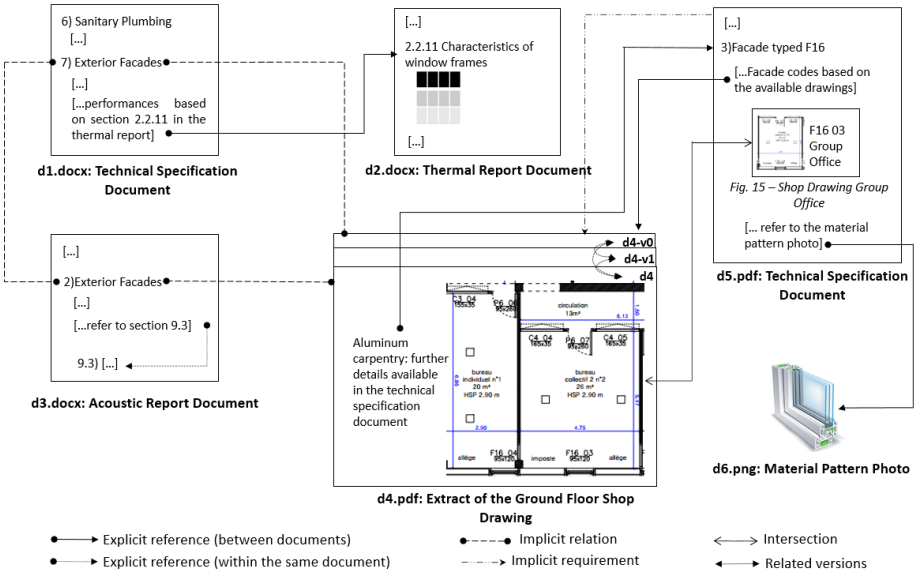

Fig. 1. Example of heterogeneous documents exchanged within a construction project.

[6] For the sake of simplicity, we only present 6 documents. However, other documents could be also involved such as videos, audios and 3D drawings.

2.2 Challenges

- **Challenge 1: Representing various inter and intra-document links** - One has to search for the relevant information in a collective knowledge built on a multitude of documents. Fig. 1 shows that a document d_i has various relations (references, mutual topics, versions, etc.) with other documents. These relations can be implicit or explicit. For instance, some textual sections of d_1 and d_3 have an implicit relation since they both describe building's exterior facades. The document d_5 has explicit reference to the image d_6 and the technical drawing d_4, which itself has several versions. This raises the need to index these documents and analyze their metadata information in order to build the relations between them as in Fig. 1.
- **Challenge 2: Handling information semantics on content and structural levels** - One needs to locate, on different depth levels, the relevant information contained in several documents recalling the same topic. For instance, d_5 describes general information regarding exterior facades, while d_1 contains detailed information of these facades together with other building services, such as plumbing and electricity. d_2 focuses particularly on the thermal properties of exterior and interior facades. This raises the need to reason over the document's content, its general metadata, and structural metadata that describes different depth levels (e.g., page, section, paragraph or sentence), thus to associate them with semantic concepts.
- **Challenge 3: Handling documents multimodality** - One has to work with heterogeneous documents: d_1, d_2 and d_3 are Word documents, d_4 is a CAD drawing in a PDF format, d_5 a PDF document and d_6 a PNG image. Therefore, it is convenient to handle different document types and formats.
- **Challenge 4: Ensuring extensibility** - One may encounter other types of construction related documents involving different structures, types of media, formats and document links. For example, an extension of the scenario illustrated in Fig. 1 may involve an audio file that analyzes the noise impact before and after the cladding of the facades. This raises the need to handle evolution of the information.

In the following section, we demonstrate that existing models are only capable of partially addressing the aforementioned challenges within a given information system.

3 Related Work

In this section, we describe existing standards and models addressing metadata and content representation (Sect. 3.1). We conclude with a discussion and a comparative study (Sect. 3.2).

3.1 Metadata and Content Representation Standards and Models

Dublin-Core - The Dublin Core Metadata Initiative (DC) is a Metadata standard for describing a wide range of online multimedia resources [6]. The Dublin

Core Metadata Element Set consists of 15 Elements describing the content of a document (e.g., title, description, etc.), the intellectual property (e.g., creator, rights, etc.), and its instantiation (e.g., date, format, etc.) [25]. This standard also offers a set of qualifiers which aim to modify the properties of the Dublin Core statements.

MPEG-7 - The Multimedia Content Description Interface is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group) [21]. Its aim is to provide a rich set of complex standardized tools describing low and high level features while also covering different granularities of data (e.g., collection, image, video, audio, segment) and different areas (content description, management, organization and navigation). The three main standardized components of MPEG-7 are: Descriptors (Ds), Description Schemes (DSs) and a Description Definition Language (DDL). While the MPEG-7 Ds are representations of features, the MPEG-7 DSs support complex descriptions and specify the structure and the semantics of the relationships among its constituents: Ds and DSs [17]. The MPEG-7 DDL is a standardized language based on XML schema. It allows the extension of existing Ds and DSs as well as the introduction of new components for specific domains [10].

Ontology-based Models - Many initiatives have been taken for the purpose of building multimedia ontologies, such as [1, 16, 24], or transforming existing formats into ontologies, such as [8]. The aim of these studies is to bridge the gap between low level features with automatically extractable information by machines and high level human interpretable features of the same information. [19]. There is often the need to combine several standards in order to meet the requirements of complex multimedia applications [18]. This boosts the efforts for building such ontologies. For instance, the Core Ontology for MultiMedia (COMM) [1] is built to satisfy multimedia annotation requirements. It is designed based on the MPEG-7 standard, DOLCE as a foundational ontology, and two ontology design patterns. In contrast to COMM, other MPEG-7 based ontologies are designed using a one-to-one automatic mapping of the entire standard to OWL (e.g., MPEG-7 Rhizomik [8]). Other ontologies, such as the Multimedia Metadata Ontology (M3O) [16], aim to provide abstract pattern-based models for multimedia metadata representation. M3O is centered on the formal upper-level ontology DOLCE+DnS Ultralight. One of the recent ontologies for describing multimedia content is the Media Resource Ontology developed by the W3C Media Annotation Working Group [24]. It is subject to multiple alignments with several existing multimedia metadata, such as MPEG-7, Dublin Core and EXIF.

LINDO Metadata Model - In the context of distributed multimedia information systems, such as the video surveillance applications that use different indexing engines, interoperability problems arise when metadata of different formats are combined together. The LINDO project (Large scale distributed INDexation of multimedia Objects) [12] took up the challenge of handling different metadata

standards within its distributed information system, such as Dublin Core, EXIF and MPEG-7. Thus, the authors in [3] define a unified XML-based metadata model that encapsulates these standards based on two levels: general metadata information describing the entire document and metadata related to multimedia contents (image, text, video and audio).

XCDF Format - Problems such as over-segmentation, additional noisy information, and lack of structure preservation produced by PDF generators are often associated with the PDF format. Great effort has been put into overcoming these issues [2]: XCDF is a canonical format which purpose is to represent the results of the physical structure extraction of the PDF documents in a single and structured format. It is based on the XML format and its DTD provides basic elements for general document representation (page, fonts, image, graphic, textblock, textline and token).

EXIF - EXIF (Exchangeable Image File Format) standard is a widely used standard for describing digital images [7]. It mainly supports a set of tags related to image-specific features (e.g., width, height, pixel composition, color), and many other general tags (e.g., image title, date and time, creator, version).

TEI - In this paper, we focus on the TEI (Text Encoding Initiative) [22] as it is a commonly adopted text-driven descriptive standard. It is based on XML format and provides a way to describe information and meta-information within a textual document. TEI offers a representational form of a text (e.g., chapters, sections, paragraphs, lists, items, page break, table, figures and graphics) as well as a set of semantically rich elements (e.g., cross reference links, title, name and address) that could improve information retrieval.

3.2 Discussion

We evaluated the existing standards and models that describe metadata and content related to a heterogeneous document corpus based on the challenges previously mentioned in Sect. 2.2. The results are depicted in Table 1.

It is obvious that the standards that focus only on text or image description are limited because they cannot handle the different types of multimedia documents (Challenge 3). Nevertheless, these standards, in particular the TEI for the text, provide relevant components that can be reused in our proposal. For example, the *<ref>* element allows us to represent document links but is limited to cross-references. As for the multimedia-based standards, the MPEG-7 defines *<StillRegion>* and *<TextAnnotation>* elements, which could be used to describe different parts of a technical drawing and its associated text legends. However, these two elements do not allow the description of relevant details that are sometimes encountered within text legends (Challenges 2 and 4), nor possible relations with descriptors from other standards, such as TEI or DC (Challenge 1).

Table 1. Evaluation of the existing standards and models with regard to the identified challenges.

Challenges	Approaches / Properties		Metadata and Content Representation Standards and Models									
			Multimedia-based								Image-based	Text-based
			Dublin Core	MPEG-7	COMM	M3O	MediaOnt	Mpeg-7 Rhizomik	LINDO Metadata Model	XCDF Format	EXIF	TEI
Challenge 1	Relations Description	Intra-document Link	Partial	Partial	Partial	Partial	Partial	Partial	Partial	x	x	Partial
		Inter-document Link	Partial	Partial	Partial	Partial	Partial	Partial	Partial	x	x	Partial
Challenge 2	Descriptive Metadata Representation		√	√	√	√	√	√	√	x	√	√
	Content Description		x	√	√	√	x	√	√	x	x	√
	Structural Metadata Representation	Image	x	√	√	√	x	√	√	x	x	x
		Text	x	Partial	Partial	Partial	x	Partial	Partial	Partial	x	√
Challenge 3	Multimodality		√	x	x	x	x	x	√	x	x	x
Challenge 4	Extensibility		Partial	Partial	Partial	Partial	Partial	Partial	Partial	Partial	x	Partial

The multimedia-based ontologies and models are also limited. Most of them aggregate descriptors of existing standards but do not offer a solution to the challenges mentioned above (specifically, Challenges 1 and 2). For instance, they are not capable of relating descriptors from MPEG-7, DC and TEI. Hence, these models inherit the deficiencies of their constituent standards.

To our knowledge, there is currently no available representation of a semantic network of linked data that can describe the collective knowledge of a heterogeneous document corpus. Our proposal, illustrated in the following section, answers this need through the design of a novel ontology relying on the most adopted metadata standards. We consider the ontologies as a reliable and an efficient solution to support Semantic Information Retrieval from heterogeneous multimedia data [9].

4 LinkedMDR: a novel ontology for Linked Multimedia Document Representation

We propose a novel ontology for Linked Multimedia Document Representation, entitled LinkedMDR[7]. Our ontology models the knowledge embedded in a corpus of heterogeneous documents. This approach is based on (i) the combination of the relevant standards for metadata representation and content description (e.g., DC [6], TEI [22] and MPEG-7 [21]), and (ii) the introduction of new semantic concepts and relations that augment the capabilities of these standards and link them through a semantic network. Our proposed ontology is made of three main layers: (i) the core layer serving as a mediator, (ii) the standardized

[7] LinkedMDR is an OWL ontology created on Protégé. Details on the LinkedMDR ontology, the overall concepts and relations are available at http://spider.sigappfr.org/linkedmdr/

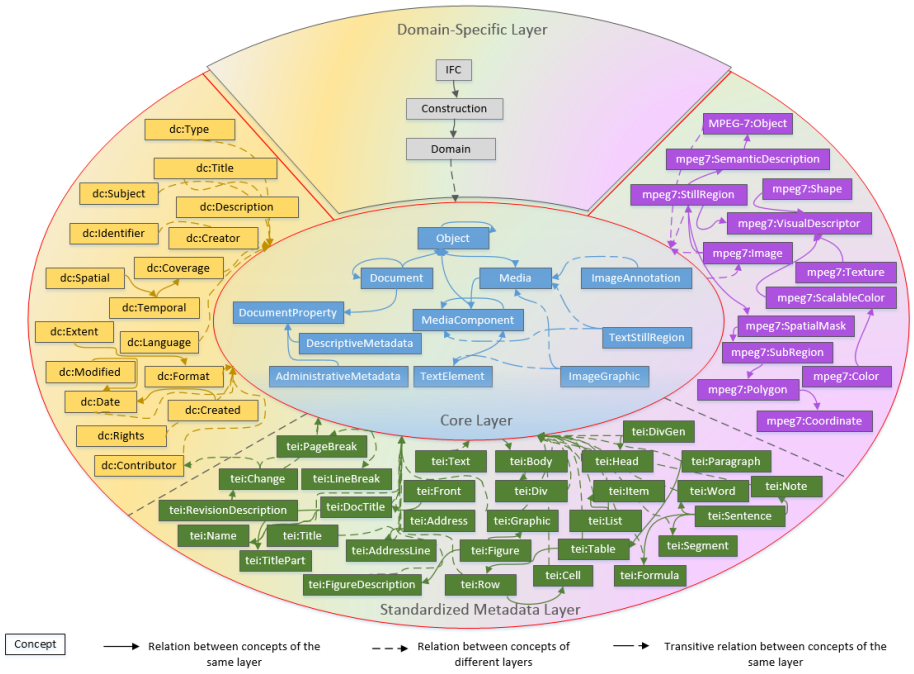

Fig. 2. LinkedMDR: the semantic model representing a corpus of multimedia documents.

metadata layer comprising elements from existing standards, and (iii) the pluggable domain-specific layer that is adaptable to domain-specific applications. The multiple layers of LinkedMDR ensure its genericity and extensibility (cf. Fig. 2).

4.1 The Core Layer

This layer introduces new concepts and relations that enrich the existing standards, mainly: (i) concepts that model the global composition of a given document and the metadata properties associated to it (e.g., *Document*, *Media*, *MediaComponent* and *DocumentProperty*), (ii) *Object* entity which abstracts *Document*, *Media* and *MediaComponent* while introducing a rich set of relations between them (i.e., *hasPart*, semantic, temporal and spatial relations), (iii) concepts that subsume other concepts involved in an adjacent layer (e.g., *DescriptiveMetadata*, *AdministrativeMetadata* and *TextElement* are super concepts generalizing metadata information of DC and TEI respectively), and (iv) concepts that extend some elements of MPEG-7 and TEI (e.g., *TextStillRegion* extends *tei:Text* and *mpeg7:StillRegion* to describe text metadata inside an image, *ImageGraphic* extends *mpeg7:Image* and *tei:Graphic* to represent the image metadata inside a text, and *ImageAnnotation* extends *tei:Text* and *mpeg7:Image* to annotate an image with another image or a structured text).

In the remainder of the paper, we use RDF [23] statements to illustrate instances of our ontology (with *lmdr* as prefix) as RDF is the most widely used data model for representing a semantic and extensible graph.

For instance, Fig. 1 shows that section 3 (*div3)* of d_5 contains a figure which itself is an excerpt of the technical drawing d_4. This can be represented by the following triples: <*lmdr:d5, lmdr:hasPart, tei:d5.div3.figure15*> and <*lmdr:d4, lmdr:contains, tei:d5.div3.figure15*>. It addresses Challenges 1, 2 and 3 (Sect. 2.2). It represents a spatial relation which is an inter-document link between documents of different types. In addition, it illustrates the semantics associated to concepts of various granularities, which allows to infer further relations and enrich the network of linked data (e.g., if <*tei:d5.div3.figure15, lmdr:contains, tei:dn.div1.figure1*>, then <*lmdr:d4, lmdr:contains, tei:dn.div1.figure1*> is deduced by transitivity). This cannot be done with existing standards since they do not handle similar relations on different levels of precision.

4.2 The Standardized Metadata Layer

This layer is made of selected metadata information defined by existing standards: Dublin Core [6], TEI [22] and MPEG-7 [21]. Consequently, it is divided into three sub-layers, each dedicated to a standard. The first corresponds to DC and comprises metadata information of the document in general. The second is related to TEI's structural text metadata. The third contains metadata information that describes the image with different levels of precision, visual features and semantic descriptors following the MPEG-7 standard. It is to note that, other standards could be easily integrated into our ontology, in the future, specifically into this layer.

This layer also involves relations between its sub-layers. For instance, we added the *isRevisedBy* relation in order to link the *tei:Change* concept (a set of changes made during the revision of a document) to the corresponding *dc:Contributor* concept (a person or organization responsible for making these changes). Further, each sub-layer is connected to the core layer (Sect. 4.1) through relations between their respective concepts. For example, the subsumption relation in <*tei:Text, lmdr:isA, lmdr:Media*>, <*mpeg7:StillRegion, lmdr:isA, lmdr:MediaComponent*>, <*dc:Title, lmdr:isA, lmdr:DescriptiveMetadata*> and <*lmdr:TextElement, lmdr:isOn, tei:PageBreak*> allow concepts of TEI, MPEG-7 and DC to inherit common properties from the core layer.

For instance, we can represent the documents of Fig. 1, such as d4, using triples related to the DC metadata: e.g., <*lmdr:d4, lmdr:hasProperty, dc:d4.title*> and <*dc:d4.title, rdf:value, "Shop Drawing"*>. Also, d_4 contains several technical drawings, each related to a specific building floor and described on different pages of the document. The MPEG-7 metadata helps in describing the different regions of the drawings but without any information on the corresponding pages. Similarly, the TEI metadata represents the pages of each drawing but without any description of their content. Hence, considering Challenge 1 (Sect. 2.2), the links between the metadata of these different standards is only possible via the LinkedMDR concepts and relations of the core layer: <*lmdr:d4, lmdr:hasPart,*

lmdr:d4.imagegraphic>, <*lmdr:d4.imagegraphic, lmdr:isOn, tei:d4.page1*>, <*lmdr:d4.imagegraphic, lmdr:hasPart, mpeg7:d4.stillregion*>.

4.3 The Pluggable Domain-Specific Layer

The previously mentioned layers are generic and independent of the type of the linked multimedia documents. However, we aim to provide a generic ontology for any domain-specific application. We introduce this pluggable layer as a means to make LinkedMDR adaptable to any domain-specific knowledge. To do so, we present a new concept entitled *Domain* and we link it to the *Object* concept of the core layer. This way domain-specific concepts can be added under the *Domain* while relating to sub-concepts of *Object* (i.e., *Document*, *Media* and *MediaComponent*).

In this paper, we introduce an example showing how we can make this layer adaptable to the construction domain. We add *Construction* concept as a sub-concept of *Domain*. We also relate the latter to *IFC* concept which comprises concepts from ifcOWL[8], which is the conversion of the IFC (Industry Foundation Classes) [4] schema into ontology. The IFC standard is the complete and fully stable open and international standard for exchanging BIM (Building Information Modeling) data, independently of the different software applications. It involves building and construction objects (including physical components, spaces, systems, processes and actors) and relationships between them [11].

As an example, section 7 (*div7*) in d_1 (Fig. 1) describes the exterior facades. It is now possible to link section 7 with the related IFC object (e.g., window4): <*lmdr:d1, lmdr:isA, lmdr:Document*>, <*lmdr:d1, lmdr:hasPart, tei:d1.div7*>, <*tei:d1.div7, tei:isA, tei:Div*>, <*tei:d1.div7, lmdr:isRelated, ifc:window4*>, and <*ifc:window4, ifc:isA, ifc:BuildingElement*>. This particularly answers Challenges 1 and 4.

5 Experimental Evaluation

We have conducted several experiments to evaluate the quality of the annotation of heterogeneous corpora based on LinkedMDR, in different scenarios, w.r.t. the previously discussed objectives (See Sect. 2.2). We show the results of 2 experiments applied on real-world construction projects provided by Nobatek. Note that, an on-going evaluation of LinkedMDR in the context of HIT2GAP[9], a European H2020 funded project, is still in progress.

[8] Available at http://ifcowl.openbimstandards.org/IFC4_ADD2.owl

[9] HIT2GAP (Highly Innovative Building Control Tools) is a large-scale project that involves 21 partners and provides an energy management platform for managing building energy behavior. Further details are available at: http://www.hit2gap.eu/

5.1 Context

Test Data - We hand-picked 5 heterogeneous corpora of real construction projects from Nobatek. Table 2 shows the document composition of each test corpus.

Table 2. Document composition of the test corpora.

Corpus	No. of Documents	Document formats	Corpus Size (MB)
1	10	3 docx, 4 pdf, 3 png	20
2	10	7 pdf, 2 png, 1 jpeg	27.4
3	17	5 docx, 10 pdf, 2 png	54.8
4	15	5 xslx, 1 docx, 7 pdf, 2 png	112.3
5	12	1 xslx, 10 pdf, 1 png	38.2

Prototype - We have developed LMDR Annotator, a java-based prototype, which purpose is to automatically annotate a document corpus based on (i) DC (using Apache Tika API), (ii) TEI (using Oxgarage Web service), (iii) MPEG-7 (using MPEG-7 Visual Descriptors API), and (iv) LinkedMDR. As for the latter, it uses the GATE (General Architecture for Text Engineering) API for the automatic generation of explicit inter and intra-document link annotations based on regular expressions encountered in text. The output XML files generated by the GATE module and the other annotation modules (related to DC, TEI and MPEG-7) are automatically transformed into LinkedMDR instances using tailored XSLT processors. The final output is an RDF file describing the entire corpus. Further details on the prototype's architecture and its different modules are available online[10].

Experiment 1 - The aim of this experiment is to compare LinkedMDR with its alternatives (i.e., DC [6], TEI [22], MPEG-7 [21], and the three combined) regardless of the annotation tools. We manually annotated one corpus (Corpus 1) using each standard separately, the three standards combined, and LinkedMDR. The manual annotations ensure the best possible document representation that can be generated from each of the data models.

We then evaluate the conciseness of each data model. More particularly, for each data model we look at the total number of annotated documents, the number of annotation elements[11], redundancies (overlapping metadata), and the number of annotation files required for covering the maximum number of relevant criteria. For the latter, we calculate F_2-scores[12] which weight the Recall measure higher than the Precision, to emphasize missed relevant annotations. Note that,

[10] http://spider.sigappfr.org/linkedmdr/lmdr-annotator

[11] The number of XML tags in the XML annotation files that we generated based on the existing standards and the number of RDF triples that we generated in the LinkedMDR ontology.

[12] F_2-measure: $(5\times P\times R)$ $/(4\times P+R)$
Recall: No. of covered relevant criteria / Total No. of expected criteria.
Precision: No. of covered relevant criteria / Total No. of annotated criteria.

we define the relevant criteria (e.g. semantic intra-document links, topological intra-document links, general metadata, text/image-specific metadata, etc.) as to address the requirements that we aim to satisfy (See Sect. 2.2). The considered factors are particularly important for future Information Retrieval application.

Experiment 2 - In a real-world application, manually annotating a document corpus is a tedious job that requires much effort and technical knowledge. In this experiment, we evaluate the effectiveness of our LMDR Annotator in automatically annotating different corpora (Corpus 2 to 5). We calculate Precision, Recall and F_2-measure. One RDF annotation file ($Corpus_m.rdf$) is generated per corpus, such that

$$Corpus_m.rdf = Corpus_m[DC].rdf \cup Corpus_m[TEI].rdf \cup Corpus_m[MPEG7].rdf \cup Corpus_m[xLinks].rdf$$

where $Corpus_m[DC].rdf$, $Corpus_m[TEI].rdf$ and $Corpus_m[MPEG7].rdf$ are the set of generated LinkedMDR instances related to the standardized metadata layer (See Sect. 4.2) for the $Corpus_m$; $Corpus_m[xLinks].rdf$ is the set of generated LinkedMDR instances, related to the Core layer (See. Sect. 4.1), representing explicit inter and intra-document links involved within $Corpus_m$; $Corpus_m$ is the set of heterogeneous documents of the m^{th} project.

5.2 Experimental Results

Experiment 1: Evaluating the Conciseness of LinkedMDR - Table 3 shows that using DC, the three standards combined, and LinkedMDR, we were able to annotate the 10 source documents involved in Corpus 1. In contrast, using TEI and MPEG-7, the annotations were not exhaustive. This is due to the incapacity of annotating images and technical drawings in TEI and textual documents in MPEG-7. The Annotation files that we generate using the DC standard contain a very small number of annotation elements since DC covers generic metadata and neglects structure and content representation of the documents (weakest f_2-score: 0.25). The combination of the three standards produces a significant number of annotation elements since the TEI and MPEG-7 standards are very verbose while not covering all of the expected annotation elements. Also, it involves many redundancies caused by common metadata elements between the DC, TEI and MPEG-7 standards (f_2-score: 0.70). Hence, the current experiment shows that the annotation based on LinkedMDR is the most concise since it provides the highest f_2-score (0.94) with a relatively small number of annotation elements, all in a single annotation file, and without any redundancy.

Experiment 2: Evaluating the Effectiveness of LMDR Annotator - Fig. 3 shows the f_2-scores evaluating the outputs of each annotation module separately (i.e., $Corpus_m[DC].rdf$, $Corpus_m[TEI].rdf$, $Corpus_m[MPEG7].rdf$,

Table 3. Results of Experiment 1: the evaluation of the conciseness of the existing standards and LinkedMDR in annotating Corpus 1.

Annotation Model	No. of Annotated Documents	Cumulative No. of Annotation Elements	No. of Annotation Files	No. of Overlapping Metadata	F_2-Scores
DC [6]	10	131	10	0	0.25
TEI [22]	5	807	5	0	0.53
MPEG-7 [21]	5	495	5	0	0.29
Combined Standards	10	1433	20	128	0.70
LinkedMDR	10	604	1	0	0.94

and $Corpus_m[xLinks].rdf$) then their union ($Corpus_m.rdf$). The results of the automatic annotation of the 4 corpora have f_2-scores that range from 0.48 ($Corpus_4.rdf$) to 0.63 ($Corpus_2.rdf$).

Looking over the individual annotation modules, the f_2-scores slightly change from one corpus to another. $Corpus_m[DC].rdf$ is in general the most effective since it involves LinkedMDR instances generated from documents' meta-tags (e.g., title, creator, date, format, etc.) which are easily extracted automatically using the Apache Tika API. On the other hand, $Corpus_m[MPEG7].rdf$ produces the lowest scores since relevant concepts, such as *mpeg7:StillRegion*, are relatively difficult to generate automatically. In fact, the MPEG-7 Visual Descriptors library is limited to the automatic extraction of low level features, such as color and texture characteristics. Advanced feature extraction requires sophisticated computer vision and machine learning algorithms (such as [14]) which, so far, are not adopted in our prototype. As for $Corpus_m[xLinks].rdf$, one can see that the GATE java API is reliable in automatically extracting explicit inter and intra-document references from regular expressions encountered in the textual documents. However, in some cases, such as in Corpus 4, we obtain a relatively lower score. This is due to some expressions presenting ambiguous references that could not be handled automatically without the use of advanced semantic disambiguation techniques [5, 20].

Fig. 4 illustrates the recall values w.r.t. to the total expected LinkedMDR instances per corpus. Intuitively, the recall scores decrease when the number of documents increases since more complex inter and intra-document links are involved, which cannot yet be resolved. This emphasizes that our LMDR Annotator, at its current state, offers relatively low recall scores when compared to the annotation potential proposed by the LinkedMDR ontology. Our current effort focuses on furthering the annotation capabilities of LMDR Annotator as experiments have shown promising results for its adoption in real world projects.

6 Conclusion and Future Work

This paper introduces LinkedMDR: Linked Multimedia Document Representation, a novel ontology for representing the collective knowledge embedded in a heterogeneous document corpus. LinkedMDR is based on the combination of the DC [6], TEI[22], and MPEG-7[21] standards with the introduction of new con-

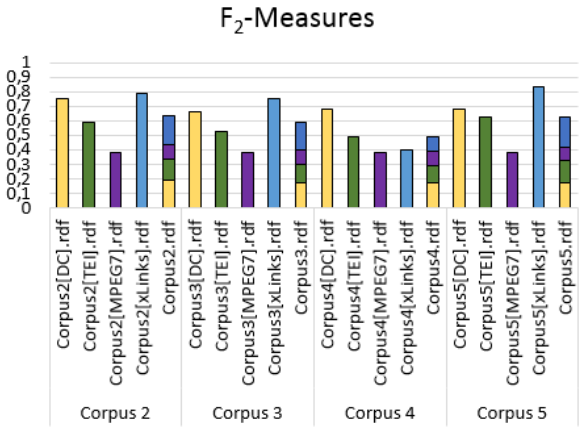

Fig. 3. Results of Experiment 2: F_2-scores measuring the effectiveness of our LMDR Annotator for Corpus 2 to 5.

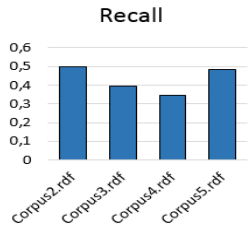

Fig. 4. Recall scores based on the total expected LinkedMDR instances per corpus.

cepts and relations augmenting the capabilities of these standards and allowing adaptation to different domain-specific knowledge. Experiments were conducted on real-world construction projects from Nobatek using our LMDR Annotator, a java-based prototype for automatic annotation of a heterogeneous document corpus based on LinkedMDR.

Building on our findings, we are working on extending our research towards the energy management domain, in the context of the European H2020 funded project HIT2GAP. Furthermore, we are extending the LMDR Annotator to handle more inter and intra-document links (such as semantic, topological, spatial) and improve the effectiveness of the automatic annotation based on LinkedMDR. We also aim at providing an interactive interface for non expert users in order to reason over the LinkedMDR ontology using non technical queries such in [15].

References

1. Arndt, R., Troncy, R., Staab, S., Hardman, L., Vacura, M.: COMM: designing a well-founded multimedia ontology for the web. Springer (2007)
2. Bloechle, J.L., Rigamonti, M., Hadjar, K., Lalanne, D., Ingold, R.: XCDF: a canonical and structured document format. In: International Workshop on Document Analysis Systems. pp. 141–152. Springer (2006)
3. Brut, M., Laborie, S., Manzat, A.M., Sedes, F.: Integrating heterogeneous metadata into a distributed multimedia information system. COGnitive systems with Interactive Sensors (2009)
4. buildingSMART: IFC-Industry Foundation Classes, IFC4 Add2 Release (2016), http://www.buildingsmart-tech.org/specifications/ifc-releases/ifc4-add2-release
5. Charbel, N., Tekli, J., Chbeir, R., Tekli, G.: Resolving XML semantic ambiguity. In: EDBT. pp. 277–288 (2015)

6. Dublin Core Metadata Initiative: DCMI Metadata Terms (2012), http://dublincore.org/documents/dcmi-terms/
7. EXIF: Exchangeable Image File Format for digital still cameras (2002), http://www.exif.org/Exif2-2.PDF
8. Garcia, R., Celma, O.: Semantic integration and retrieval of multimedia metadata. In: 5th International Workshop on Knowledge Markup and Semantic Annotation. pp. 69–80 (2005)
9. Guo, K., Liang, Z., Tang, Y., Chi, T.: Sor: An optimized semantic ontology retrieval algorithm for heterogeneous multimedia big data. Journal of Computational Science (2017)
10. Hunter, J.: An overview of the MPEG-7 Description Definition Language (DDL). IEEE Transactions on circuits and systems for video technology 11(6), 765–772 (2001)
11. Huovila, P.: Linking IFCs and BIM to sustainability assessment of buildings. In: Proceedings of the CIB W78 2012: 29th International Conference (2012)
12. ITEA: LINDO-Large scale distributed INDexation of multimedia Objects (2010), https://itea3.org/project/lindo.html
13. Klinger, M., Susong, M.: The construction project: phases, people, terms, paperwork, processes, chap. Phases of the Contruction Project. American Bar Association (2006)
14. OpenCV: Open Source Computer Vision Library (2011), http://opencv.org
15. Pankowski, T., Brzykcy, G.: Data access based on faceted queries over ontologies. In: International Conference on Database and Expert Systems Applications. pp. 275–286. Springer (2016)
16. Saathoff, C., Scherp, A.: 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. pp. 831–840. ACM (2010)
17. Salembier, P., Smith, J.R.: MPEG-7 Multimedia Description Schemes. IEEE Transactions on circuits and systems for video technology 11(6), 748–759 (2001)
18. Scherp, A., Eissing, D., Saathoff, C.: A method for integrating multimedia metadata standards and metadata formats with the multimedia metadata ontology. International Journal of Semantic Computing 6(01), 25–49 (2012)
19. Suarez-Figueroa, M.C., Atemezing, G.A., Corcho, O.: The landscape of multimedia ontologies in the last decade. Multimedia tools and applications 62(2), 377–399 (2013)
20. Tekli, J., Charbel, N., Chbeir, R.: Building semantic trees from XML documents. Web Semantics: Science, Services and Agents on the World Wide Web 37, 1–24 (2016)
21. The Moving Picture Experts Group: MPEG7-Multimedia Content Description Interface (2001), http://mpeg.chiariglione.org/standards/mpeg-7
22. The Text Encoding Initiative Consortium: TEI-Text Encoding Initiative (1994), http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf
23. W3C: Resource Description Framework (2004), https://www.w3.org/RDF/
24. W3C: Ontology for Media Resources 1.0 (2012), http://www.w3.org/TR/mediaont-10/
25. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin Core metadata for resource discovery. Tech. Rep. 2070-1721 (1998)