Scientific Data and Document Processing in $Chem_XSeer$

Prasenjit Mitra, C. Lee Giles, Bingjun Sun, Ying Liu, Anuj R. Jaiswal
The Pennsylvania State University, University Park, PA 16802.

Abstract

$Chem_XSeer$ is a digital library and a data repository for the chemistry domain. The data deposited into our repository is linked with digital documents to create aggregates of resources representing the links between the data and the articles in which the data is reported. $Chem_XSeer$ enables the user to annotate the data using a metadata capturing tool. The metadata is indexed and searched to return relevant datasets to the user. $Chem_XSeer$ extracts chemical formulae and chemical names, disambiguates them and indexes them to allow for domain-knowledge enhanced search capabilities. As search engines mature, we foresee such vertical search engines, employing domain-specific knowledge to perform information extraction and indexing, especially for scientific domains, become more popular. Though substantial research has been pursued on information extraction from text, extracting information from tables and figures has received little attention. In the $Chem_XSeer$ project, we are building tools that allow automatic extraction of tables and figures.

Introduction

Our $Chem_XSeer$ system comprises of a database and a digital library. The database hosts data related to chemical kinetics. The digital library hosts scholarly articles related to the domain of chemistry. Most of the data sets hosted in the database have also been published in papers that are stored in our digital repository.

We maintain an index of the scholarly articles and the published data and its associated metadata. Indexing articles, the data, and its associated metadata is necessary for providing users the ability to quickly search for articles of interest. Currently, we have crawled over 150,000 articles from the Royal Society of Chemistry repositories and indexed them. Experimental data are published at the website and an integrated search capability allows users to search the data using its several features. Scientists can easily publish their data by uploading it onto the repository and providing optional metadata to improve the search capabilities. The data is linked to the published articles such that the user can access an article and then drill down to examine the raw data after reading the article, or access a data set and then examine the article to find out the detailed experimental conditions, conclusions, discussions, and hypotheses that were validated using the experiments and their resultant data.

Enabling search and data retrieval in scientific domains is a hard problem because of the difficulty of automatically determining the semantics of the vocabulary and natural language used in the scientific documents and the metadata for the data sets. Retrieving documents that are relevant to the users' search queries (expressed using keywords) is a hard problem. Scientific domains like chemistry have additional sources of complexity in that the same formula can be written in different ways, e.g., CH3COOH or C2H4O2, and multiple chemical names are used for the same chemical. Our tools have to handle the problem of automatically disambiguating chemical formula, e.g., OH (hydroxyl) from abbreviations, e.g., OH (Ohio) and utilizing the disambiguated chemical formula while performing search.

While some search engines enable search on the text of digital documents, to the best of our knowledge, there exists no tool that can be used to search for tables across a large set of documents. A tool that enables an user to quickly search for tables reporting results for a particular experiment under certain conditions and shows the tables as search results laid out on one (or more) pages is extremely useful to scientists. As part of the $Chem_XSeer$ project, we have built TableSeer, a tool that enables searching for tables in digital documents.

A cyberinfrastructure, like $Chem_XSeer$ will enable (a) scientists to easily store their data, archive them, and publish them to the community, (b) allow users to search and retrieve data and articles easily. Allowing users intelligent and more accurate search capabilities enables users to find information that they require. Besides, due to tools like the formula search engine and TableSeer, the information is more readily available to the user, thereby reducing unproductive time for him or her. Currently, (we have been told that) NIST employs post-doctoral scholars to extract information from tables in documents with significant manual effort using crude tools. The scholars enter the data into a database. TableSeer currently identifies tables in PDF documents automatically, extracts their captions and references to the tables in text, and indexes these to enable searching for data. Extracting the data from these tables and identifying the semantics of the columns in the tables is currently being pursued.

Scientists find it difficult to publish scientific data on the Web because (a) the data comes in multiple formats, and (b) different types of metadata is required for different datasets. For each different type of data format, we need different parsers. On the other hand, we can create a "universal" metadata standard to capture all the metadata.

Our cyberinfrastructure(CI) has two main tools: a search engine that allows searching using keywords across digital documents including formula and table search and a database that is populated with experimental data derived from chemical kinetics. The database can be queried. $Chem_XSeer$ employs a mediated architecture where the end user can use the same query interface to query across the digital library and the database (across different

heterogeneous datasets). The keywords are looked up across text indexes that index the digital documents, formula indexes that index formulae in the documents, and the metadata and data of all the data in the database. All matching documents are ranked and presented to the end-user.

The $Chem_X$Seer project is a sister project of $CiteSeer_X$[11], the next generation CiteSeer project. Advances made in $CiteSeer_X$ will be incorporated into $Chem_X$Seer. Document management, indexing and citation management software developed in CiteSeer will be directly used in $Chem_X$Seer. However, CiteSeer does not have the capability to host experimental data nor can it link the data with scholarly articles where the results of the experiments are published. The novelty off $Chem_X$seer lies in the integrated data and document management capability it provides and the innovative information extraction algorithms it uses to extract and index information from tables and figures as well as domain-specific information like chemical formulae, chemical names and chemical structures. Other related digital libraries like Rexa[14] and Google Scholar[15] also do not provide the domain-specific search capabilities that are required to support scientists.

Knowledge-Enhanced Semantic Data Retrieval

Semantic Web technologies [17] have been deployed successfully in scientific applications [18]. These applications require the use of domain knowledge for successful deployment. Chemists gather and save their data in databases and repositories. These databases need to be annotated with metadata that explains the semantics of the data stored in the repositories.

Integration and Interoperation: End-users, often, have a need to integrate data and interoperate among digital libraries and scientific data repositories. In $Chem_X$Seer, chemists need to interoperate between the digital library and the scientific data repositories. The chemist is interested in searching for data and examining the data in the database. Then, upon observing some interesting phenomena from the data in the database, the chemist wants to read the article in the digital library that publishes the data, some aggregation of the data, or related results. The article provides important information, e.g., the conditions under which the experiment was conducted and the resultant data produced. The end-user uses such information to interpret the data.

Consequently, in order to allow the chemist to interoperate seamlessly, we have to address the following two cases:

(1) ***Table contains data reported in multiple articles***: When the table reports data from multiple articles, the data table has a column referencing the article that contains the data or its aggregate.

(2) ***Table contains data reported from a single article***: For a table that contains data only from a single article, the reference to the table is kept as metadata.

The creator of the dataset enters the metadata using an Excel Metadata capture tool. Using our metadata capturing tool built as a macro on Excel, the user entering the data can provide the reference to the article in which the data was published. The metadata for each Excel worksheet is stored along with the data in XML format in our database. The semantics of the metadata fields that describe the data must be explicated in a domain ontology. We are currently in the process of building such an ontology for the domain of chemical kinetics.

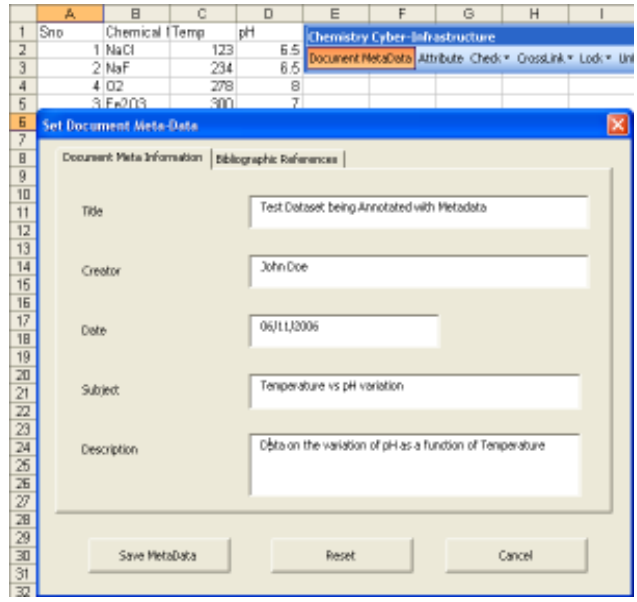

Figure 1: The Excel Toolbar and user form to input document metadata.

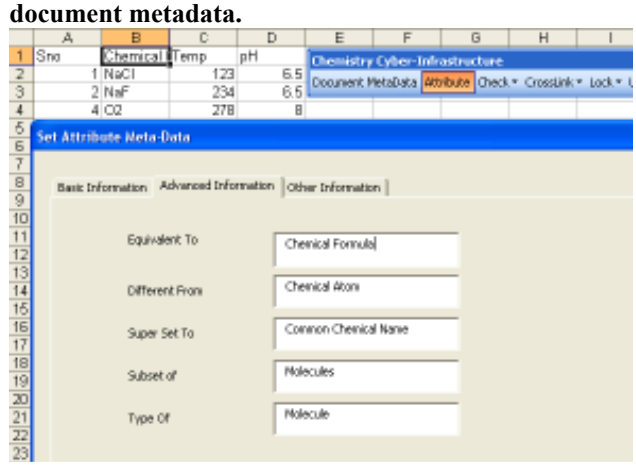

Figure 2: The user form to input attribute semantics and metadata.

The metadata for all tables are stored in a metadata table that associated a table identifier with its captured metadata. When the user wants to examine the article related to a dataset, the system queries the metadata table using the table-id of the related table to obtain the reference to the

article where the data was reported. Chem$_X$Seer fetches the article and returns it to the user. Internally, we use the Digital Object Identifier (DOI) of an article as reference to the article.

Digital Objects, ORE, and Interoperation: Chem$_X$Seer will use the Open Archives Initiative Object Reuse and Exchange protocol (ORE) to represent digital objects. The formal specification of ORE is released in March, 2008. Our initial work uses the Alpha release of OAI-ORE specifications[19]. ORE is built on top of the Resource Description Framework (RDF) [20].

Our repository consists of ORE aggregations. The ORE specification[19] indicates:

> An aggregation is a set of Resources which together form a "logical unit. Each of the individual Resources that compose an Aggregation is referred to as an Aggregated Resource.

We model each dataset as a Resource and each digital document as a resource. We create Aggregates of linked Resources, i.e., if an article contains information related to a dataset, an aggregate off the article and the database is created. An end-user can submit a dataset, an article, its preprints, associated supplementary notes, working notes as an aggregate with each component designated as an individual resource. The links among the aggregated elements are thus established during data input. ORE uses a resource map to record these links. The ORE specification[19] indicates:

> A Resource Map (ReM) is a Resource that specifies a URI for an Aggregation and describes its composition, properties, and relationships to other Resources.

Use of a standard like ORE makes it easy to share digital objects across repositories and enables interoperation.

Knowledge-enhanced Semantic Information Retrieval

The requirement for semantic annotation of the web has been well documented in numerous articles propounding the semantic web [17,18]. We believe that the most important impediment to the success of the semantic web is the lack of semantic annotations to the several billion webpages that exist and the lethargy of users towards using web annotation tools to annotate the webpages with semantic tags anchored in ontologies. We could not convince our chemists to tag all their webpages with semantic tags. Consequently, we took the approach that we must generate tools to automatically tag, at least, the basic entities in chemistry.

The entities that we have extracted, disambiguated, and automatically tagged with high accuracy are chemical names, and chemical formulae. Our tagging tool uses a Conditional Random Field (CRF) based classification algorithm. An expert tags documents to indicate which words are chemical formulae and which are chemical names. These tagged documents are used to train the supervised learning algorithm (CRF) to detect these entities. Our empirical evaluation shows that the entity extraction and disambiguation has x/y precision/recall.

Our system converts the chemical formulae to their canonical form. A canonical form, e.g., one where the elements in the formulae are lexicographically sorted, is chosen. Thus, CH4 and H4C would both be converted to their canonical form CH4. Apart from indexing each document using all their keywords, the Chem$_X$Seer maintains an index of chemical formulae. The key to an entry is the canonical form of a chemical formula and the inverted index points to all documents where the chemical formula and its various other representations appear.

This is an example where domain knowledge from the chemistry field has been inserted into the Chem$_X$Seer system to enhance information retrieval. Now, when the user searches for CH4 or H4C, using the chemical formula index, the user will retrieve all documents having variations of CH4, e.g., H4C, CD4 (D represents Deuterium, an isotope of hydrogen), etc. Our ranking function is also knowledge-enhanced in that not only does it attempts to rank the similarity between the user query term and the terms in the documents, but, it also allows for fuzzy searches, partial matches, and similarity measures that make sense in the chemistry domain. We believe that instead of a one-size fits all search engine, such domain-specific search engines that have been enhanced using domain knowledge and knowledge of the preferences of users in the domain will be rampant. General-purpose search engines do not precisely satisfy the requirements of domain scientists and users. Automatic entity extraction and disambiguation resulting in automatic semantic tagging will improve search engines and information retrieval. Chem$_X$Seer's search engine demonstrates that.

Figure 1: A screenshot of Chem$_X$Seer formula search

Figure 1 shows a screenshot of the $Chem_X$Seer formula search engine[2]. $Chem_X$Seer enables several types of approximate search queries. Initially, it lists the different types of formulae found that were similar to the search query. The user has the option of clicking on any of these formulae and of refining the search to obtain desired documents. The formula search engine uses a novel ranking function to rank the documents and return a ranked list of relevant documents to the user's search query [2].

Because $Chem_X$Seer allows the user to enter partial formulae and enables searches on that, there has to be an efficient index built to aid that. Specifically, it is impossible to index all possible sub-formulae of all chemical formulae occurring in the documents. Therefore, we propose to index only the formulae and sub-formulae that occur frequently independently (not as part of another frequent formula) in the entire set of documents being indexed. We show that our feature-selection based indexing method reduces the size of the index significantly with negligible reduction in the accuracy of the search results (due to space limitations, please see [2] for details).

Apart from chemical formula search, we also allow the search using chemical names [3]. Every document in the digital library is processed to identify the occurrence of chemical names and indexed. An innovative segmentation algorithm has been proposed that automatically identifies meaningful segments of chemical names. For example, Bis(2-chloro-1-methylethyl)ether is segmented into 'Bis', 'chloro', 'methyl', 'ethyl', and 'ether'. All of these are then indexed to allow users to specify partial chemical names and yet be able to obtain a set of search results that are meaningful.

Beyond Textual Information Extraction

Currently, primarily information available in the text of scientific documents has been extracted and indexed. Consequently, users use search engines to access such information. There have been advanced technologies, like semantic web technologies, that have been proposed to assist the search of textual data.

However, there is a large amount of information that is typically not indexed. For example, authors use tables and figures in digital documents to present the most important information, like summaries of experimental findings, etc. Arguably, these formats contain the most amount of information; however, they are not indexed by traditional search engines.

In the $Chem_X$Seer project, we aim to resolve two problems. The first is to develop algorithms and tools to extract such information automatically or semi-automatically from tables and images with minimal user interaction. The extracted information must be information that is useful to the domain scientist. In order to be effective, it must also use domain knowledge that a human reader would use to interpret the information in the tables and images.

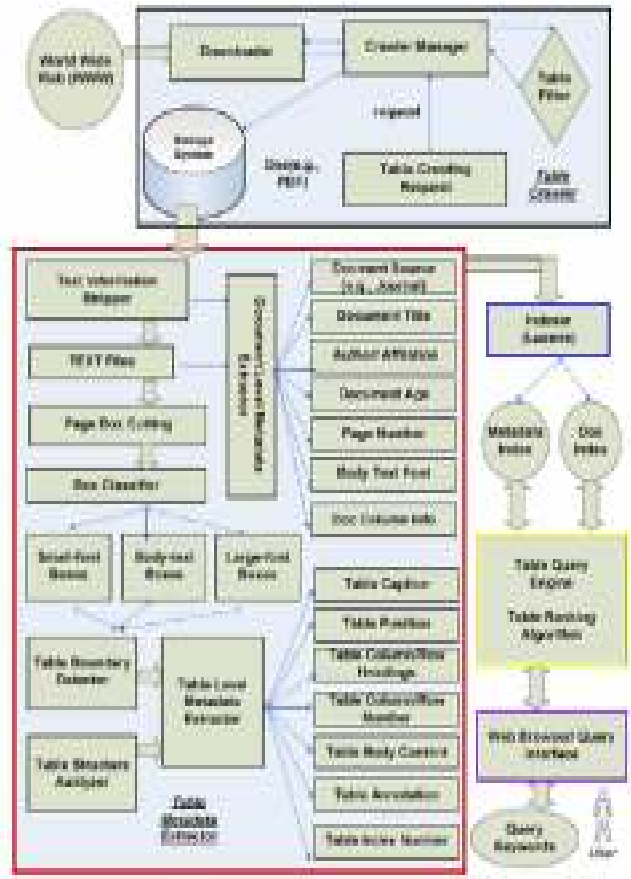

Figure 2: System Architecture for TableSeer

TableSeer: Table Extraction and Search: Figure 2 shows the architecture of the TableSeer toolkit [1,5,10] that allows users to enable automatic extraction and search for tables from digital documents. TableSeer extracts tables using its features like existence of columns, white space between columns and rows, differences in font-sizes and captions that often mention the term 'Table'. The major innovation of TableSeer lies in the fact that it identifies and associates references to a table in the document and uses the text associated with a reference as metadata to the table. This metadata is very useful to enable accurate search for tables across the digital library. For example, if there is a reference that says, "Table 1 shows the dissolution rate of Kaolinite … ", then this information is associated with Table 1 and if one searches for "dissolution rate" and "Kaolinite", this table will be returned as a result of that search. Such information is often available in the caption of the Table. However, in some papers, the table caption is not very descriptive but the description in the text contains much more information. By capturing both pieces of metadata and associating it with tables in the documents, TableSeer achieves high precision and recall for table search queries.

There is a substantial amount of heterogeneity in the way tables are formatted. Detecting the correct data and

semantics is a formidable problem. Several tables have multiple levels of column headings. That is, at the top level, the columns may have two headings, the first one covering groups of columns 1-5, and the next one 6-10. The next level of headers are at the column-level where each column has a heading. In a sense, the multi-level column headings define a hierarchy or ontology of terms. These terms and their semantics have to be captured. While it is possible to automatically identify certain aspects of the semantics of columns, like, units, the rest of the semantics has to be provided using manual input (at least currently). Automating this process fully is a challenging future task (that may never be fulfilled to 100% automation level). Domain-specific ontologies describing the classes and subclasses of concepts can be used to aid the detection off the semantics of the classes and their subclasses referred to by multi-level column headings. We are developing table metadata to describe attributes of tables.

TableSeer also employs an innovative ranking function that utilizes metadata like the terms that appear in the caption and the references, as well as terms that appear elsewhere in the document, like the document title, the publication date, etc., to determine the ranking of the tables retrieved. For example, it rewards recent articles over older ones because of an inherent assumption that more people will be interested in the latest results. However, if other metadata (like caption) of a table in an old document matches the search query very closely while the metadata off a second table in a new document does not match the search query, just being a new article will not cause the second document to be ranked higher. Our ranking function strives to achieve a good balance among the various table-level and document-level features that can be useful for ranking tables across digital documents[5].

Most existing work on table extraction and identification has been performed on documents that have been scanned. The closest work to ours is that of Pyreddy and Croft, who proposed the TINTIN system for retrieval of text tables [21] and that of Wei, Croft, and MacCallum [22]. Our work differs in that we use the references to the tables in text to extract additional metadata [1] and propose a novel ranking function [5] that improves the quality of the table search.

Extraction of Data from Figures

We are also pursuing extraction of datapoints from two dimensional (2-D) plots in figures in documents. Two dimensional plots are the visual counterparts of tables where, typically in the chemistry domain, several dependent variables are measured and plotted against values of an independent variable that is systematically varied. We are developing a metadata standard for figures to represent the various metadata elements associated with figures in digital documents. Furthermore, our extraction tools extract (a) axes in the figures, (b) their labels, (c) the tics on the X and Y axes, (d) legends, (e) line plots, and (f) data points in figures. After extracting the labels of the axes they are parsed using heuristics to identify the variables that have been plotted. These will then be matched to a domain-specific ontology to explicate the semantics of the variable. The label is also parsed to extract the units of the variable specified in it (if such a unit is specified in the label of the axes; otherwise the unit information has to be extracted from the text document). The legend is parsed to extract the symbols corresponding to the data points and their corresponding textual description. Again, these textual description will be matched to an ontology to disambiguate their semantics. The extraction of the data points poses a problem in the presence of overlapping data points. We have designed a novel segmentation algorithm to separate the overlapping data points. The line extraction has disambiguation issues when it has to determine the continuation of a line from a point where multiple lines intersect. We use a heuristic that matches line segments at an intersection point by matching pairs of segments with similar slopes such that the matching reduces an overall slope-difference measure [16]. All text segments in the figures are pass through Optical Character Recognition (OCR) software to identify the text. As should be reasonably evident, all these steps are error-prone and currently we have a 60% accuracy rate --- research is under way to improve that. However, even if 60% of the data is automatically extracted, it reduces the labor involved to identifying data and text from figures manually.

We are working on creating ontologies in the chemical kinetics and geochemistry domain to assist us in our work. Because the work is not mature, we have not reported on its progress.

Conclusion

In this paper, we have highlighted $Chem_X$Seer, an integrated data repository and digital library and have shown the architecture of a formula search engine designed for $Chem_X$Seer. We also present TableSeer a novel Table search utility that enables scientists to search for tables occurring in a large collection of digital documents and a brief outline of our ongoing work on data extraction from figures.

Acknowledgements

This work was partially supported by the U.S. National Science Foundation grant CHE-0535656.

References

[1] Liu, Y., Bai, K., Mitra, P., Giles, C.L. 2007. **"TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries"** *ACM IEEE Joint Conference on Digital Libraries (JCDL'07)*, Vancouver, British Columbia, Canada.

[2] Sun, B., Tan, Q., Mitra, P., Giles, C.L. 2007. **"Extraction and Search of Chemical Formulae in Text Documents on the Web"** *The 16th International World Wide Web Conference (WWW'07)*, Banff, Alberta, Canada, 2007.

[3] Sun, B., Mitra, P., Giles, C.L. 2008. Mining, Indexing, and Searching for Textual Chemical Molecule Information on the Web. *The 17th International World Wide Web Conference (WWW'08),* Beijing, China.

[4] Zhao, Q., Mitra, P., Giles, C.L. 2007. **"Image Annotation by Hierarchical Mapping of Features"** *The 16th International World Wide Web Conference (WWW'07), (Poster Track)*, Banff, Alberta, Canada.

[5] Liu, Y., Bai, K., Mitra, P., Giles, C. L. 2007. TableRank: A Ranking Algorithm for Table Search and Retrieval. Twenty-Second AAAI Conference, Vancouver, Canada, pp. 317-322.

[6] Bolelli, L., Lu, X., Liu, Y., Jaiswal, A., Bai, K., Councill, I., Mitra, P., Wang, J.Z., Mueller, K., Kubicki, J., Garrison, B., Bandstra J., Giles, C.L. 2007. **"ChemxSeer: A Chemistry Web Portal for Scientific Literature and Datasets"** *Open Repositories Conference*, San Antonio, Texas.

[7] Lu, X., Mitra, P., Wang J.Z., Giles C.L. 2006. **"Automatic Categorization of Figures in Scientific Documents"** *ACM and IEEE Joint Conference on Digital Libraries*, Chapel Hill, NC, June, 11-15.

[8] Jaiswal, A., Giles, C.L., Mitra, P., Wang, J.Z. 2006. An Architecture for Creating Collaborative Semantically Capable Scientific Data Sharing Infrastructures, *8th International Workshop on Web Information and Data Management (WIDM2006)*, Arlington, VA.

[9] Liu, Y., Mitra, P., Giles, C.L. Bai, K. 2006. **"Automatic Extraction of Table Metadata from PDF Documents (Poster)"** *ACM and IEEE Joint Conference on Digital Libraries*, Chapel Hill, NC, June, 11-15, 2006.

[11] Isaac G. Councill, C. Lee Giles, Ernesto Di Iorio, Marco Gori, Marco Maggini, Augusto Pucci 2006. Towards Next Generation CiteSeer: A Flexible Architecture for Digital Library Deployment. *ECDL 2006*: 111-122.

[12] C. Lee Giles 2006. The Future of CiteSeer: CiteSeer^x. *PKDD 2006*: 2.

[13] Huajing Li, Isaac G. Councill, Wang-Chien Lee, C. Lee Giles 2006. CiteSeerx: an architecture and web service design for an academic document search engine. *WWW 2006* pp. 883-884

[14] The Rexa Digital Library. Available at http://rexa.info/ on August 31st, 2007.

[15] Google Scholar. Available at http://scholar.google.com on August 31st, 2007.

[16] Lu, X., Wang, J., Mitra, P., Giles, C.L. 2007. Automatic Extraction of Data from 2-D Plots in Documents, 9th International Conference on Document Analysis and Recognition (ICDAR'07), Curitiba, Brazil, September, 2007.

[17] Berners-Lee, T., Hendler, J., Lassila, O. 2001. The Semantic Web. Scientific American.

[18] Hey, T., Trefethen, A.E. 2005. Cyberinfrastructure for e-Science, Science.

[19] Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R., Warner, S. ORE Specification and User Guide. Available from http://www.openarchives.org/ore/0.1/toc on February, 11th, 2008.

[20] Resource Description Framework, Available at http://www.w3.org/RDF/#specs on February, 11th, 2008.

[21] Pyreddy, P., Croft, W.B. 1997. TINTIN: a system for retrieval in text tables. *International Conference on Digital Libaries. Second ACM International Conference on Digital Libraries, DL'97,* Philadelphia, PA, ACM Press, New York, NY, pp. 193-200.

[22] Wei, X., Croft, W.B., and McCallum, A. 2006. Table extraction for answer retrieval. *Information Retrieval Journal*, volume 9, issue 5, pages 589-611, November 2006.