oreChem ChemXSeer: A Semantic Digital Library for Chemistry

Na Li*, Leilei Zhu*, Prasenjit Mitra*, Karl Mueller†, Eric Poweleit†, C. Lee Giles*
*College of Information Sciences and Technology †Department of Chemistry
The Pennsylvania State University
University Park, PA 16802, USA
{nzl116, luz113, pmitra, ktm2, etp113, clg20}@psu.edu

ABSTRACT

Representing the semantics of unstructured scientific publications will certainly facilitate access and search and hopefully lead to new discoveries. However, current digital libraries are usually limited to classic flat structured metadata even for scientific publications that potentially contain rich semantic metadata. In addition, how to search the scientific literature of linked semantic metadata is an open problem. We have developed a semantic digital library oreChem Chem$_x$Seer that models chemistry papers with semantic metadata. It stores and indexes extracted metadata from a chemistry paper repository Chem$_x$Seer using "compound objects". We use the Open Archives Initiative Object Reuse and Exchange (OAI-ORE))[1] standard to define a compound object that aggregates metadata fields related to a digital object. Aggregated metadata can be managed and retrieved easily as one unit resulting in improved ease-of-use and has the potential to improve the semantic interpretation of shared data. We show how metadata can be extracted from documents and aggregated using OAI-ORE. ORE objects are created on demand; thus, we are able to search for a set of linked metadata with one query. We were also able to model new types of metadata easily. For example, chemists are especially interested in finding information related to experiments in documents. We show how paragraphs containing experiment information in chemistry papers can be extracted and tagged based on a chemistry ontology with 470 classes, and then represented in ORE along with other document-related metadata. Our algorithm uses a classifier with features that are words that are typically only used to describe experiments, such as "apparatus", "prepare", etc. Using a dataset comprised of documents from the Royal Society of Chemistry digital library, we show that the our proposed method performs well in extracting experiment-related paragraphs from chemistry documents.

[1]http://www.openarchives.org/ore/

JCDL'10, June 21–25, 2010, Gold Coast, Queensland, Australia.

Categories and Subject Descriptors

D.2.12 [**Software Engineering**]: Interoperability—*Data mapping,Distributed objects*; E.2 [**Data Storage Representations**]: Linked representations; H.3.7 [**Information Storage and Retrieval**]: Digital Library—*Collection,Dissemination*

General Terms

Design, Experimentation, Management

Keywords

Digital library, OAI-ORE, metadata extraction, semantic web, Support Vector Machines, Chem$_x$Seer SeerSuite

1. INTRODUCTION

Easy and fast access to scientific artifacts is important as the amount of scientific literature continues to increase. Vertical search engines such as Google Scholar[2] provide efficient access to scientific papers based on user provided keywords. Digital libraries such as the ACM digital library[3], Libra[4] and Chem$_x$Seer[5], take advantage of well-defined metadata from catalogues, taxonomies and domain-specific information in scientific papers. Nevertheless, researchers find difficulties in retrieving desired content from a large retrieved document set.

Metadata are a set of controlled vocabularies stored in databases with a fixed schema. However, a flat metadata structure does not properly represent the rich information structure in scientific publications. Fig. 1 illustrates an example of a chemistry paper linked to a set of relevant heterogeneous information. A chemistry paper has several important features and attributes, e.g., authors, referenced papers, figures, tables, chemical formulae, experimental sections, etc. An author may have other papers and may have a homepage with the name, title, affiliations, contact information, research interests, etc. A chemical formula may be described using a graphical chemical structure or may be described by its name. An experimental section can describe instruments, chemical formulae, and other important information. Therefore, modeling unstructured textual information with a set of properly structured metadata is crucial for rich and efficient information access and knowledge aggregation.

[2]http://scholar.google.com/
[3]http://portal.acm.org/portal.cfm
[4]http://libra.msra.cn/
[5]http://chemxseer.ist.psu.edu/

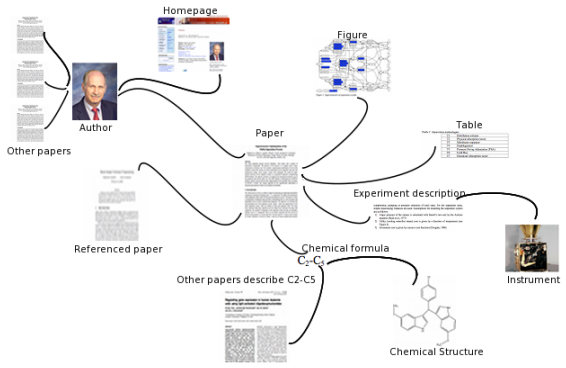

Figure 1: Example of the linked information structure of a scientific publication.

We present a semantic digital library oreChem Chem$_x$Seer[6] that intends to model chemistry papers with semantic metadata and semantic relations. In our system, we use a new data model, the Object Reuse and Exchange (OAI-ORE)l[7] proposed by the Open Archives Initiative (OAI) [9, 5] to aggregate metadata related to documents in our digital library. The aggregation of the set of sources are called "compound objects". The compound objects represent information across cooperating digital repositories, registries and services. The ORE model is a graph model based on Named Graphs [4] that are extensions of RDF graphs. Named Graphs consist of nodes and arcs within a node set. When applied to compound objects, the nodes correspond to related resources; and the arcs correspond to typed relations. Because the nodes and arcs are stored as RDF triples with unique URIs, these resources can be identified and referenced unambiguously through the URIs.

An example of a typical compound object in the ORE model is similar to the graph in Fig. 1. In that example, the chemistry paper is an aggregation. External objects such as referenced papers, authors' homepages, and internal objects such as figures, tables, formulae, data, etc. are aggregated. The same data can be modeled in different ways.

The advantage of ORE is that a set of related information can be searched or operated on as one object. Users can create these aggregates and share them both with other users as well as across their own user sessions. Users who wish to save or print a set of related pages, an individual paper, a presentation of that paper, and all documents referenced by that document can print all of them together simply by parsing an ORE file [8].

Our digital library oreChem Chem$_x$Seer is built on top of the Chem$_x$Seer system. Documents in the Chem$_x$Seer digital library and their associated metadata are packaged as ORE objects. The metadata of each document in the Chem$_x$Seer repository was obtained from the Chem$_x$Seer system that, in turn, used (i) the SeerSuite system [17] to extract the metadata automatically and (ii) scraped the metadata from webpages associated with documents in digital libraries such as the RSC digital library[9]. We store the metadata associated with each document in a RDF repository. The ORE objects are generated on demand and the RDF graph is then visualized. End-users can then directly access these ORE objects.

As a case study, we explore how to extract and integrate new types of data into oreChem Chem$_x$Seer metadata aggregates along with document-related metadata. We propose a method to identify paragraphs from documents describing experiments or containing information related to how the experiments were conducted. Once the paragraphs containing experiments have been identified, then end-users could extract detailed information related to experiments, e.g., methods describing how the experiments were performed, the reactants used, etc. Such detailed information extraction will enable end-users to search for documents based on the type of experiment, reagents used, etc. Work on automated experiment paragraph extraction is a first step toward that goal.

Identifying experiment paragraphs is a nontrivial problem since, despite advances in natural language processing, understanding the semantics of text and resolving the inherent ambiguity automatically is still an unsolved problem. Determining which parts of a document talk about experiments and which do not is hard in the absence of a good semantic understanding of the document, which, in turn, is hard to achieve using automated methods. The experiment paragraph tagging module uses a supervised machine-learning algorithm with training sets generated by domain experts and uses single word features like "apparatus", "reagents", etc. to identify experiment paragraphs. The experiment detection module performs the tagging with high accuracy.

Our system for oreChem Chem$_x$Seer has been made available[10]. Users can navigate ORE objects and examine the documents and their associated attributes.

Our contributions are listed as follows. First, to the best of our knowledge, we are the first to establish a platform using the ORE data model to manage automatically extracted data from scientific publications. Second, we have designed a faceted interactive interface powered by the ORE infrastructure. Our interface provides faceted navigation across a large set of metadata so that chemists can navigate the digital library and examine the literature with added ease. Third, we have proposed and demonstrated an ontology-based supervised machine learning method for extracting important experimental information from chemistry papers.

The rest of the paper is organized as follows. In Section 2, we introduce related work on semantic digital libraries. In section 3, we present the system infrastructure of oreChem Chem$_x$Seer. In Section 4, we propose an ontology based experimental information extraction approach using support vector machines (SVMs), and give our experimental results. We conclude and propose future work for experimental information extraction in Section 5.

2. RELATED WORK

To the best of our knowledge, oreChem Chem$_x$Seer is the first system that provides end-users the capability to share aggregated metadata. In this section, we briefly provide an overview of work related to semantic digital libraries.

ScholOnto [15] uses ontologies to model relationships among research documents and enriched the citation relationship with an ontology called "Claim". In "Claim", a document

[6]http://130.203.146.147:8081/oreChem/
[7]http://www.openarchives.org/ore/
[8]http://www.openarchives.org/ore/1.0/primer
[9]http://www.rsc.org/
[10]http://www.cxs03.ist.psu.edu:8890/

can have many relationships with other documents in the literature, e.g, a document can be *an example of* another document, can be *inconsistent with* another document, can *entend* the content of another document, etc.

JeromeDL [14] used a collection of ontologies to model different aspects of bibliographic information. In JeromeDL, not only general ontologies, like Dublin Core[11] and foaf[12] were used, but also event-based ontologies were included with relationships such as "isReviewed", "hasSubmissionStatus", "isUploadedBy", etc. JeromeDL used an interactive interface for publishers and creators to annotate pieces of data with the provided ontologies during the upload process.

Greenstone 3 [19] proposed how ontologies can be fully integrated into digital libraries. Greenstone used the FRBR [2] framework to model data. FRBR uses four entities: works, expressions, manifestations and items. With the four core entities and other attributes for expressing the identities of entities, FRBR is able to model data types, relationships among them, and data sources from different repositories. FRBR was first used for data ingestion, where the data were annotated with the FRBR vocabulary and indexed in the system. Greenstone 3 supports typed search using FRBR resulting in more accurate results than when FRBR was not used.

The Fedora [10] data model is used in the National Science Digital Library. The Fedora data model is most similar to the data model that we use. Fedora is also a graph-based data model for exposing a repository as a network of objects. It is also flexible in that it allows overlaying statements from multiple ontologies. Another common feature of Fedora and ORE is that both enable fine-grained digital objects accessible through an architecture of remixed data sources and transformations. However, ORE is better than Fedora in two aspects: (i) ORE takes a resource-centric view that defines clear logical boundaries between resources and enhances the interoperability of information, and, (ii) ORE provides a standard for identifying web services and agents through resource maps[13] with which information can be easily collected across different repositories without losing their provenance. ORE grants the power to users to easily choose their preferred repositories and services.

Finally, digital library search engines such as CiteSeerX and others automatically extract OAI metadata and other metadata resources such as citations. With the exception of the CiteSeerX model, populating ontologies with data and metadata has been left to publishers, creators and authors and is often a manual process. This work attempts to automatically extract general data and metadata with a particular focus on chemistry.

3. SYSTEM INFRASTRUCTURE

oreChem Chem$_x$Seer is built on a related project Chem$_x$Seer. Chem$_x$Seer had access to over 130,000 articles from the Royal Society of Chemistry repository[14]. Not only does Chem$_x$Seer index such documents, but, using the SeerSuite software (from which CiteSeerX is built), it also automatically extracts and indexes bibliographic data, tables, figures and chemical formulae, which can be represented in oreChem Chem$_x$Seer as well.

3.1 ORE Data Model

The ORE abstract data model consists of four entities: *aggregation*, *aggregated resouce*, *resouce map*, and *proxy*. An *aggregation* is a resource of type *ore:Aggregation* that is a set of other resources. An *aggregated resource* is a resource that is a constituent of an *aggregation*. A *resource map* describes a single *aggregation*, and an *aggregation* can have many *resource maps*. A *resource map* retains provenance information of the constituents described in an *aggregation*.

We implement an ORE data model that underlies the oreChem Chem$_x$Seer system. Fig.2 illustrates an example of a compound object with *aggregation*, *aggregated resources*, and *resource map* in oreChem Chem$_x$Seer. The triples producing the *resource map* in Fig. 2 are serialized in RDF/XML as follows:

```
<rdf:Description rdf:about="
 http://chemxseer.ist.psu.edu/
   rem/rdf/document/10.1039/b402145m">
  <dcterms:modified rdf:datatype="
   http://www.w3.org/2001/
   XMLSchema#date">2009-08-25T11:08:26-0400
  </dcterms:modified>
  <dc:creator rdf:nodeID="A5"/>
  <ore:describes rdf:resource="
   http://chemxseer.ist.psu.edu/
   document/10.1039/b402145m"/>
  ...
  <rdf:Description rdf:nodeID="A5">
  <rdf:type rdf:resource="
   http://purl.org/dc/terms/Agent"/>
  <foaf:name rdf:datatype="
   http://www.w3.org/2001/
   XMLSchema#string">OreChem ChemXSeer
  </foaf:name>
</rdf:Description>
```

The *resource map* describes a document aggregation in this example. As illustrated in Fig. 2, the document aggregation contains a set of resources that describe and support it. Ontological metadata are used to describe the information associated with each resource. Several existing ontologies are reused in the oreChem Chem$_x$Seer model, for example, the foaf ontology is used to manage author information, Dublin Core is used to manage general publishing information, and the ChemAxiom Metrology ontology[15] is used to manage experiment information. However, we expand this document ORE model further with a set of vocabularies particular to chemistry such as chemical formulae, tables, and figures that are aggregated along with the metadata of the documents. Currently, we have defined 12 classes of 27 concepts and 10 relationships.

The following RDF/XML code shows an example of how we model chemical formulae in the oreChem Chem$_x$Seer ORE data model. Each *rdf:Description* represents either a primary resource or a secondary resource. The secondary resource formula *6-propionyl-2-(N,N dimethylamino)naphthalene* is an aggregated resource denoted by *ore:aggregates* for the document *aggregation* named b402145m, which is a primary

[11]http://dublincore.org/
[12]http://www.foaf-project.org/
[13]http://www.openarchives.org/ore/1.0/primer
[14]http://www.rsc.org/
[15]http://bitbucket.org/na303/chemaxiommetrology/

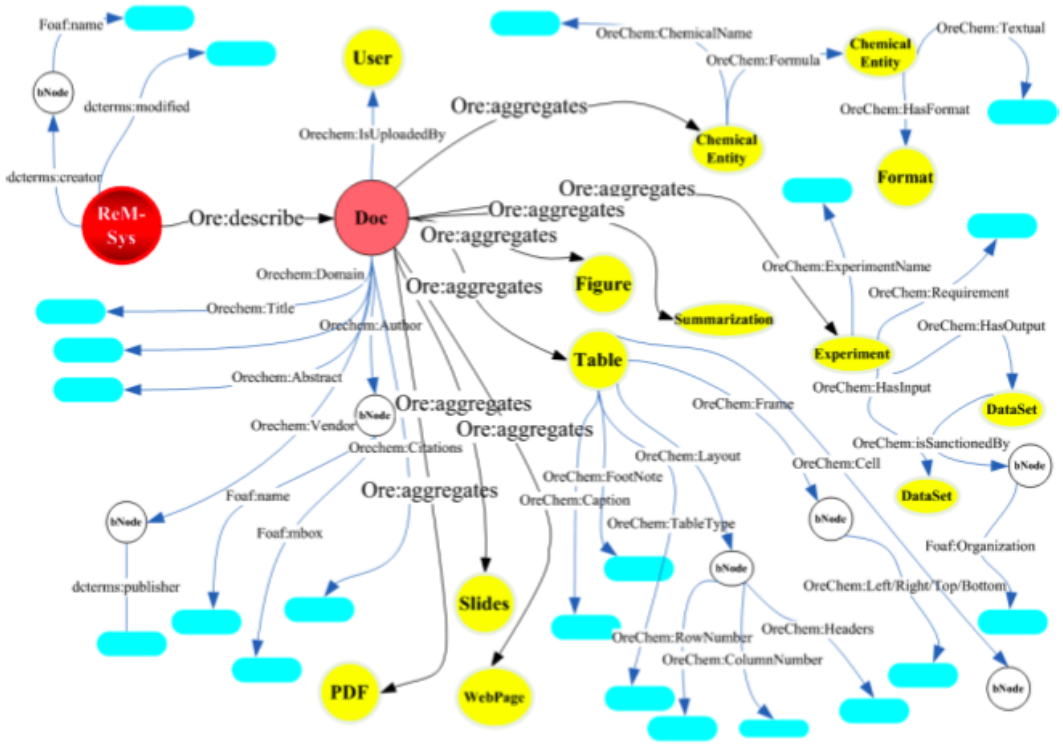

Figure 2: The comprehensive ORE data model underlying oreChem Chem$_x$Seer

resource. There is only one property in the chemical formula resource: *formula name*.

It is very easy to add more properties to the resource by adding additional triples within the corresponding *rdf:Description* tag.

```
<rdf:Description rdf:about="
 http://chemxseer.ist.psu.edu/
   document/10.1039/b402145m">
  <ore:aggregates rdf:resource="
   http://chemxseer.ist.psu.edu/
                formula/6-propionyl-2-(N,
                N-dimethylamino)naphthalene"/>
  <orechem:contains rdf:nodeID="A18"/>
</rdf:Description>
<rdf:Description rdf:nodeID="A18">
  <dcterms:title>6-propionyl-2-(N,
              N-dimethylamino)naphthalene
  </dcterms:title>
  <rdfs:seeAlso rdf:resource="
   http://chemxseer.ist.psu.edu/
       formula/6-propionyl-2-(N,
              N-dimethylamino)naphthalene"/>
</rdf:Description>
```

This more general model used in oreChem Chem$_x$Seer includes many of the standard metadata features of current digital scientific publications. The metadata is serialized in RDF/XML and stored in a triple store. In the next section, we will illustrate how this is integrated into the oreChem Chem$_x$Seer architecture.

3.2 System Architecture and Implementation

Fig. 3 shows the system architecture of oreChem Chem$_x$ Seer. On the front end, a user can input typed queries, e.g., "author:Peter keyword:carbon venueyear:Analyst-2004". Then, the system returns a list of document compound objects. At the back end, the ORE data model is built on Chem$_x$ Seer's data including bibliographic data, table/figure data, formula names, statistics and other information. Specifically, RDF triples are generated from these data using the ORE data model's vocabulary that we defined. Each triple is then populated into a Sesame native RDF store [16]. Note that Sesame supports "Named Graph"s by providing a field called "context" besides the "subject," "object" and "predicate" fields for each triple. Therefore, we can easily construct a document aggregation while populating the data by putting a document aggregation URI into the "context" field for each triple. In addition, the Sesame native RDF store will build three triple indexes *spoc*, *posc*, and *cspo*, where *s* denotes subject, *p* denotes predicate, *o* denotes object and *c* denotes context. After the database is initialized and populated, we can query any triple or document aggregation using the SeRQL (Sesame RDF Query Language) [17] query language. When a user sends a query to the server, the query is first converted to a SeRQL query, and then a set of constructed RDF subgraphs (termed "graphs") are returned from the Sesame database. The system then sends the output of Sesame to Foresite [18], a toolkit for constructing and serializing ORE resource maps into different formats. Fore-

[16] http://www.openrdf.org/
[17] http://www.openrdf.org/doc/sesame/users/ch06.html
[18] http://code.google.com/p/foresite-toolkit/

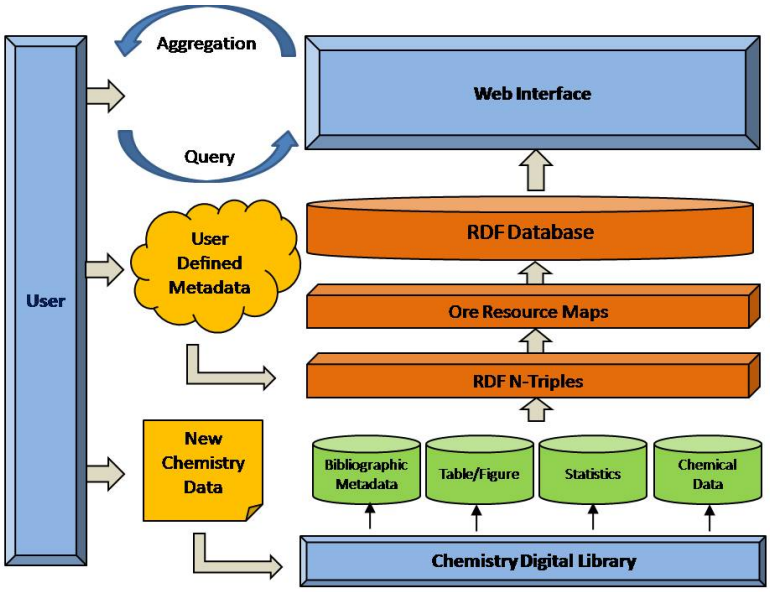

Figure 3: oreChem Chem$_x$Seer architecture

site will parse the RDF files and serialize them into ORE files in six formats: ATOM, RDF/XML, N3, N-Triples, Turtle and RDFa. We also used a new format named SVG for visualizations. The SVG ORE files are transformed from RDFa files through XSLT. An SVG ORE file is a searchable interactive graph showing metadata and their relationships. We describe how the new format helps represent oreChem Chem$_x$Seer's data in the next section.

3.3 oreChem Chem$_x$Seer User Interfaces

In this subsection, we show a working example of oreChem Chem$_x$Seer. We give a general idea of major functions of oreChem Chem$_x$Seer. The technologies we use to implement the system include JSF, JSP, Sesame, JavaBeans, XSLT and SVG. The bibliographical metadata extraction module is derived from SeerSuite[19], and the chemical metadata extraction module is derived from Chem$_x$Seer [16].

Fig. 4 shows the portal of oreChem Chem$_x$Seer. Note that we only present the screen showing how to search a document aggregation. Other views showing additional ORE objects, e.g., experiment aggregation will be added in the future. In this example, a user wants to search document aggregations that contains "Carbon" in the document title, and that was written by "Mike" in "1999".

Fig. 5 shows the result page corresponding to the above query. Users can collaborate and help maintain the records by clicking the "Modify" link. Fig. 6 shows fields that we currently allow the user to modify.

Next, we show multi-faceted views of oreChem Chem$_x$Seer. Faceted navigation is thought to be a powerful tool that reduces the mental work of searching an information collection by promoting recognition over recall and suggesting logical but perhaps unexpected alternatives. Furthermore, it also acts as important scaffolding for exploration and discovery, while seamlessly integrating free text search within the category structure. [7].

[19]http://sourceforge.net/projects/citeseerx/

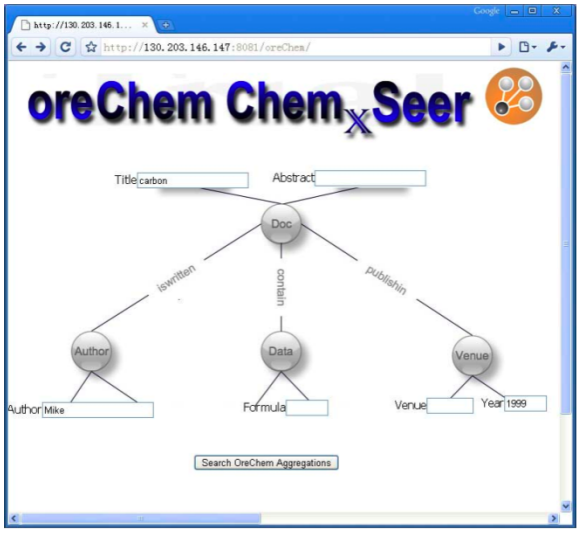

Figure 4: oreChem Chem$_x$Seer portal

Fig. 7 shows a hyperlink-based multi-faceted view for a document aggregation. Each indexed data is clickable and searchable. If a user wants to learn a set of resources related to the formula "aluminium", the user can simply click on the hyperlink "aluminium" to get a page like Fig. 8. Or if a user wants to learn a set of resources related to the author "Michael Thompson", the user can click on the hyperlink "Michael Thompson" to get a page such as Fig. 9. Constructing these different views is simple using named graphs in Sesame.

Fig. 10 is a graphical multifaceted view for a document aggregation. The colored nodes represent available resources in a document aggregation. The nodes are expandable and clickable, so that users can easily interact with the graph.

We have shown seven snapshots of our system. These

Figure 5: A result page from oreChem Chem$_x$Seer

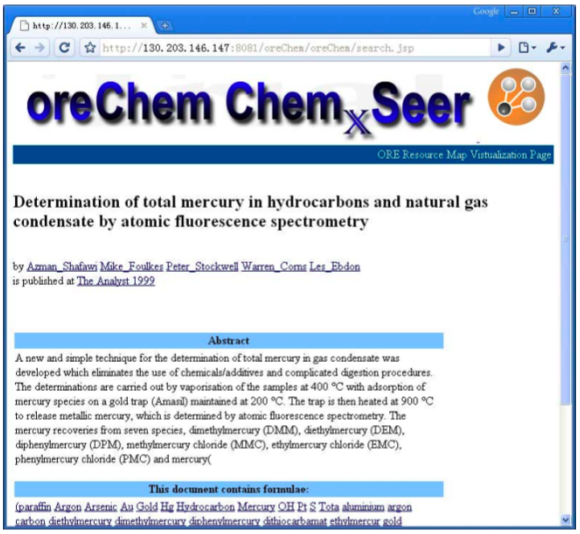

Figure 7: Document view of oreChem Chem$_x$Seer

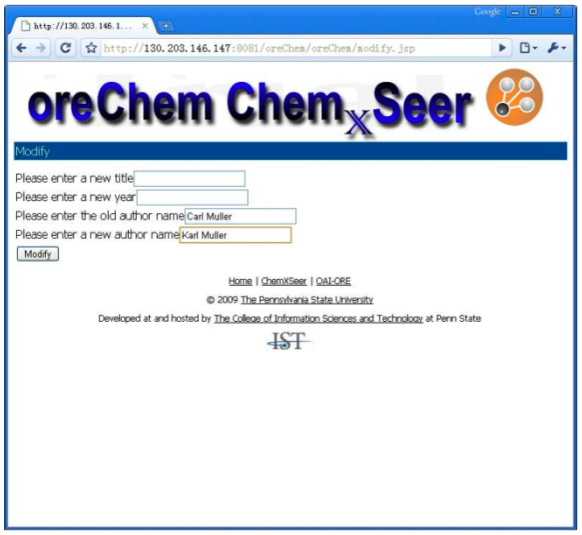

Figure 6: A modify page from oreChem Chem$_x$Seer

Figure 8: Formula view of oreChem Chem$_x$Seer

seven scenarios highlights our system: a fully functioned semantic digital library that supports faceted search and navigation, user interaction, and seven document formats for downloading.

4. EXPERIMENT INFORMATION EXTRACTION

An important part of documents related to empirical chemistry and other empirical sciences is the experiment section that reports methods and materials, observations, etc. Ideally, we want one (or more) "experiment" ORE object to be associated with a chemistry document that provides information about the experiments reported in the document. In order to do this, we investigate the task of extracting paragraphs describing experiments in chemistry documents. To the best of our knowledge, extracting such information automatically has not been previously investigated. We use a machine-learning-based approach to identify experiment-related paragraphs in documents. We tested our methods on chemistry papers from various journals, but it can be easily extended to other domains.

4.1 Problem Formalization

Generally speaking, experiment-related information consists of sentences or paragraphs that describe components of an experiment; they could describe the experimental environment, experimental data, experimental procedure and experimental results. In experimental Chemistry papers, researchers often describe an experiment in the following way: a description of instruments used such as model, instrument characteristics, and instrument calibration. They also describe how reagents are prepared. Researchers then describe experimental procedure and reactions observed during the experiment. Finally, they discuss results and compare results from different conditions. Accordingly, the problem is how to automatically detect and extract such information.

Figure 9: Author view of oreChem Chem$_x$Seer

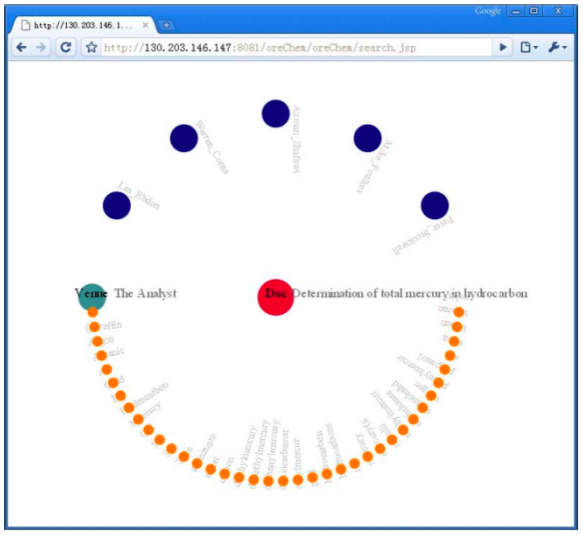

Figure 10: Graphical view of oreChem Chem$_x$Seer

In this paper, we will focus on how to extract experiment-related paragraphs, because we have empirically found that information about experiments are usually organized closely in consecutive paragraphs and non-experiment information seldom occurs in such paragraphs in chemistry papers. Taking each paragraph as an instance, which we denote as p_i, each paragraph is either related to experiment or not, which we denote as E or NE respectively. For each paragraph, there is a set of features $\{f_{ij}|j = 1 \ldots n\}$, where i denotes which paragraph the feature belongs to and n denotes the number of features.

The problem is then reduced to a classification problem where we want to classify the instances into two categories, either E or NE. We use Support Vector Machines(SVMs) [3]; they have been widely used for classification tasks. However, other classifiers could also be explored.

4.2 Classification Model

4.2.1 *Support Vector Machines*

SVM is a binary supervised learning method used for classification. For learning, the input are two sets of instances represented by feature vectors in a n-dimensional space and classification labels corresponding to the instances. They can be represented as $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, in which x_i denotes an instance (a feature vector) and $y_i \in \{-1, +1\}$ denotes a classification label. An SVM will try to find an optimal separating hyper-plane in the n-dimensional space that maximally separates the two classes of training instances (more precisely, maximizes the margin between the two classes of instances).

We use LibSVM [20], a library for support vector machines. We choose the RBF (Radial Basis Function) kernel, because our preliminary experimental results show that it works best for the current task. There are two parameters while using RBF kernels: C and γ. We use a "grid-search" using cross-validation to find the optimal C and γ as recommended in LibSVM guide [21].

4.2.2 *Feature Sets*

Appropriate selection of features is crucial to good classification. We emphasize the use of ontology concepts as features. Our features can be classified into two categories: Keyword Features and Concept Features.

Keyword Feature: These features include representative words that occur frequently in experiment-related paragraphs while seldom occurring in other paragraphs. They are "procedure", "experiment", "apparatus", "reagent", "react", "prepare", "instrument" and "calibrate". We observed that these keywords always appeared in the subtitle and the body of an experiment section of papers in our corpus.

Concept Features: These features represent concepts that are often used in chemistry experiments and are seldom used in other sections in a paper. We choose the ChemAxiom Metrology ontology[22], which describes the concept and relations of named techniques and instruments in chemistry experiments. It contains 470 classes in total.

Our intuition for using ontologies as features is that an ontology that tries to model and describe concepts and relations for a particular domain would well represent the features of that domain. In addition, the ORE data model easily represents and integrates ontologies. Thus, ontology-based extraction, ontology-based data management and ontology-based navigation can be readily synthesized into one system.

4.3 Experiments and Results

4.3.1 *Data Set*

We collect experiment-related paragraphs from chemistry papers in PDF formats. We chose 174 experiment paragraphs and 820 non-experiment paragraphs from three journals: the *Analyst*, *Organic & Biomolecular Chemistry* and *Chemical Communications*, as our training data. The first two journals have rigid format requirements such that subtitles should proceed each section. The later one does not

[20]http://www.csie.ntu.edu.tw/ cjlin/libsvm/
[21]http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf
[22]http://bitbucket.org/na303/chemaxiommetrology

have such requirements and, as a consequence, most subtitles are missing. Structure-wise, papers in *Chemical Communications* are not as easy to parse as the other two. In addition, we chose 85 experiment paragraphs, and 662 non-experiment paragraphs for our experiment as our test data. Some pre-processing is done before extraction. We convert the papers from PDF to plain text using PDFBox[23]. Then, we automatically detect paragraph boundaries, and mark them. Then, chemists annotated experiment-related paragraphs for all papers.

4.3.2 Performance of Experiment Extraction

We conduct a 5-fold cross validation to evaluate the performance of the SVM classifier. We also used several rule-based approaches as baseline methods. We developed a simple rule-based approach to assign a positive label to a paragraph if the paragraph contains at least k keywords. The keyword set is the same as the keyword feature set in the SVM classifier. We measure the performance of both the SVM and rule-based methods in terms of precision, recall and F1-measure. Given the number of the correctly-labeled true experiment-related paragraphs A, the number of true experiment-related paragraphs but mis-labeled as non-experiment paragraphs B, and the number of non-experiment paragraphs that are mis-labeled as experiment-related paragraphs C, we can derive: the *Precision* is $\frac{A}{A+C}$, the *Recall* is $\frac{A}{A+B}$, and the $F-measure$ is $\frac{2\times Recall\times Precision}{Recall+Precision}$.

Table 1 shows the performance measurements of both SVM and rule-based methods. For rule-based methods, we vary k from one to eight.

Methods	Precision	Recall	F-measure
SVM	**83.9%**	**85.9%**	**84.9%**
RL(k=1)	24.9%	78.8%	37.9%
RL(k=2)	29.9%	50.6%	37.6%
RL(k=3)	30.2%	30.6%	30.4%
RL(k=4)	29.5%	15.3%	20.2%
RL(k=5)	36.4%	9.4%	15.0%
RL(k=6)	31.3%	5.9%	9.9%
RL(k=7)	40%	4.7%	8.4%
RL(k=8)	28.6%	2.3%	4.3%
RL(k=9)	6.7%	2.4%	4.5%

Table 1: Performance measurements of rule-based baseline methods and SVM classifier

Figure 11 and Figure 12 show the precision and F-measure of rule-based methods respectively. From Figure 11, we can see that the best performance is when k is equal to seven. When k is greater than seven, the performance quickly gets worse. From Figure 12 we can see that it is almost a monotonically declining value; when k gets larger, the F-measure gets smaller. Therefore, the performance is best when k is equal to one.

Comparing the performance measurements of the rule-based methods when k is equal to one and seven with the SVM classifier in table 1, we can conclude that SVM classifier gets the best performance in terms of precision, recall and F-measure, and significantly outperforms the rule-based methods at least for feature set and paragraphs used.

[23]http://incubator.apache.org/pdfbox/

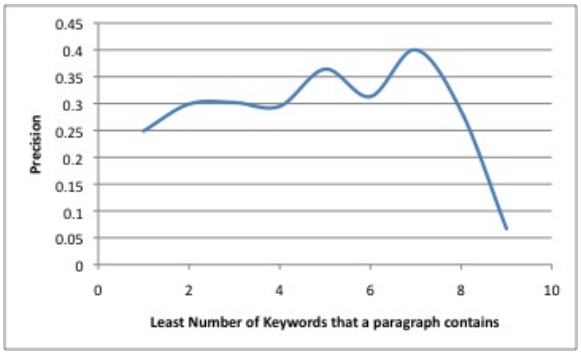

Figure 11: Rule-based methods precision plot

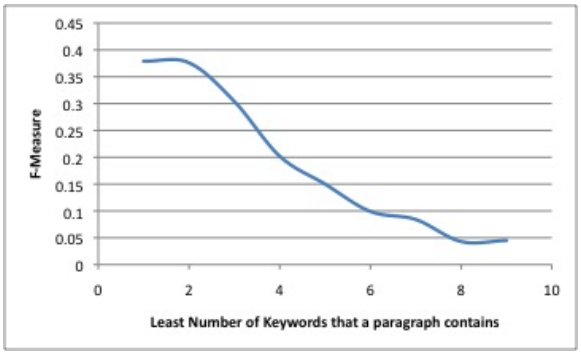

Figure 12: Rule-based methods F-measure plot

4.3.3 Discussion

Further investigation of the true negatives and false positives leads us to three reasons for errors in classification:

1. The feature sets as specified may not be complete. We could potentially improve the performance further by adding missing features. We could expend the ChemAxiomMetrology.owl ontology with another two important classes. One would be a subclass of *Instrument*, which we denote as "general instrument" (such as flask); the other should be a subclass of process, which we denote as "general process" (such as distillation and separation).

2. Tables and Figures are identified as non-experiment paragraphs. Actually some tables or figures belong to the experiment section. However, because we have not extracted the content of the tables or figures and their captions yet, we annotate them as non-experiment paragraphs. This can be fixed by performing table extraction [11] and figure extraction [8] in the future.

3. Information from other paragraphs have been wrongly classified, mostly from the Discussion/Analysis sections. There are sentences that briefly discuss different results obtained under different conditions in these sections. Our chemists annotate those paragraphs as non-experiment paragraphs, because most of the content in those paragraphs are not related to describing experiments. This observation indicates that we may need a finer grained experiment extractor, one that performs the classification at the sentence level.

In the future, we will vary our feature sets and evaluate the results to find an optimal feature set for the task. We may also need to explore sentence-level experiment extraction or step-wise experiment extraction methods to compliment the current experiment-extraction methods.

5. CONCLUSION

Automatically transforming unstructured scientific literature to structured knowledge is a nontrivial task. We have proposed a system oreChem Chem$_x$Seer based on the OAI-ORE data model. We show that our system is capable of automatically populating a chemistry ontology of metadata consisting of authors, scientific papers, chemical formulae, data and others. We have also defined and explored the problem of experimental information extraction. Using Support Vector Machines and chemistry ontologies, we have been able to extract experiment-related paragraphs from PDF documents. Experiments show that our approach is able to extract most experiment-related paragraphs from various academic chemistry papers. Future research can take many directions. For example, we can add a table extraction module and figure extraction module to the system. We can explore a sentence-level related experiment extraction method and step-wise experiment information extraction methods.

6. ACKNOWLEDGEMENTS

We gratefully acknowledge support by Microsoft Corporation and the NSF funded Chem$_x$Seer project (Grant No. 0535656). We also acknowledge useful discussions with Nico Adams, S.J. Coles, Jim Downing, J.G. Frey, Carl Lagoze, Peter Murray-Rust, and Marion Pierce.

7. REFERENCES

[1] D. Banville. Mining chemical structural information from the drug literature. *Drug Discovery Today*, 11(1-2):35–42, January 2006.

[2] G. Buchanan. Frbr: enriching and integrating digital libraries. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 260–269, New York, NY, USA, 2006. ACM.

[3] C. J. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[4] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Named graphs, provenance and trust. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 613–622, New York, NY, USA, 2005. ACM.

[5] H. V. de Sompel, C. Lagoze, M. L. Nelson, S. Warner, R. Sanderson, and P. Johnston. Adding escience assets to the data web. *CoRR*, abs/0906.2135, 2009.

[6] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 37–48, Washington, DC, USA, 2003. IEEE Computer Society.

[7] M. A. Hearst and E. Stoica. Nlp support for faceted navigation in scholarly collections. In *2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 62–70, 2009.

[8] S. Kataria, W. Browuer, P. Mitra, and C. L. Giles. Automatic extraction of data points and text blocks from 2-dimensional plots in digital documents. In *AAAI'08: Proceedings of the 23rd national conference on Artificial intelligence*, pages 1169–1174. AAAI Press, 2008.

[9] C. Lagoze, H. V. de Sompel, M. L. Nelson, S. Warner, R. Sanderson, and P. Johnston. Object re-use and exchange: A resource-centric approach. *CoRR*, abs/0804.2273, 2008.

[10] C. Lagoze, S. Payette, E. Shin, and C. Wilper. Fedora: an architecture for complex objects and their relationships. *Lecture Notes in Computer Science*, 6(2):124–138, 2006.

[11] Y. Liu, P. Mitra, C. L. Giles, and K. Bai. Automatic extraction of table metadata from digital documents. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 339–340, New York, NY, USA, 2006. ACM.

[12] V. Monev. Introduction to similarity searching in chemistry. institute of organic chemistry. In *Bulgarian Academy of Sciences, Sofia 1113, Bulgaria. Match-Communications in Mathematical and in Computer Chemistry 51*, pages 7–38, 2004.

[13] P. Murray-rust, H. S. Rzepa, and M. Wright. Development of chemical markup language (cml) as a system for handling complex chemical content. *New J. Chem*, 25:618–634, 2001.

[14] L. Z. Sebastian Ryszard Kruk, Stefan Decker. Jeromedl - adding semantic web technologies to digital libraries. *Lecture Notes in Computer Science*, 3588:716–725, 2005.

[15] S. B. Shum, E. Motta, and J. Domingue. Scholonto: An ontology-based digital library server for research documents and discourse. *International Journal on Digital Libraries*, 3:237–248, 2000.

[16] B. Sun, P. Mitra, and C. L. Giles. Mining, indexing, and searching for textual chemical molecule information on the web. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 735–744, New York, NY, USA, 2008. ACM.

[17] P. B. Teregowda, I. G. Councill, J. P. F. R., M. Kasbha, S. Zheng, and C. L. Giles. Seersuite: Developing a scalable and reliable application framework for building digital libraries by crawling the web. In *Proceedings of the 2010 USENIX Conference on Web Application Development*, page 12. USENIX Association, 2010.

[18] P. Willett. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, 38(6):983–996, 1998.

[19] I. H. Witten and Et. Greenstone: A platform for distributed digital library applications. In *Research and Advanced Technology for Digital Libraries*, volume 2163/-1. Springer, 2001.

[20] J. Zhao, C. Goble, and R. Stevens. Semantic web applications to e-science in silico experiments. In *In Proceedings of WWW*, pages 284–285. ACM Press, 2004.