Abstract: This paper presents our approaches in integrating heterogeneous metadata records in Peer-to-Peer (P2P) based digital libraries (DL). In this paper, the advantages of adapting P2P network over other approaches are to be presented in searching information among moderate-sized digital libraries. Before we present the semantic integration solution, we describe the P2P architecture built in JXTA protocol. By adopting JXTA protocol, peers can automatically discover the other candidates which can provide most appropriate answers. Such feature is realized by the advertising functionality which is introduced in the query process in the paper. As to the metadata integration, since resources may adopt distinct metadata, standardized or non-standardized, we employ the most widely adopted Dublin Core [17] as a globally shared metadata to sponsor the interoperation. This paper also describes the mechanism of applying inference rules to convert heterogeneous metadata to local repository.

Categories and Subject Descriptors
H.3 [Information Storage and Retrieval]: H.3.7 [Digital Libraries]: C.2.1 [Network Architecture and Design]

General Terms
Digital Library, Peer to peer network

Keywords: peer to peer network, metadata, digital libraries, semantic integration

Received 30 Oct. 2004; Reviewed and accepted 30 Jan. 2005

1 Introduction
The near universal availability of high-speed network connected to the Web makes it possible to have the distribution and access of electronic resources from anywhere in the Web. Various strategies have been rendered to collect, organize and process such enormous data and information efficiently. For instance, database and XML technologies enable us to organize and search data and information easier according to specific features, such as structured information. However, even if data has already been in structured format, such a huge data volume still prevents users to consume them efficiently. Under such situation, metadata, briefly explained as ‘data of data’, has emerged as a strategy to assist people to get directly to the basic but critical context of data/information [15]. It is achieved by comparably concise metadata information.

However, there are many metadata standards created in isolated organizations for various topics/domains and different purposes. It is just like data was managed 30 years ago - there are many data available but few methods can be used to process them efficiently. As to heterogeneous metadata sets, probably it is affordable for globally shared metadata to sponsor the interoperation. In this paper, the concepts of ‘node’ and ‘peer’ are interchangeable.

When a global mediating schema is conducted to accommodate all local schemas, Peer-to-Peer (P2P) based data management has yet received much attention recently as a more scalable and flexible architecture for sharing heterogeneous data. The major difference from primitive approaches is that there is no central mediator node. One peer can harvest data from another peer if the mapping between the two schemas is defined. Hence, the user can be set free from concentrating on the centralized mediator node. But such architecture suffers one problem of how to discover the peers that hold relevant data. Currently, search mechanisms in P2P network can be roughly classified into three categories. First, a central server maintains a global index for all participating peers and all queries are directed to the server, for Napster [4] as such. Such solution suffers the same performance problem with exponential growth of the data. The second type of systems adopts some form of flooding [5] or gossiping [6], which has known limitations on network utility. A distributed hash table (DHT) based solution is the third type where the nodes’ structurally organized by a hash function. The DHT-based mechanism is widely adopted to locate data sources in large scale systems [7]. One example is Chord [8], which is scalable in locating data by adopting a hash table. However, metadata could not be simply treated as keywords and located through a hash function in Chord or other DHT-based systems.

In this paper, we aims to alleviate the problem of semantic-based content search in federated DLs. Especially, we focus on the mechanisms in integrating heterogeneous schemas in P2P network. We propose to use the JXTA Search [9] as an experimental platform. We argue that JXTA Search is appropriate for searches of distributed data sources that actively produce data, such as the news website or ordinary DL systems. The system builds indices on the queries a data source can answer and distributes them across JXTA hubs to which data sources are connected. Thus, the way catalog information is distributed across the hubs is determined by where the data providers connect. It is anticipated that providers of similar topics will connect to the same hubs. In our case we implement it in a service discovery style which can be also easily extended to Web services. We assume in this paper the collections in independent and distributed nodes are XML collections. This premise originates from the fact that many textual contents in DL are and will be encoded with XML document markup [10][11]. However, one example will be described as well in converting from database tables to XML records. It is clear that XML alone cannot meet the requirements of functionality and interoperability originally anticipated, though. The problem is that XML has not been designed for representing the fundamental semantic relationships among document components and features; even it is able to identify a document’s meaningful structure. Coming back to the starting point, in order to support semantics-based searching, it is inevitable to consider schema and ontology mapping issues in the retrieval procedure. Herein, we differentiate for this paper the meaning of schema and ontology slightly. Schemas, which is to be carried out by document type definitions (DTD) or XML Schema [12], focus on the structural and syntax of an XML vocabulary adopted in collections, while ontologies [13] are leveraged for basic facts and semantic relationships represented by the XML constructs.

The rest of paper is organized as follows. Section 2 introduces the P2P infrastructure we are to work with. Section 3 demonstrates an example to demonstrate the necessity of semantic markups in independent collections located peers; and as well indicates our intended approaches and tentative solutions. Section 4 reviews current related works. Conclusions and future works are drawn finally in Section 5.

In this paper, we present our approaches in integrating heterogeneous metadata records in Peer-to-Peer (P2P) based digital libraries (DL). In this paper, the advantages of adapting P2P network over other approaches are to be presented in searching information among moderate-sized digital libraries. Before we present the semantic integration solution, we describe the P2P architecture built in JXTA protocol. By adopting JXTA protocol, peers can automatically discover the other candidates which can provide most appropriate answers. Such feature is realized by the advertising functionality which is introduced in the query process in the paper. As to the metadata integration, since resources may adopt distinct metadata, standardized or non-standardized, we employ the most widely adopted Dublin Core [17] as a globally shared metadata to sponsor the interoperation. This paper also describes the mechanism of applying inference rules to convert heterogeneous metadata to local repository.
2 Architecture

2.1 General Architecture

The advantages of adopting P2P architecture has been discussed in [14] from the perspectives of three Internet assets, namely, information, bandwidth and computing resources. P2P technologies can adopt a network-based computing style that neither excludes nor inherently depends on centralized control points. Additionally, the data and information resources on the Web are also very dynamic, such as the collections in personal digital libraries, news website, product information and auctions. Generally Web crawlers used in search engines are ideal for static contents but not for dynamic ones. Thus, a complementary data sharing and integration mechanism is highly required for such applications.

The nodes in the P2P network are more or less identical from the aspect of functionality. However, considerable co-operations among nodes are expected since relevant data and information may be scattered over distributed nodes.

Figure 1 indicates that a peer can join several peer communities according to the topics she may be interested in. So long as peer supports JXTA communication protocol, she can instantly ask and join any other peer groups under some topics, even if such peer has already got involved in other peer groups. From the other side, if certain peer shares her resources to the public community, such as a collection bibliographic search can be shared, she can just send an advertisement of her expertises to the public. Generally, such expertises are a list of controlled vocabulary or ontology terms which can be employed in selecting appropriate peers. Section 2.2 will discuss the query procedure in a more detailed way.

2.2 Peer Selection

This section describes the query procedure in our prototype by adopting the JXTA protocol. Figure 2 illustrates the peer selection procedures when conducting queries.

Generally peer broadcasts queries just inside the peer group it belongs to. In Figure 2, a peer group is formed by four peers A, B, C and D. When a query is generated in A, peer A will firstly send out a broadcast to request for other peers' expertises; Any connected peer should respond with her expertise to A; then A matches those expertises with the conducted query automatically and then selects suitable peers to send the query to. Intuitively, such mechanism can relieve the communication load in the network considerably, since it avoids blind broadcasting to all connected but irrelevant peers. Consequently, it saves useless computing resources on peers which can not provide any suitable results. The advertisement is used as the primary tool for making general peer, peer group, and resource configuration information available to the network, peers, and peer groups. We can use XML, which is default format in JXTA, to pass advertisement from peer to peer, easily providing expandable and hierarchical representation of information which is compatible with any classification systems.

2.3 Peer Architecture

This section describes in detail the structure of an average peer shown by Figure 3.

Rule Engine: On the left side, the rule engine is illustrated and the inner components are indicated as well. The rule engine contains an inference engine, a rule base, and a working memory. The purpose of introducing rule engine is to nail down the implicit relationships among heterogeneous metadata terms. We integrate the JESS [16], one of the Java based rule engines, into our prototype. The benefits of adopting rule engine in semantic search will be argued in next section. The Rule Base contains all the rules the system knows. Practically, we present the mapping information in a set of rules and store them in the rule base. Meanwhile, ontologies, standard metadata information (i.e., Dublin Core [17]) and as well self defined (such as that in personal library) metadata elements must be referred to when mapping heterogeneous terms. The crucial part of the rule engine is obviously the inference engine, which controls the whole process of applying rules to working memory and generating corresponding results. The working memory, which is transparent to users, contains all pieces of premises and conclusions of rules we provide beforehand.

Local Resources (LOS): In order to simplify the problem, we assume that all collections are or can be formatted in XML. Yet in reality, many other collections may have different format, for instance, a travel agency may have the tourism information in the format of Excel, Word, PDF files or even emails. We will not consider these perspectives, though. This component stores the physical information to be shared in the P2P network. A peer can allocate self-defined tasks to its crawler to collect information on the Web, which is especially common in moderate-sized libraries or personal ones, where the users may collect distinct documents according to their own preferences.

Figure 3 Individual Peer Structure

Wrapper: We will not query over the XML collections in different metadata schemas directly because the variety in metadata standards prevents us to do so. Alternatively, we use wrapper or converter to translate varied metadata records to a common understandable/interpretable one. Dublin Core is the most widely used one in digital library community. It is rather simple and general, though. So some detailed and special elements in source metadata sets may not find the correspondents in Dublin Core. Such a result is inevitable in
translating a complex schema to a simple one if no extra maintenance procedures have been conducted. However, we can still achieve interoperability to some degree since major descriptive information on a digital object can be described in Dublin Core. Inside wrapper, specific crosswalk which is used for translation is given by peers. According to the Canadian Heritage Information Network [1], crosswalks are used to translate between different metadata element sets. The elements (or fields) in one metadata set are correlated with the elements of another metadata set that have the same or similar meanings. This is also sometimes called semantic mapping. Additionally, wrapper will also use the results extracted from rule engines, re-process XML collections in LOS and convert to RDF [19] documents in LPR. The RDF collection LPR is the real resource for searching and designed to support reasoning functionality in querying metadata records.

Local Peer Repository (LPR): The information searching procedure is to be carried out in LPR, which also supports browsing and navigating by local users. Strictly speaking, this component is not only a metadata annotated resource but also a hierarchical- and knowledge-based repository. In LPR, we adopt Sesame [15] for storing, querying and reasoning with RDF and RDF Schema. Supported by Sesame, a LPR can deduce over the terms in incoming queries and help retrieve the related metadata records, and return them to original peers. We argue our preference of adopting RDF/RDFS to format the metadata records as follows. First, compared with DAML+OIL [18] or OWL [20], RDF/RDFS allows greater flexibility, and this may be an advantage. Next, tools supporting RDF/RDFS is perhaps at this stage more mature. DAML+OIL and OWL allow definitions of specific constraints on the data model, and have more expressive power. Therefore, if there are more complicated requirements on such aspects or more powerful tools come to birth, we should migrate to OWL solutions for better precision.

User Interaction (UI): UI allows users to import, export local repositories. Besides, she can also import new crosswalks or edit available ones. Meanwhile, she can even select from connected peers to send the queries to. Communication between peers can be rendered automatically as illustrated by Service Publishing and Resource Requesting in Figure 3.

3 Operating Semantics – Walking Through Examples

The critical part in the whole system is obviously how to extract, express and exploit the semantic relationships in the heterogeneous metadata schemas. Such procedure is two layered. The anterior one is that structure and syntax heterogeneities should be removed; and the posterior one is to process and inference over the semantic dissimilarities. To demonstrate such process, we adopt working examples as follows.

3.1 Example: Semantic Markup

We use the term 'Semantic Markup' in this paper to clarify it from many interesting works in 'semantic annotation', which focuses on identifying the semantics of textual or multimedia streams. Herein, we concentrate on extracting the semantics embedded in document markup as well as making it 'processable. Markup semantics are defined in [11] as applying knowledge representation technologies to the problem of making the structures, relationships and properties explicit.

Let us consider the following fragment of an XML-tagged document from Financial Times (FT) Collection in TREC 4 [21].

```xml
<DOC>
  <DOCNO>FT911-376</DOCNO>
  <HEADLINE>FT 13 MAY 91/Survey of Cardiff(2):Selling on the road - The financial sector</HEADLINE>
  <BYLINE>By ANTHONY MORETON</BYLINE>
  <TEXT>Although the day-long event was one of a series that will ... (Omitted)</TEXT>
  <PUB>The Financial Times</PUB>
  <PAGE>London Page 16 Photograph The Bank of Wales was set up in 1972, and moved to its new building in September (Omitted).
  </PAGE>
</DOC>
```

From the literal in the fragment, it is moderate for readers to deduce the meaning of tag names. For instance, "DOCNO" can be roughly regarded as "Identifier" in the Dublin Core, "BYLINE" is subsumed by "Creator", and "PAGE" as "Relation". Two tags are complicated. One is the "HEADLINE" which combines the "Date" and "Title" information as are in Dublin Core. Since the "Date" segment is not standard, so we just simply assign it to "dc.title". "TEXT" is another tag which does not have correspondent in Dublin Core since it is context body, so we just remain it unchanged. We illustrate the crosswalk information between FT metadata and Dublin Core metadata as follows.

In Figure 4, the lower part illustrates the syntactic view of this XML fragment; and the relationships against Dublin Core are indicated by the dotted lines in the upper part. We have converted the mapping information into the following rules:

```
(defrule RuleFTDC
  (DCCollection ?d) => (Assertion (DocumentName ?d) (CollectionName "FT")))
```

Such rules can be deduced in inference engines, such as JESS. In JESS, we can indicate any explicable relationships in the rule base besides direct crosswalk. Take one example, we can present an inverse relationship which indicates that an author who has several publications in an XML fragment, may be the child node of publication in another XML fragment. It is due to that one author may have several publications and a publication should have at least one author's name.

```
(defrule RuleInverse
  "Indicating the inverse relationship between author and publication"
  (Tag (Document1Name ?d1) (TagName "DCIdentifier")
       (AuthorName ?authorname)) =>
  (assert (InverseAxiom (Document1Name ?d1) (DCIdentifier ?docno)
                       (DCCreator ?byline) (DCTitle ?headline) (DCRelation ?pagename))))
```

Figure 4. The Syntax Tree of Doc. 'FT911-376' and its Enhancement with Semantics
3.2 Expressing Semantics

With the XML results as inputs in LPR, we can transform them into local RDF repositories by XSLT or Java APIs. The query language to the local RDF repository is SeRQL (Sesame RDF Query Language) [28]. Actually, there are also many other RDF query languages, such as RQL [30], QEL [23], etc., but SeRQL addressed some practical requirements that were not sufficiently met by other query languages. We will not describe SeRQL’s capabilities in full detail, but refer readers to SeRQL’s manual (http://www.openrdf.org/doc/SeRQLmanual.html) for a complete overview.

Herein, we provide a RDF segment and take it for an example to present the way of expressing and deducing semantics.

```
<rdf:Description rdf:nodeID="A6">
  <dc:creator>Ashraf Aboulnaga, Surajit Chaudhuri</dc:creator>
  <title>Suggested additional metadata schema to improve a query workload that arises in applications, such as network monitoring and telecommunications...</title>
  <dc:source>Conferences on Information and Knowledge Management</dc:source>
  <dc:publisher>ACM Press New York, NY, USA</dc:publisher>
</rdf:Description>
```

Firstly, SeRQL creates the query in a style of ‘select-from-where’ or ‘construct-from-where’, where the select or construct clauses specify projections, the from clause specifies a graph match template, and the where clause allows the definition of additional boolean constraints on matched values in the path expressions. In our example, we can easily conduct the query over ‘dc:creator’ or ‘dc:title’ in such form.

Secondly, we can extract the implicit semantics from transitive relation, such as the ‘dc:references’ or ‘dc:isReferencedBy’ which are defined in Qualified Dublin Core element sets [29]. Finally, Dublin Core records in local LPR repository may be incomplete. SeRQL allows distinguishing between optional and required elements in the query.

4 Related Works

Information searching over P2P network is not a new scenario, but the semantic searching over such network is. The conventional P2P networked realized searching by doing string matching between the query and the file names in the peer’s repository. Consequently searches are restricted to strings that can be contained in a file name.

Perhaps the earliest metadata based information searching over P2P network is work done by Sumeeet Thadani [22]. [22] proposed a technique for allowing richer queries on metadata associated with files in the repository. Thus, requests will encode content richer queries and responses will contain results based on the rich query searches in addition to the regular results. The proposed scheme will ensure that the protocol continues to work with older clients, which do not understand the embedded rich queries. However, such approaches do not touch the heterogeneity issue in the metadata schemas.

Edutella [23] is an excellent work which provides an RDF-based infrastructure for exchanging metadata in P2P applications. The Edutella Query Services (QEL [23]) is intended to be a standardized query exchange mechanism for RDF metadata stored in distributed RDF repositories. The Edutella project focuses on the education community. [24] is an extended work based on Edutella platform, it adopts TRIPLE language[25] to express suitable rules and axioms, and extends the purely syntactic definition given in the Learning Objects Metadata (LOM) specifications by such axioms and inference rules so as to avoid redundant metadata annotation and derive additional metadata from existing ones as well. Haley et.al. considered the problem of schema mediation in a peer data management system (PDMS) – Piazza [31]. They proposed a flexible language for mediating between peer schemas, which extends known data integration formalisms to more complex architecture. They realized algorithms for chaining through multiple peer mappings in order to locate data relevant to a query. Such approach work well when the collections on varied peers are rich and similar, and there is still a problem in retrieving information from peers with small collections.

Our work differs in several ways. First, our focus is how to integrate the semantic reasoning functionality into the conventional information retrieval procedure against P2P network. Second, we use rule engine to do the reasoning over XML files by predefining the mapping rules in rule base. Third, local peer has more flexibility in creating its own rules and facts according to her preferences and characteristics, which is crucial to the library systems.

Lakshmanan [26] proposes to enrich local sources with semantic declarations so as to enable interoperability. Rather than determining a commonly agreed upon global schema (Global as View - GAV) or translating schemas among local schemas (Local as View - LAV), the approach adopts these declarations to expose the semantics of the information content of sources by mapping the concepts present therein to a common and application specific vocabulary. All participating sources must map its data to the standard ontologies that already exists, though.

5 Conclusions and Future Work

In this paper, we presented a framework for semantic data integration against P2P network where many heterogeneous metadata schemas exist. We discuss the general P2P architecture, especially under the digital library application scenario. A detailed structure of a specific peer has been described. We assumed that all of the collections in the P2P network are XML formatted. A proposed mechanism in dynamically expressing the implicit relationships in RDF is presented and a RDF query language – SeRQL is used to conduct the searching. In the future work, we are going to implement an adaptive information retrieval method which can modulate the searching according to the results generated from the rule engine. Likewise, consideration on the query features of general P2P networks is highly required in order to couple with the IR component. Many researches have recently focused on tackling some challenges in P2P network with respects to autonomous pooling resources (e.g., files, storage, and compute cycles), high availability and fault-tolerance, and self-organization of peers [27]. We recognize such strong needs for some researches, but the development of such tasks is not part of our work. Another critical task in our future work is to evaluate the approach of rule-based information searching in P2P network. The whole and concrete evaluation will be carried out by comparing with simple keyword-based information searching in P2P networks.

Acknowledgements

The authors would like to express their gratitudes to the anonymous reviewer for their invaluable comments and suggestions about prior versions of this paper. This work was funded in part by Norwegian Research Council grants WEB-TEK IKT2000.
References

Journal of Digital Information Management  Issue: Volume 3 Number 2  June 2005 75