MuMIe: Multi-level Metadata Mapping System

Samir Amir, Ioan Marius Bilasco, Chabane Djeraba
LIFL UMR CNRS 8022, University of Lille1 and Telecom-Lille1, Villeneuve d'Ascq, France
Email:{samir.amir, marius.bilasco, chabane.djeraba}@lifl.fr

Abstract—**The recent growth of multimedia requires an extensive use of metadata for their management. However, a uniform access to metadata is necessary in order to take advantage of them. In this context, several techniques for achieving metadata interoperability have been developed. Most of these techniques focus on matching schemas defined by using one schema description language. The few existing matching systems that support schemas from different languages present some limitations, especially when using semantic and structural information. In this paper we present a new integration system supporting schemas from different description languages. Moreover, the proposed matching process makes use of several types of information (linguistic, semantic and structural) in a manner that increases the matching accuracy.**

I. Introduction

The ubiquitous presence of multimedia in our lives has generated an increasing interest in the use of metadata to improve the efficiency and effectiveness of retrieval, filtering and managing procedures of these types of contents. Metadata is a machine processable data that is used for the interpretation and the processing of multimedia content for their adaptation, filtering, or semantic knowledge extraction. Metadata describes different types of information: multimedia contents (e.g., video, image, audio, etc.), semantics of these contents (e.g., concept in image, context of video, etc.), characteristics of devices consuming or transmitting these contents (e.g., networks, TV, mobile, etc.) and consumer characteristics (e.g., consumer profile, consumer preference, etc.).

Metadata has extensive use in real world applications. Multimedia services platform often combine metadata issue from several domains. For instance, The BBC[1] offers an online program information service for its TV and radio channels, where The TV-Anytime and MPEG-7 standards [15] were chosen for representing program information and audio-visual metadata respectively. HARMONY[2] (part of International Digital Libraries Initiative) is another example of a framework that combines metadata from several domain. The framework addresses the challenge of describing networked collections of highly complex and mixed-media digital objects. HARMONY uses RDF, XML, Dublin Core, MPEG-7 and INDECS standards for achieving its goals.

By anticipating the increase of metadata in the upcoming years, we can foresee that it will become more and more difficult to achieve a uniform access to media objects since a multimedia object can be described using different metadata standards, developed by independent communities. So, a given user or service cannot extract a given attribute without a ability of interpretation. In this context, *metadata interoperability* becomes crucial.

Over the last decades, a various of interoperability techniques have been proposed and are analyzed in [14]. Most of these works focus on the creation of a core ontology which contains common information on metadata to be integrated. This ontology acts as a mediated schema, which is the common interface used for querying all metadata encoded in different formats. After designing this core ontology, a manual mapping is performed between the latter and other metadata formats [1].

The works mentioned above have not been a real success since the integration process is performed manually, which is costly and time consuming. Besides, the integration must be updated every time a new metadata format appears. In this context, several semi-automatic techniques have been developed to facilitate the integration process. *Schema matching* techniques play a central role in these approaches [27].

Due to the complexity of schema matching, it was mostly performed manually by human experts. However, manual reconciliation tends to be a slow and inefficient process especially in large-scale schemas (e.g., MPEG-7, MPEG-21, etc.) and dynamic environments such as multimedia where new metadata standards are appearing constantly. Another problem that should be taken into consideration is that metadata are heterogeneous on two levels: *schema* and *schema definition language*. The structural and semantic heterogeneity on the schema level occurs when the same information is represented differently on different schemas (e.g., naming conflicts, multilateral correspondence, abstraction level incompatibility, etc.). The schema definition language heterogeneity is due to the substantial structural and semantic discrepancies of schema definition languages. For instance, several XML Schema structural descriptions can not be expressed using RDF Schema (RDFS) and several semantic descriptions of the latter are not supported by XML Schema (XSD) [14].

This paper aims at automating the integration process of metadata in order to achieve interoperability. This is done by proposing a new integration system named MuMIe (Multi-level Metadata Integration). The proposed system can be used to help developers in their integration process (semi-automatic integration). However, if we consider the definitions given in the [14], where the authors define

[1]http://backstage.bbc.co.uk/feeds/tvradio/doc.html
[2]http://www.ilrt.bris.ac.uk/discovery/harmony/

interoperability as *the ability to exchange metadata between two or more systems without or with minimal loss of information and without any special effort on either system*, we can say that the proposed solution can be considered as automatic solution in the case of a high matching accuracy.

MuMIe works on the schema level and schema definition language level. Moreover, our system combines several types of semantic and structural information so as to increases the matching accuracy. The paper is organized as follows: Section II discusses the limitations of existing schema matching techniques and describes related work. In Section III, we describe the proposed integration system. An evaluation study is presented in Section IV. Finally, Section V gives concluding remarks and future work.

II. Related Work

Since metadata heterogeneity can occur on the schema definition language and the schema levels, it is necessary to specify the mappings for each level. In order to perform the mapping for the schema definition language level, several methods have been developed. Examples include the work done in [12] [32], in which the authors propose conversion rules for XML Schemas to OWL. These rules insure the capture of a part of XML Schema implicit semantics. The authors in [11] describe the conversion of schemas into an object-oriented representation and propose a methodology for translating schemas and converting relational instance data. The Ontology definition metamodel specification [17] offers a set of metamodels and mappings for translating between the UML metamodel and ontology languages such as RDF Schema and OWL. the authors in [23] present an approach for automatically generating RDF Schema from XML Schema specification. Other approaches for language translation have been proposed, some of them are analyzed in [14]. Because of the substantial structural and semantic discrepancies among schema definition languages, the works mentioned previously have not been very successful as the translation from one language into another causes the loss of valuable semantic and structural information.

Concerning the heterogeneity problem on the schema level, tools and mechanisms are needed in order to achieve interoperability. These tools and mechanisms must resolve the semantic and structural heterogeneity of schemas and align terms between metadata. To do so, several schema matching techniques have been proposed over the last decade, but up to now few existing systems aim at being general and support schemas from different languages, among them we can note: Similarity Flooding [22](relational, XML and RDF), Cupid [20](relational and XML), S-Match [13] (relational, XML and Ontologies) and Clio [24] [26](relational and XML). These approaches do not make use of most of the structural and semantic information (e.g., equivalence properties, generalization features, disjointness classes, etc.). Consequently, it is difficult to detect the complex matching such as one-to-many and many-to-many. Moreover, the main drawback of these methods is how to use the structural information. For instance, the authors in [20] consider that much of the information content is represented in leaves and that the leaves have less variations between schemas than the internal structures. Thus, the similarity of inter-nodes is based on the similarity of their leaf sets. This is not always true as we can find equivalent concepts occurring in completely different structures, and completely independent concepts that belong to isomorphic structures. The drawback of the method developed in [22] is that is based on the idea of similarity propagation, the basic concept behind the algorithm is that adjacency contributes to similarity propagation. Thus, the algorithm will perform unexpectedly in cases when adjacency information is not preserved. The approach proposed in [24] only makes use of parent-child relationship to calculate structural similarity contexts. The work done in [13] is limited to a tree-like structure and does not consider properties or roles.

Motivated by the above challenges, we discuss below a simplified approach for achieving the interoperability on two levels. The approach can be used to integrate several heterogeneous schemas, defined using different definition languages. We combine several syntactic, semantic and structural resources available on schemas to detect the mappings between metadata terms. Moreover, our approach overcomes the structural gap of the matching systems mentioned above by using several similarity contexts [16].

III. Proposed Approach

This section shows the two main parts of the proposed system. The first one is the *projection* step which is responsible for achieving interoperability on the language level. This is done by the transformation of all schemas into a common representation model. The second part is the *matching* step which is responsible for finding the correspondences between metadata terms by using several types of information.

A. Projection

As mentioned in Section II, up to now, few existing matching systems support schemas from different languages. In this section, we propose transformation rules that allow the projection of different schema languages on the same representation space. This space is an abstract model that serves as a foundation to represent conceptually several types of schemas whatever their description language. We model schemas as a directed labeled graph (Section III-A.1) and save semantic and structural information in order to use them for the matching process (Section III-A.2). Figure 1 illustrates the projection process. Note that examples on this paper are limited to three description languages (XML Schema, RDF Schema and OWL) [9] [6] [21], since they are commonly used by metadata communities and present a significant heterogeneity (a high structural description for XSD and a high semantic expressiveness for RDF Schema and OWL).

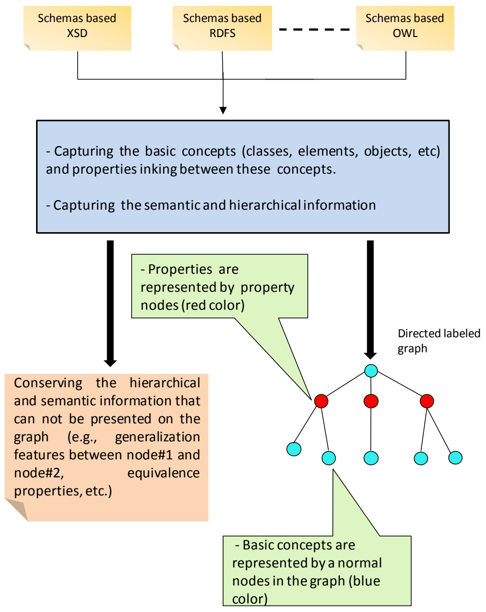
Schemas based XSD
Schemas based RDFS
Schemas based OWL
- Capturing the basic concepts (classes, elements, objects, etc) and properties inking between these concepts.
- Capturing the semantic and hierarchical information
- Properties are represented by property nodes (red color)
Directed labeled graph
Conserving the hierarchical and semantic information that can not be presented on the graph (e.g., generalization features between node#1 and node#2, equivalence properties, etc.)
- Basic concepts are represented by a normal nodes in the graph (blue color)

Figure 1. Illustration of the projection step.

1) Schema Graph Representation: Due to the high heterogeneity of schema definition languages, it is not possible to find one representation model which supports all schema features. In our approach, we model these schemas as a directed labeled graphs representing only basic schema concepts and properties linking between them. These two entities are the common and basic information for all description languages. In order to create the graph, the description concepts of schema definition languages must be known. For instance, all classes (*rdfs*:*class*) and properties (*rdf*:*property*) in RDF Schema must be represented by nodes. Table I shows the basic concepts and properties for the three different languages. Concerning the languages based on taxonomy classification (e.g., XML Schema), we use the idea proposed in [23] to capture the semantics of implicit properties. So, the name of property node is the combination of parent and child nodes.

We categorize nodes into *normal nodes* and *property nodes*. Normal nodes correspond to the basic concepts (e.g., *xsd*:*complexType*, *rdfs*:*class*, *xsd*:*element*, etc.) and property nodes correspond to properties linking between schema concepts (e.g., *rdf*:*property*, *owl*:*ObjectProperty*, etc.). Figure 2 illustrates a schema graph example of the RDF metamodel described in Figure 3.

2) Semantic and Structural Information: Different types of structural and semantic information can be specified with the different schema definition languages. However, not all of this information is necessary for the matching process. In our approach, we make use of a part of this information that can help to detect mappings between metadata terms. The datatype comparison for

TABLE I.
LANGUAGE MAPPING RULES

Node Types	XSD	RDFS	OWL
Normal Nodes	xsd:element xsd:attribute xsd:attributeGroup xsd:group	rdfs:class	owl:class
Property Nodes	parentName and childName	rdf:property	owl:objectProperty

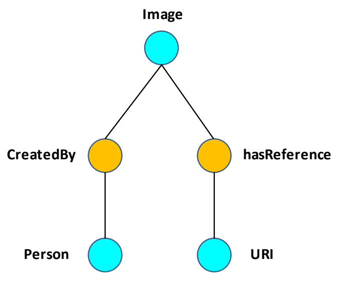

Figure 2. A schema graph example.

```
<rdf:RDF>
-------------------
<rdfs:Class rdf:about="Image">
       <rdfs:subClassOf rdf:resource="&abs;Object"/>
</rdfs:Class>
<rdf:Property rdf:about="hasReference">
       <rdfs:domain rdf:resource="Image"/>
       <rdfs:range rdf:resource="URI"/>
</rdf:Property>
<rdf:Property rdf:about="createdBy">
       <rdfs:range rdf:resource="Image"/>
       <rdfs:domain rdf:resource="Person"/>
</rdf:Property>
--------------------
</rdf:RDF>
```

Figure 3. RDF fragment example.

instance is ignored because we can find a node n_i which matches another node n_j but one of them corresponds to a leaf node (simple type) and the other is a complex node (complex type). Moreover, datatype is not available on all schema definition languages (e.g., all simple types are *rdfs*:*literal* for RDF Schema). Element cardinalities are also another example of ignored information. Table II shows the semantic and structural information used in our matching process. Examples of utilization of some types of this information are given in Section III-B.2

TABLE II.
SEMANTIC AND STRUCTURAL INFORMATION USED ON THE MATCHING PROCESS

XSD	xsd:restriction, xsd:abstraction, xsd:extension, xsd:substitutionGroup
RDFS	rdfs:seeAlso, rdfs:isDefinedBy, rdfs:subClassOf
OWL	owl:unionOf, owl:complementOf, owl:intersectionOf, owl:TransitiveProperty, owl:someValuesFrom, owl:equivalentClass, owl:disjointWith, rdfs:subClassOf owl:distinctMembers, owl:differentFrom owl:AllDifferent, owl:sameAs, owl:equivalentProperty

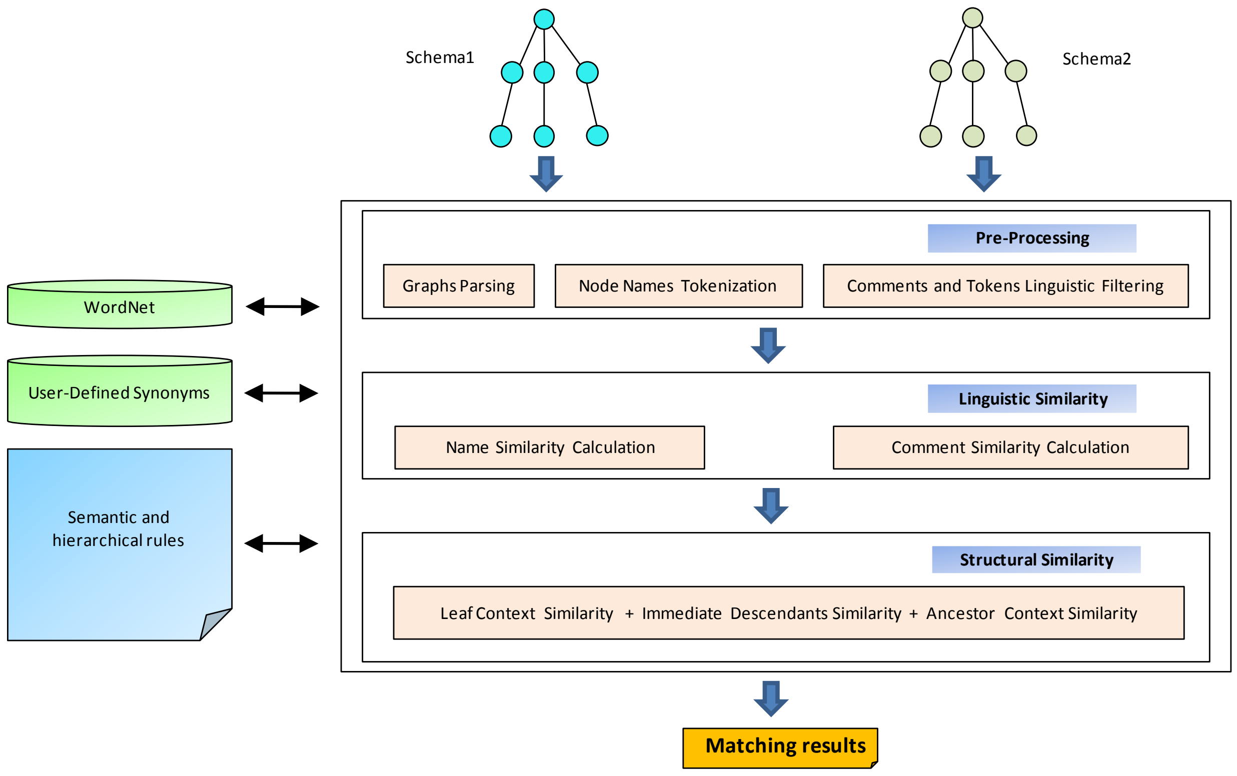
Schema1
Schema2
Pre-Processing
Graphs Parsing
Node Names Tokenization
Comments and Tokens Linguistic Filtering
WordNet
User-Defined Synonyms
Linguistic Similarity
Name Similarity Calculation
Comment Similarity Calculation
Semantic and hierarchical rules
Structural Similarity
Leaf Context Similarity + Immediate Descendants Similarity + Ancestor Context Similarity
Matching results

Figure 4. Matching process phases

B. *Matching System*

In this section, we describe the different steps of the proposed matching system as shown in Figure 4. The system is composed of three main parts: *pre-processing*, *linguistic* and *structural similarity computation*. The system takes as input two schemas presented as directed labeled graphs as well as the structural and semantic information got from the projection step. Then, all irrelevant node names in both schema graphs are eliminated and the useful words are normalized (Section III-B.1). After the *pre-processing* step, the system calculates the linguistic similarity between nodes in both schemas by exploiting the semantics of their corresponding names and comments as well as semantic information obtained from the projection step (Section III-B.2). Finally, the structural similarity is computed and the correct mappings are selected according to the linguistic and structural similarity scores (Section III-B.3).

1) Pre-Processing: In this step, we start by parsing all entities involved in the matching process, including normal nodes, property nodes as well as comments (i.e: textual description available on *xsd:documentation*, *rdfs:comment*, etc.) corresponding to these entities. Then, node names and comments are normalized in order to make their semantics useful for the linguistic similarity calculation step.

a) Node Names: Normally, each entity in the graph is modeled by a node with a name. A node name is a string, without blank characters (space), that may be a word, a term, or an expression (a combination of words). In order to calculate the similarity between node names, a normalization step is necessary. First, each entity name is broken into a set of tokens *M* with a customizable tokenizer using punctuation, upper case, special symbols, and digits, e.g. *MediaRegionLocator* becomes (*Media*, *Region*, *Locator*). Once the tokenization step is over, tokens are lemmatized. Namely, they are morphologically analyzed in order to find all their possible basic forms. Thus, for instance, *Locations* is associated with its singular form, *Location*. A user-defined dictionary is also used to deal with acronyms and abbreviations, e.g., *ID* becomes *Identifier*.

b) Comments: As mentioned above, comments are used as other semantic information. This can be performed via information retrieval techniques. To do so, comments must be linguistically filtered by eliminating the words carrying little useful information, such as articles, prepositions, conjunctions, pronouns and modal verbs [30].

2) Linguistic Similarity Computation: This phase concerns with the linguistic similarity computation between every graph node pairs belonging to two schemas to be mapped. In order to form the linguistic similarity matrix, a string-based technique is used to map the node names. WordNet is used for the explicitation of the words' meaning [10]. In addition to the node names, comments are used as a second semantic resource for the matching process. We apply the TF/IDF technique [30] to these comments in order to extract the most pertinent information. The linguistic similarity score is the weighted sum of both similarities. Finally, the semantic and structural information got from the projection step are used in order to deduce other mappings.

a) Name Matching: The purpose of this phase is to find an initial matching by calculating the similarity distance between the names of all node pairs in the two schemas to be mapped. Each node is represented by a set of tokens. Because of the richness of natural language, we first start with the explicitation of tokens' meaning by using WordNet [19]. Several synonyms can be found for a given term. This helps to resolve problems of terminological conflicts occurring when metadata standards are developed by different communities which may describe the same information using different terms. For instance, some multimedia metadata communities [31] use the term "*type*" to describe the type of a given content. Some others [17] use the terms "*format*" or "*genre*" to describe the same information. Each node n_i represented by a set of tokens M_i will have a set of synonyms *synset* for each token m_i after the explicitation step. $M_i^{'}$ is the final result that regroups all synsets returned by M_i explicitation.

$$M_i^{'} = M_i \bigcup \{m_k | \exists m_j \in M_i \bigcap m_k \in synset(m_j)\} \quad (1)$$

The explicitation is a necessary step because it gives all possible interpretations for each token. However, since a term with multiple senses belongs to multiple synsets, the explicitation process can also result in a wide enlargement of the field of words meaning. Word sense disambiguation (WSD) is a needed operation which allows the selection of the real meaning of a node (the correct synset). In order to achieve WSD we use the method proposed in [18] and enhanced in [25]. This method takes advantage of the network of relations provided in WordNet measure of semantic relatedness between word senses based on the notion of extended gloss overlap (the similariry between text descriptions corresponding to sysnsets). The extended gloss overlap measure takes two synsets as an input, describing two adjacent nodes and computes a gloss overlap score. Synsets which have a maximum score are considered as the best corresponding ones to the nodes they describe [18] [25].

Once the explicitation step is performed, we compute the similarity S_{name} between all node pairs belonging to the two schemas to be mapped. To do so, for each node pair (n_1, n_2) we calculate S_{name} by using Jaro-Winkler metric (JW) [4] between each token $m_i \in M_1^{'}$ and all tokens $m_j \in M_2$ (and vice versa) [19]. The Jaro-Winkler distance is given by:

$$JW(m_i, m_j) = \frac{1}{3}\left(\frac{r}{|m_i|} + \frac{r}{|m_j|} - \frac{r-t}{r}\right) \quad (2)$$

where r is the number of matching characters, t is the number of transpositions and ($|m_i|$, $|m_j|$) are the number of string characters corresponding to tokens (m_i, m_j). We choose Jaro-Winkler distance in our matching precess based on the comparative study done in [7] where the authors made an evaluation of several string distances on the matching process.

We take the maximum score (MJW) of each token m_i:

$$MJW(m_i, M^{'}) = max_{m_j \in M^{'}} JW(m_i, m_j) \quad (3)$$

Finally, the average of the best similarities is calculated to get the name similarity between nodes:

$$S_{name}(n_1, n_2) = \frac{\sum_{m_i \in M_1} MJW(m_i, M_2^{'}) + \sum_{m_j \in M_2} MJW(m_j, M_1^{'})}{|M_1| + |M_2|} \quad (4)$$

where $|M_1|$ and $|M_2|$ are the number of tokens in M_1 and M_2.

b) Comment Matching: Due to the use of technical vocabularies by multimedia metadata communities, node names do not always provide a sufficient semantics. The comments related to each entity are also another semantic resource. We apply the TF/IDF technique [30] used in the information retrieval domain in order to calculate the similarity between comments. To do so, all comments on the two schemas to be mapped are considered as documents, each node will be represented by a vector whose coordinates are the results of TF/IDF. Hence, the similarity between two nodes is the distance between vectors corresponding to their comments.

In order to illustrate how to calculate these vectors, let us consider *v* = (w_1, w_2,...., w_P), a vector representing a certain node *n*. *P* = $|U|$ is the number of distinct words in all comments in the two schemas to be mapped. The i_{th} element w_i in the vector *v*, which represents the node *n* in a schema, is calculated as follows:

$$w_i = tf_i * idf_i \quad (5)$$

$$idf_i = \log_2 \frac{N}{b_i} \quad (6)$$

where tf_i is the term frequency. tf_i represents the number of times that the i_{th} word in *U* appears in the comment corresponding to n_i. idf_i (inverse document frequency) is the inverse of the percentage of the concepts which contain the word w_i. *N* is the number of comments in *U* in both schemas. b_i is the number of comments which contain the word w_i at least one time. As we have mentioned previously, the similarity $S_{comment}$ between two nodes n_i and n_j is the distance between vectors corresponding to their comments v_i and v_j. This distance is a cosine similarity σ [28]. It is calculated as follows:

$$\sigma(v_i, v_j) = \frac{\sum_{k=1}^{P} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{P} (w_{ik})^2 * \sum_{k=1}^{P} (w_{jk})^2}} \quad (7)$$

The result of above processes is a linguistic similarity matrix *lSim*, where:

$$lSim(n_i, n_j) = \mu_1 * S_{name}(n_i, n_j) + \mu_2 * S_{comment}(n_i, n_j) \quad (8)$$

$\mu_1 + \mu_2 = 1$ and $(\mu_1, \mu_2) \geq 0$

c) Semantic and Structural Information Utilization: Structural and semantic information obtained from the projection step (Section III-A) are used in our system to detect other mapping candidates, especially complex ones (n:m mappings). In the following, we present some examples of such features and show how they can be used to deduce other mapping candidates:

- **Generalization Features:** The generalization relationship between two types indicates that one type is a subset of another (e.g., *xsd*:*extension*, *rdfs*:*subClassOf*, *xsd*:abstraction, etc.) [9] [21]. This information can successfully help the matching algorithm to infer complex matches. Let us consider that the attribute *ms*: *Agent* is defined at the schema to be matched with MPEG-7 schema, this attribute linguistically map to *mpeg7*:*AgentType* complex element (*lSim* between both node is greater than a given threshold). In this context, the union of two complex elements $mpeg7{:}PersonType$ and $mpeg7{:}OrganizationType$ is also considered as a mapping candidates for *ms*:*Agent*. This is done because *mpeg7*:*OrganizationType* and *mpeg7*:*PersonType* are extensions of *mpeg7*:*AgentType*. Other structural properties are used (e.g., *owl*:*disjointWith*, *owl*: *unionOf*, etc.)

- **Semantic Features:** Let us consider that n_i and n_j are two nodes linguistically matched, and n_k is another node corresponding to a class having an equivalence property with the one corresponding to n_i (e.g., *owl*: *equivalentProperty*, *owl*:*equivalentClass*, *xsd*:*substitutionGroup*, etc.) [9] [21]. This information allows us to deduce that n_k is another mapping candidate for n_j. Other semantic properties are also used (e.g., *owl*:*sameAs*, *owl*:*differentFrom*, etc.)

The use of semantic and structural information helps to discover more mappings but may also increase the number of false positive mappings that can be eliminated in the structural similarity computation step (Section III-B.3).

3) Structural Similarity Computation: The linguistic similarity computation may provide several node candidates. There can be multiple matching candidates which differ in the structure but have a high linguistic similarity value. For instance, *mpeg7: MediaRelTimePoint* and *mpeg7: MediaRelIncrTimePoint* [15] are two elements which may map to the same entity *ms:MediaTimePoint* defined in the mediated schema. Thus, in order to deal with this case, the structural similarity is computed in order to prune these false positive candidates. Our structural matching algorithm is based on the node context, which is reflected by its ancestors and its descendants. In this paper, as in [16], we consider three kinds of node contexts depending on their position in the ontology tree: *ancestor context*, *immediate descendant context* and *leaf context*. The context of a node is a combination of these three contexts.

a) Ancestor Context Similarity Measure: The ancestor context of a node n_i is defined as the path p_i extending from the root node of the schema to n_i. Consequently, in order to compare two ancestor contexts, we essentially need to compare their corresponding paths. The authors in [16] have introduced the concepts of path context coefficient to capture the degree of similarity in the paths of two elements. However, this solution has no high matching accuracy. For this reason, we use the ideas of path similarity measure described in [8]. The authors in [8] have relaxed the matching condition by allowing the matching of paths even if their source nodes do not match and their nodes appear in a different order. In addition, paths can also be matched even if there are additional nodes within the path, meaning that the child-parent edge constraint is relaxed into ancestor-child constraint. Such relaxations are inspired by ideas in query answering to approximate answering of queries. The authors in [8] consider two paths p_1 and p_2 being matched, p_2 is the best match candidate for p_1 if it fulfils the following criteria:

- The path p_1 includes most of the nodes of p_2 in the right order.
- The occurrences of the p_1 nodes are closer to the beginning of p_2 than to the tail, meaning that the optimal matching corresponds to the leftmost alignment.
- The occurrences of the p_1 nodes in p_2 are close to each other, which means that the minimum of intermediate non matched nodes in p_2 are desired.
- If several matching candidates that match exactly the same nodes in p_1 exist, p_2 is the shortest one.

In order to answer to the creteria mentioned above, four parameters have been defined in [8]: $lcs_n(p_i, p_j)$, the longest common subsequences between two paths normalized by the length of the first path, $pos(p_i, p_j)$ which considers that the optimal matching between (p_i, p_j) is the matching that starts on the first element of p_i and continues without gaps, $gap(p_i, p_j)$ used to ensure that the occurrences of two path nodes are close to each other, and $ld(p_i, p_j)$ used to give higher values to source paths whose length is similar to target paths.

Based on the criteria for the paths matching mentioned above, we introduce a new adaptation and relaxation of the approach as follows:

- For the fourth parameters, the *longest common subsequence*(*lcs*) must be calculated according to the linguistic similarity matrix computed in the previous step. It means that two nodes (n_i, n_j) are considered identical if their $lSim(n_i, n_j)$ is greater than a given threshold e.g. 0.80. That is, ordinary string comparison is now relaxed into string and text comparison $(lSim)$ that is based on similarity threshold.
- The parameters in [8] have been defined for XML which is based on a taxonomy classification (relations between nodes are not labeled, generally containment type). Therefore, the transformation of all schemas into a directed labeled graphs needs more relaxation for the *lcs* calculation. For instance, let us consider the two RDF triples (*Objet*, *ContentOn*, *URL*) and (*Object*, *hasReference*, *URL*). We can easily note that the predicate value of the first triple matches the subject value (*Content* is equivalent to *Object*). However, the predicate value in the second triple matches the object (*Reference* is equivalent to *URL*). We introduce another type of relaxation by considering triple predicates (property nodes) as nodes which are switchable with the immediate child node or the immediate parent node (object or subject according to RDF interpretation). To do so, we consider each path as a set of segments

$p = \{g^1, ..., g^N\}$, each segment is composed of two adjacent nodes (n_k, n_{k+1})(one of them in a property, see Figure 5). In this context, $lcs^{'}$ is the new value of *longest common subsequence* between two paths after the relaxation step. Its calculation is showed in Algorithm 1.

Algorithm 1 $lcs^{'}$ calculation

```
1: Input parameters $(p_1, p_2)$.
2: Each path is a set of segments $p_1 = \{g_1^1, ...., g_1^N\}$, $p_2 = \{g_2^1, ...., g_2^M\}$.
3: Each segment is a composed of two adjacent nodes $g_i = (n_i, n_{i+1})$.
4: int $lcs^{'}$ = 0. int i = 0.
5: calculate $lcs(p_1, p_2)$.
6: $lcs^{'} = lcs(p_1, p_2)$.
7: for all segments $g_i^1$ $(i \in N)$ do
8:     switch nodes $(n_i, n_{i+1})$ of segment $g_i^1$ and update $p_1$.
9:     calculate $lcs(p_1, p_2)$
10:    if $(lcs(p_1, p_2) > lcs^{'})$ then
11:        $lcs^{'} = lcs(p_1, p_2)$
12:        pass to the after next segment (i := i + 2).
13:    else
14:        return $(n_i, n_{i+1})$ to their previous positions and update $p_1$.
15:        pass to the next segment (i := i + 1).
16:    end if
17: end for
18: return $lcs^{'}$
```

After the relaxations introduced above, the new version of equations presented in [8] becomes:

$$lcs_n(p_i, p_j) = |lcs'(p_i, p_j)|/|p_i| \tag{9}$$

$$pos(p_i, p_j) = 1 - ((ap - aop)/(|p_j| - 2 * aop + 1)) \tag{10}$$

$$gap(p_i, p_j) = gaps/(gaps + lcs'(p_i, p_j)) \tag{11}$$

$$ld(p_i, p_j) = (|p_j| - lcs'(p_i, p_j))/|p_j| \tag{12}$$

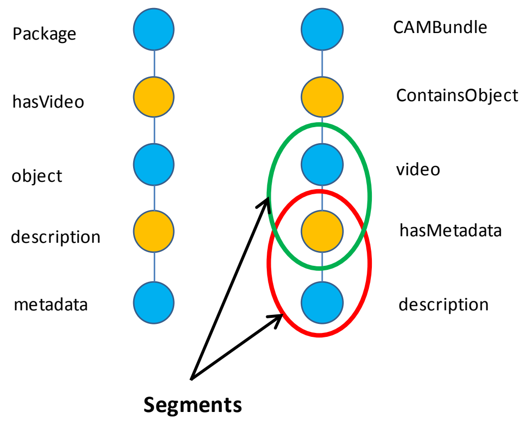

Figure 5. Example of paths and segments.

If we consider the two paths showed in Figure 5 as the path being matched, the *longest common subsequence* is 5 instead of 0 after the two steps of relaxation introduced above.

Finally, the similarity *ps* between two paths (p_i, p_j) is obtained by combining the scores resulting from the equations (9, 10, 11 and 12):

$$ps(p_i, p_j) = \delta lcn_n(p_i, p_j) + \varphi pos(p_i, p_j) - \theta gap(p_i, p_j) - \lambda ld(p_i, p_j) \tag{13}$$

where δ, φ, θ and λ are positive parameters representing the comparative importance of each factor. These parameters are given based on experimental results from [8]: δ = 0.75; φ = 0.25; θ=0.25; λ=0.2.

Finally, the ancestor context similarity *ancSim* between two nodes (n_i, n_j) is the path similarity of their ancestors (p_i, p_j) weighted by their linguistic similarity $lSim(n_i, n_j)$:

$$ancSim(n_i, n_j) = ps(n_i, n_j) * lSim(n_i, n_j) \tag{14}$$

b) Immediate Descendant Context Similarity Measure: To obtain the immediate descendant context similarity *immSim* between two nodes (n_i, n_j), we compare their two immediate descendant sets $S = \{s_1, s_2, \cdots, s_n\}$ and $S^{'} = \{s_1^{'}, s_2^{'}, \cdots, s_m^{'}\}$ (immediate descendant nodes are normal nodes, property nodes are included only for path similarity calculation). This is done by using the similarity *SimI* between each pair of children in the two sets. Where:

$$SimI(s_i^{'}, s_j) = ps(s_i^{'}, s_j) * lSim(s_i^{'}, s_j) \tag{15}$$

$ps(s_i^{'}, s_j)$ is the similarity between paths extending from $(s_i^{'}, s_j)$ to (n_i, n_j).

Then, we select the matching pairs with maximum similarity values.

$$MaxSimI(s_i, S^{'}) = max_{s_j^{'} \in S'} SimI(s_i, s_j^{'}) \tag{16}$$

$$MaxSimI(s_i^{'}, S) = max_{s_j \in S} SimI(s_i^{'}, s_j) \tag{17}$$

Finally, the average of the best similarity values is taken to get the immediate descendant similarity value *immSim*:

$$immSim(n_i, n_j) = \frac{\sum_{i=1}^{|S|} MaxSimI(s_i, S^{'}) + \sum_{i=1}^{|S^{'}|} MaxSimI(s_i^{'}, S)}{|S\prime| + |S|} \tag{18}$$

c) Leaf Context Similarity Measure: The leaf context of a node n_i is defined as the set of leaf nodes of subtrees rooted at n_i. If $l_i \in$ *leaves*(n_i) is a leaf node, then the context of l_i is given by the path p_i from n_i to l_i. The leaf context is given by:

$$leafSim(l_i, l_j) = ps((p_i, p_j)) * lSim(l_i, l_j) \tag{19}$$

To obtain the leaf context similarity between two leaves $l_i \in$ *leaves*(n_i) and $l_j \in$ *leaves*(n_j), we compute the leaf similarity *leafSim* between each pair of leaves in the two leaf sets. We then select the matching pairs with the maximum similarity values. The average of the best similarity values is taken (see Section III-B.3b) *d) Node Similarity:* The node similarity *nodeSim* between normal nodes (property nodes are included only for path similarity calculation) can be obtained by the combination of *ancestor context*, *immediate descendant context*, and *leaf context* similarities unless one of the two nodes being compared is a leaf node or a node child of root. In this case, the node similarity calculation considers that the context of both nodes depends only on their ancestors or descendants. The node similarity is given by:

$$nodeSim(n_i, n_j) = \alpha * ancSim(n_i, n_j) + \beta * immSim(n_i, n_j) + \gamma * leafSim(n_i, n_j) \quad (20)$$

$$\alpha + \beta + \gamma = 1 \text{ and } (\alpha, \beta, \gamma) \geq 0$$

Once the structural similarity computation is made, the system returns for each source node n_i the k node candidates that have the maximum values of *nodeSim*. In order to select the k candidates, one of the following strategies can be used:

Threshold: Returns all node pairs showing a similarity exceeding a given threshold value τ_2. This strategy may return too many matched candidates.

MaxDelta: Returns the node pair having a maximum similarity value *nodeSim* which is determined as candidate plus all pairs with a similarity differing at most by a tolerance value *d*.

MaxN: The N node pairs with maximal similarity *nodeSim* are selected as matching candidates.

In our approach, we support considering several criteria at the same time, in particular MaxN in combination with a low threshold e.g. 0.80.

IV. Experimental Evaluation

In this section, we describe the experiments that we have carried out to evaluate our proposed system. Firstly, we describe the data sets which we have used through the evaluation. Secondly, we show the experimental results which allow us to evaluate the proposed system.

A. Data Sets

The system has been tested using several metadata standards (MPEG-7, MPEG-21, EXIF, MIX, DIG35, XMP, MPEG-7 Ontology, PeCMan Ontology, TV-Anytime, DIG35 Ontology and CAM4Home) [15] [3] [2]. These standards have a significant heterogeneity on two levels: the metadata specification is performed using different languages and schemas that present a high structural and semantical heterogeneity (Table III).

In order to compare our results with Cupid [20], we must first choose only metadata based XML Schema. To do so, we calculated the mapping between TV-Anytime and all standards based XSD. Secondly, so we can compare our system with SF [22] we calculated the mapping between CAM4Home metadata model presented in [3] and all metadata standards in Table III. CAM4Home metadata model is based on RDF Schema and it is a part of the CAM4Home ITEA2 project [3]. A group of twenty multimedia academic and industrial practitioners from TV, 3G and Internet application fields defined a large set of metadata requirements in order to support the convergence of multimedia contents in Digital Home environments. Metadata defined under CAM4Home project describes information related to content semantics, user characteristics and device profiles. The ground truths of metadata standard mappings is obtained from [17]

TABLE III.
Characteristics of metadata used in the experimentation

Metadata Standard	Max Depth	# Nodes	Language
MIX	7	41	XSD
DIG35	9	57	XSD
EXIF	6	63	XSD
MPEG-7	13	115	XSD
MPEG-21	11	143	XSD
PeCMan Ontology	7	46	OWL
DIG35 Ontology	9	57	OWL
CAM4Home	17	153	RDFS
TV-Anytime	8	71	XSD
XMP	6	53	XSD/RDFS
MPEG-7 Ontology	10	98	OWL

B. Experimental Results

The experimentation starts by setting the $\mu_1, \mu_2, \alpha, \beta$ and γ parameters to know their effect on the matching process. Then, the matching quality measure is showed in Section IV-B.3. Finally, a comparative study is given in Section IV-B.4.

1) Tuning the Parameters μ_1 and μ_2: In this Section we show the effect of combining the two similarity measures. We have selected two sets of attributes from several metadata standard (T_1, T_2), each element of T_i has an equivalent attribute in T_j. We have calculated the linguistic similarity ($lSim$) between the elements in two sets by using several combinations of the parameters μ_1 and μ_2. The attributes are considered as matched if their linguistic similarity value are greater than 0.9 (τ_1). Figure 6 shows the experimental results in terms of precision (percentage of correct mappings), where we can note that the name of the attribute has more importance to the comments. However, the comments provide also information that helps to discover the mapping.

2) Tuning the Parameters α, β and γ: In order to know the influence of each similarity context, we calculated the mapping between several metadata standards using several combinations of the α, β and γ parameters. The mapping results between (MPEG-7, DIG35) and (DIG35, EXIF) in terms of *F-Measure* [20] is showed in Figure 7 and 8 respectively.

The experimental evaluation shows that the greatest amount of structural information is contained in the ancestor context (Figure 7 and 8) where the highest values of F-Measure are located when $\alpha \in$ [0.45 0.6]; this explains

[3]http://www.cam4home-itea.org/

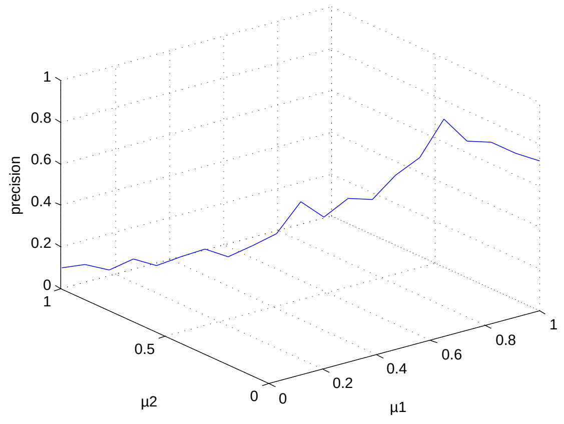

Figure 6. The effect of the parameters μ_1 and μ_2 on the linguistic similarity calculation

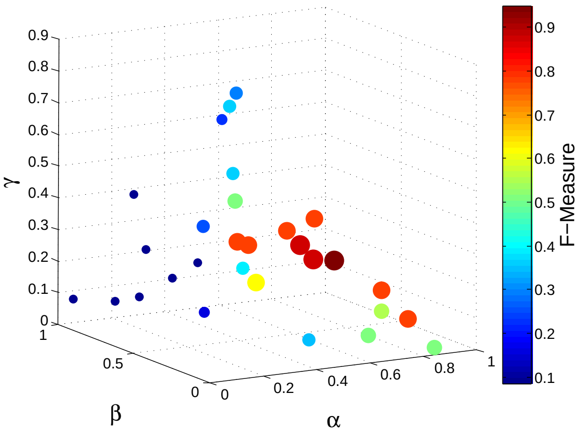

Figure 8. DIG35, EXIF mapping results in term of F-measure

the interest of some matching strategies which consider that the context of only nodes depend on their ancestors [5] [24] [29]. Besides, the experimentations showed that the immediate descendant context and the leaf context have also an effect on the matching process, where the highest values of F-Measure is located when $\beta \in [0.1\ 0.15]$ and $\gamma \in [0.25\ 0.4]$. However, we believe that an automatic choice of the α, β and γ values for each pair of nodes according to the node positions in the graph can increase the matching accuracy. This will be part of our future works.

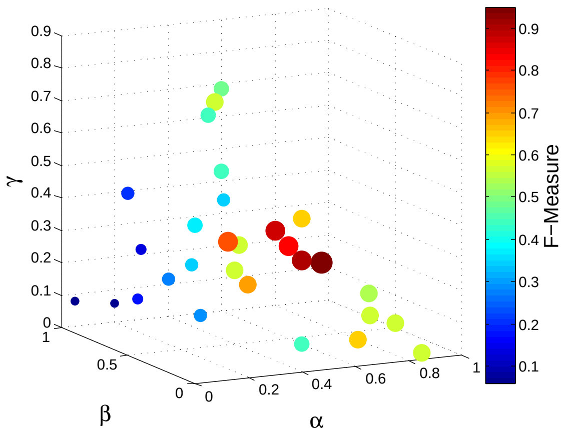

Figure 7. MPEG-7, DIG35 mapping results in term of F-measure.

3) Matching Quality Measure: In order to calculate the similarities between nodes, equation (20) parameters are given according to the experiment results from the previous section ($\mu_1 = 0.85, \mu_2 = 0.15, \alpha = 0.55, \beta = 0.15, \gamma = 0.30$). The proposed solution has been evaluated by considering the matching quality based on the criteria described in [20] [22], including *Precision*, *Recall*, *F-measure*. The experiment results of the mapping between TV-Anytime and other standards based XSD in terms of *Precision*, *Recall* and *F-measure* are shown in Table IV. Table V shows the mapping results between CAM4Home and other standards. By analyzing the two tables above, we can notice that the results are important from our point of view where the average value of F-measure is 62% and 81% for the Tables IV and V respectively. The results in Table V are higher than the results in Table IV, this is due to the type of schema used to integrate metadata (TV-Anytime for Table IV and CAM4Home for Table V) where the complexity of metadata vocabulaires (terms) change from one schema to another.

TABLE IV.
MAPPING RESULTS BETWEEN TV-ANYTIME AND STANDARDS BASED XSD

Metadata standards	Precision	Recall	F-measure
MIX	50%	55%	52%
EXIF	62%	65%	63%
MPEG-7	75%	70%	72%
MPEG-21	43%	47%	45%
DIG35	77%	82%	79%

TABLE V.
MAPPING RESULTS BETWEEN CAM4HOME AND OTHER STANDARDS

Metadata standards	Precision	Recall	F-measure
MIX	92%	89%	90%
DIG35	87%	90%	88%
EXIF	90%	81%	85%
MPEG-7	77%	64%	70%
MPEG-21	72%	60%	65%
PeCMan Ontology	94%	89%	91%
DIG35 Ontology	85%	90%	87%
TV-Anytime	77%	72%	75%
XMP	57%	65%	61%
MPEG-7 Ontology	70%	68%	69%

4) Comparison With Other Systems: In order to compare the proposed solution with other systems, we have implemented SF [22] and Cupid [20]. The choice of Cupid and SF is done because they are schema-based, and they all utilize linguistic and structural matching techniques. The result of this comparative study in terms of F-measure score is showed in Figure 9 and 10 where our solution is

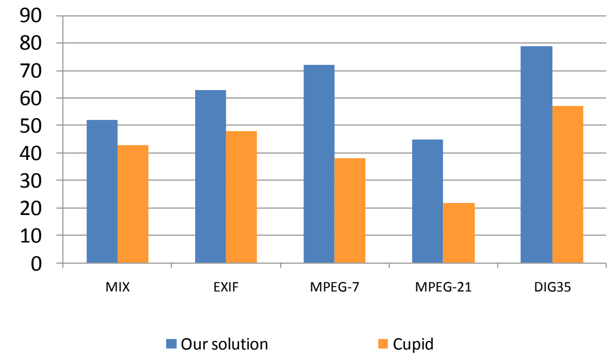

Figure 9. Comparative study with Cupid in term of F-measure.

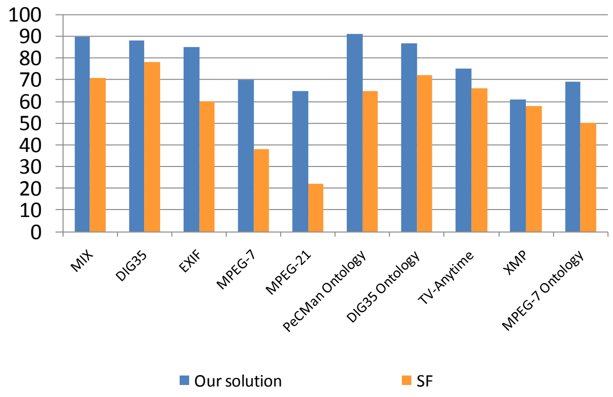

Figure 10. Comparative study with SF in term of F-measure.

better than SF and Cupid for all tested metadata standards. This is due to the use of a single context for both systems (leaf context for Cupid and child context for SF). Whereas, in our solution, we have considered all ancestor-context, child context and leaf-context. The use of the semantic and hierarchical rules has also has improved the accuracy of our solution (Section III-B.2), where 68 % of complex matchings have been successfully detected. Besides our system is general and not limited to one or two schema definition languages.

V. Conclusion

Due to the extensive use of metadata and their semantic and structural heterogeneity, there has been a great interest to develop an integration system for achieving metadata interoperability. The existence of such model is crucial for media object uniform access fulfillment. To do so, we proposed and implemented in this paper a new integration technique for achieving metadata interoperability whatever the definition language. We essentially proposed a matching system that supports metadata schemas whatever their description language. It uses several types of syntactic, semantic and structural information to match between metadata. Our experiments showed that the combination of the linguistic and structural similarities plays a significant role in deriving a correct mapping.

In our ongoing works, we plan to enhance the proposed matching system through a better use of structural information. This can be achieved by adding a new structural matching technique to the system. We mainly explore the use of adjacency nodes to detect other mappings that cannot be detected by the current matching strategy. We also plan to enhance the proposed approach by taking into account the mappings already validated by the user as another structural information which may help to find other mappings.

VI. Acknowledgements

This work was supported by the ITEA2 CAM4HOME European project.

References

[1] S. Amir, “Un système d’intégration de métadonnées dédiées au multimédia,” in *INFORSID*, 2009, pp. 491–492.

[2] S. Amir, I. M. Bilasco, T. Urruty, J. Martinet, and C. Djeraba, *Designing Intelligent Content Delivery Frameworks Using MPEG-21*. UK: John Wiley and Sons, 2010, ch. 20, pp. 455–475.

[3] I. M. Bilasco, S. Amir, P. Blandin, C. Djeraba, J. Laitakari, J. Martinet, E. Martínez-García, D. Pakkala, M. Rautiainen, M. Ylianttila, and J. Zhou, “Semantics for intelligent delivery of multimedia content,” in *SAC*, 2010, pp. 1366–1372.

[4] M. Bilenko, R. J. Mooney, W. W. Cohen, P. D. Ravikumar, and S. E. Fienberg, “Adaptive name matching in information integration,” *IEEE Intelligent Systems*, vol. 18, no. 5, pp. 16–23, 2003.

[5] P. Bouquet, L. Serafini, and S. Zanobini, “Semantic coordination: A new approach and an application,” in *International Semantic Web Conference*, 2003, pp. 130–145.

[6] D. Brickley and R. Guha, “Rdf vocabulary description language 1.0: Rdf schema,” W3C, Amsterdam, `http://www.w3.org/TR/rdf-schema/`.

[7] W. W. Cohen, P. D. Ravikumar, and S. E. Fienberg, “A comparison of string distance metrics for name-matching tasks,” in *IIWeb*, 2003, pp. 73–78.

[8] D.Carmel, Y. S. Maarek, M. Mandelbrod, Y. Mass, and A. Soffer, “Searching xml documents via xml fragments,” in *SIGIR*, 2003, pp. 151–158.

[9] J. P. C. S.-M. E. M. F. Yergeau, T. Bray, “Extensible markup language (xml) 1.0 (third edition) w3crecommendation.” W3C, Amsterdam, `http://www.w3.org/TR/2004/REC-XML-20040204/`. [Online]. Available: ¡http://www.w3.org/TR/2004/REC-XML-20040204/¿

[10] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[11] J. Fong, “Converting relational to object-oriented databases,” *SIGMOD Record*, vol. 26, no. 1, pp. 53–58, 1997.

[12] R. Garca and O. Celma, “Semantic integration and retrieval of multimedia metadata,” in *2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media*, 2005, pp. 69–80.

[13] F. Giunchiglia, P. Shvaiko, and M. Yatskevich, “S-match: an algorithm and an implementation of semantic matching,” in *ESWS*, 2004, pp. 61–75.

[14] B. Haslhofer and W. Klas, “A survey of techniques for achieving metadata interoperability,” *ACM Comput. Surv.*, vol. 42, no. 2, 2010.

[15] M. Hausenblas, “Multimedia vocabularies on the semantic web,” W3C, Amsterdam, `http://www.w3.org/2005/Incubator`.

[16] M.-L. Lee, L. H. Yang, W. Hsu, and X. Yang, “Xclust: clustering xml schemas for effective integration,” in *CIKM*, 2002, pp. 292–299.

[17] W. Lee, T. Brger, F. Sasaki, and V. Malaisé. (2008, Jul.) Use cases and requirements for ontology and api for media object. W3C. Amsterdam. [Online]. Available: http://dev.w3.org/2008/video/mediaann/mediaont-1.0/

[18] M. Lesk, “Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone,” in *SIGDOC ’86: Proceedings of the 5th annual international conference on Systems documentation*. ACM Press, 1986, pp. 24–26.

[19] F. Lin and K. Sandkuhl, “A survey of exploiting wordnet in ontology matching,” in *IFIP AI*, 2008, pp. 341–350.

[20] J. Madhavan, P. A. Bernstein, and E. Rahm, “Generic schema matching with cupid,” in *VLDB*, 2001, pp. 49–58.

[21] D. L. McGuinness and F. v Harmelen, “Owl web ontology language overview,” W3C, Amsterdam, `http://www.w3.org/TR/owl-features/`.

[22] S. Melnik, H. Garcia-Molina, and E. Rahm, “Similarity flooding: A versatile graph matching algorithm and its application to schema matching,” in *ICDE*, 2002, pp. 117–128.

[23] I. Miletic, M. Vujasinovic, N. Ivezic, and Z. Marjanovic, “Enabling semantic mediation for business applications: Xml-rdf, rdf-xml and xsd-rdfs transformations,” in *IESA*, 2007, pp. 483–494.

[24] R. J. Miller, M. A. Hernández, L. M. Haas, L.-L. Yan, C. T. H. Ho, R. Fagin, and L. Popa, “The clio project: Managing heterogeneity,” *SIGMOD Record*, vol. 30, no. 1, pp. 78–83, 2001.

[25] S. Patwardhan, S. Banerjee, and T. Pedersen, “Using measures of semantic relatedness for word sense disambiguation,” in *CICLing*, 2003, pp. 241–257.

[26] L. Popa, Y. Velegrakis, R. J. Miller, M. A. Hernández, and R. Fagin, “Translating web data,” in *VLDB*, 2002, pp. 598–609.

[27] P. Shvaiko and J. Euzenat, “Ten challenges for ontology matching,” in *OTM Conferences (2)*, 2008, pp. 1164–1182.

[28] D. A. Simovici and C. Djeraba, *Mathematical Tools for Data Mining: Set Theory, Partial Orders, Combinatorics*. Springer Publishing Company, Incorporated, 2008.

[29] G. Stumme and A. Maedche, “Fca-merge: Bottom-up merging of ontologies,” in *IJCAI*, 2001, pp. 225–234.

[30] C. J. van Rijsbergen, *Information Retrieval*, 2nd ed. London: Butterworths, 1979.

[31] X. working group. (2008, Feb.) Extensible metadata platform specification. XMPSpecificationPart2.pdf. [Online]. Available: http://www.adobe.com/devnet/xmp/pdfs/

[32] K. Yang, R. Steele, and A. Lo, “An ontology for xml schema to ontology mapping representation,” in *iiWAS*, 2007, pp. 101–111.

Appendix

A. Equation Definitions

TABLE VI.
Descriptions of equations used by the mapping system

Equation	Description
$lcs_n(p_i, p_j)$	it is the longest common subsequences between two paths (p_i, p_j), normalized by the length of the first path
$pos(p_i, p_j)$	it considers that the optimal matching between two paths (p_i, p_j) must starts on the first element of p_i and continues without gaps
$gap(p_i, p_j)$	it is used to ensure that the occurrences of two path nodes are close to each
$ld(p_i, p_j)$	used to give higher values to source paths whose length is similar to target paths.
$S_{name}(n_i, n_j)$	name similarity between the nodes (n_i, n_j)
$S_{comment}(n_i, n_j)$	represents the similarity value between comments corresponding to nodes (n_i, n_j)
lSim(n_i, n_j)	linguistic similarity between nodes (n_i, n_j)
$ps(p_i, p_j)$	the similarity between two paths (p_i, p_j)
leafSim(n_i, n_j)	the leaf context similarity between two nodes
ancSim(n_i, n_j)	the ancestor context similarity between two nodes
immSim(n_i, n_j)	the immediate descendant context similarity between two nodes
nodeSim(n_i, n_j)	the final similarity between two nodes (n_i, n_j)

B. Parameters Definitions

TABLE VII.
Descriptions of parameters used by the mapping system

Parameter	Description
μ_1	the comparative importance of name similarity
μ_2	the comparative importance of comment similarity
τ_1	linguistic similarity threshold
δ	represents the comparative importance of lcs_n
φ	represents the comparative importance of pos
θ	represents the comparative importance of gap
λ	represents the comparative importance of ld
α	represents the comparative importance of *ancSim*
β	represents the comparative importance of *immSim*
γ	represents the comparative importance of *leafSim*
τ_2	the final similarity threshold