An Architecture for Creating Collaborative Semantically Capable Scientific Data Sharing Infrastructures

Anuj R. Jaiswal[1], C. Lee Giles[2, 3], Prasenjit Mitra[2, 3], James Z. Wang[2, 3]
[1]Department of Electrical Engineering
[2]College of Information Sciences and Technology
[3]Department of Computer Science and Engineering
The Pennsylvania State University
University Park, Pennsylvania, USA
{ajaiswal, giles, pmitra, jwang}@ist.psu.edu

ABSTRACT

Increasingly, scientists are seeking to collaborate and share data among themselves. Such sharing is can be readily done by publishing data on the World-Wide Web. Meaningful querying and searching on such data depends upon the availability of accurate and adequate metadata that describes the data and the sources of the data. In this paper, we outline the architecture of an implemented cyber-infrastructure for chemistry that provides tools for users to upload datasets and their metadata to a database. Our proposal combines a two level metadata system with a centralized database repository and analysis tools to create an effective and capable data sharing infrastructure. Our infrastructure is extensible in that it can handle data in different formats and allows different analytic tools to be plugged in.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Scientific Databases, H.3.5 [**Online Information Services**] – Data Sharing, Web-based Services.

General Terms

Design, Management.

Keywords

Scientific Databases, Architecture for Cyber-Infrastructures, Research Dataset Integration, Inter-operation.

1. INTRODUCTION

Scientific communities in domains such as chemical kinetics, computational chemistry and geo-sciences, consist of individual researchers and organizations who are data producers and users. Data is typically generated as a result of experiments, experimental devices and/or simulation programs. Data is then excerpted or summarized (in a publication whose author has analyzed the data to support a scientific research claim), or published to external end-users (i.e., is utilized by researchers other than the producer).

Researchers often build on prior inventions and discoveries using work done by their peers. For future research, the availability of the data produced from prior experiments is of great value. Such data sharing can bring around a significant increase in productivity; facilitate discovery, understanding, assessment, and validation. However, the true potential of effective utilization and analysis of any scientific data is limited because the data obtained as a result of experiments is not readily available to the research community and thus not easily reused.

Large aggregated data collections exist for research communities in domains as varied as global atmospheric and climatic research, computational chemistry, genomics and analytical physics allowing for effective utilization of such data for scientific research. Typically, such data collections are a result of data aggregation across large organizations such as NASA[1], NIST[2] and NOAA[3]. Researchers that intend to access, store and analyze this data form a large community of distributed members. Most often this data is locally downloaded and analyzed by researchers who utilize a variety of analysis tools for varying research requirements. However, such data collections do not allow datasets from similar domains to be integrated. Most often, the experimental data generated by an end-user (i.e., an individual scientist) will have a database schema that is different from the pre-existing ones. Typically, an organization such as NASA or NOAA collects data which follows their own database schema or else convert -other experimental data into their schema before the data is loaded into their databases. Sometimes, to avoid the effort of converting data conforming to different schemas, the collected data is simply stored in flat files. The disadvantage of such a choice is that either this data cannot be efficiently queried using indexes and query optimization routines that come with databases, requiring that structures and routines must be custom-built over the flat files.

Research experiments in a scientific domain typically have a vast number of variables, a subset of which a researcher might decide to use within his experimental framework or simulation program. For example, an experiment in the chemical kinetics domain might involve the following variables: “Chemical Name”, “pH” and “Temperature”. It is most likely that even two similar research experiments will have a different number of variables, though a majority of them might be common between the two experiments. For example, researcher A conducting an

WIDM ’06, November 10, 2006, Arlington, Virginia, USA
Copyright 2006 ACM 1-59593-524-8/06/0011…$5.00.

[1] http://www.nasa.gov

[2] http://www.nist.gov

[3] http://www.noaa.gov

experiment might have the variables “Temperature”, “Chemical Name” and “pH” while researcher B conducting the same experiment might have the same variables (and a fourth variable “Common Chemical Name”. Furthermore, individual researchers might name the same variable differently. For example, researcher A might name a variable for temperature “Temp”, while researcher B might name the same variable “Temperature”. All members of a scientific community may not easily agree on a fixed vocabulary or a standardized database schema. Reaching consensus can be costly and in the absence of efforts by a powerful organization (like a funding agency) scientists may not make the effort to adopt a common standard. Even more problematic, deciding upon a fixed database schema for such an application could prove difficult, since the number of variables that could exist in the database schema could be expansive.

We propose an implemented architecture for a cyber-infrastructure for the dissemination, sharing, querying and searching of scientific data on the World-Wide Web. In our system, scientists upload their data that is to be stored in databases accessible via the Web. Unlike the paradigm in databases where the schema is known and end-users pose queries based on that schema, the schema of all the data tables for our online database is not known. Consequently, there is a need for *data search* in any cyber-infrastructure for scientific data.

In order to retrieve data accurately in response to queries on the web, the data must be augmented with metadata. Without the semantics of the data, meaningfully querying of the data is difficult. For example, there must be sufficient metadata to answer at least the following questions:

(1) What experimental data does the table contain? What were the experimental conditions? What is the source of the data? When was the data uploaded?

(2) What are the semantics of each of the columns (rows)? What concept does each column (row) represent? What are their units?

This metadata, available to the end-user from the online repository, assists in the data search. The challenge lies in building tools that requires the scientist to know as little of the technology as necessary in order to populate the metadata. Ideally, the tools should generate the metadata automatically; with the scientist should verifying the metadata before uploading the data. However, totally automated metadata generation can be difficult. Therefore, we propose a semi-automatic metadata generation toolkit.

We have constructed an annotation tool that can be coupled with a standard data editor to elicit annotations that are stored as metadata in the database. This tool greatly reduces the manual effort required in constructing online scientific databases.

1.1 Related Work

There exists a variety of techniques to build scientific dataset infrastructures. Shosani et al. [12] proposed a storage architecture to optimize access to the large datasets on disk, that result from high energy physics experiments. Swiss-Prot [1], a protein sequence database and Online Mendelian Inheritance in Man (OMIM) [10], is a database of descriptions of human genes and genetic disorders widely used in genetics research. Both databases are similar in hierarchical structure and are heavily curated, i.e., they are maintained by extensive manual input from domain experts in the field. Both databases strive to reduce redundancy in genetics data. Pfam [2] is a large dataset of protein families and domains and is built on top of Ecobase [3] and OMIM. Buneman et al. [4] propose an architecture for archiving Scientific Datasets using XML.

The Ecobase project [3] proposes a distributed architecture to extract information from a number of autonomous and heterogeneous data sources (or providers) over the internet by use of mediators. Cavalcanti et al. [5] proposed an architecture which allows management of distributed scientific models and data. The Data grid [6] architecture uses an LDAP-based metadata implementation to manage datasets stored within the storage system. These systems extensively use metadata to query and locate datasets. They can be used primarily to aggregate datasets across an organization or expect the datasets to be integrated to follow a uniform metadata scheme. However, none of these solutions can be applied to create a community-based information aggregation system for datasets that have been obtained using different variables or schemas while allowing the capability to locate datasets by querying on variable names.

1.2 Overview of Our Proposed Architecture

Our goal is to create an infrastructure that allows experimental datasets from a domain to be aggregated and shared via a central data repository. Furthermore, the capability to search within experimental datasets corresponding to a set of variables is a must, since data reuse can occur more effectively if researchers can find datasets that suit their research needs. Additionally, the infrastructure targeting a specific scientific community must provide necessary data analysis tools that cater to that community.

- Metadata on variables allows for better semantic feature search since researchers can provide more semantic information on variables. Metadata on variables (or dataset attributes), can also remove the inconsistencies that are created by different users such as similar variables (or attributes) named differently and resolution of semantic differences between variables (or attributes) which have similar names but different meanings. Such metadata can then be utilized to generate dynamic collaborative ontologies on variables within the system, allowing for, inter-operability, ease of use and semantic correctness of variable names associated with a dataset.

- We propose a system architecture in Figure 1 which combines a metadata annotator with a central repository and portal. Our aim is to design an infrastructure that not only allows for experimental datasets to be aggregated, stored and shared but searchable not only at the dataset level but within the dataset itself. Specifically, we propose an architecture that allows researchers to annotate their datasets and provide metadata for the document (i.e., datasets) as well as for variables (i.e., dataset attributes). This will allow semantic search functionality at the dataset attribute level. To archive data, researchers then submit the URL (address) of the web available datasets. The infrastructure pulls datasets from the given addresses and metadata is automatically generated. Users can search for datasets and query them using a web portal-based front-end. Additionally, end-uses can also use the tools available via the portal to analyze the queried data within datasets (e.g., plot two variables on a graph, compute the correlation of the dependent variable pair-wise with each independent variable to determine which independent variable influences the dependent variable, etc.).

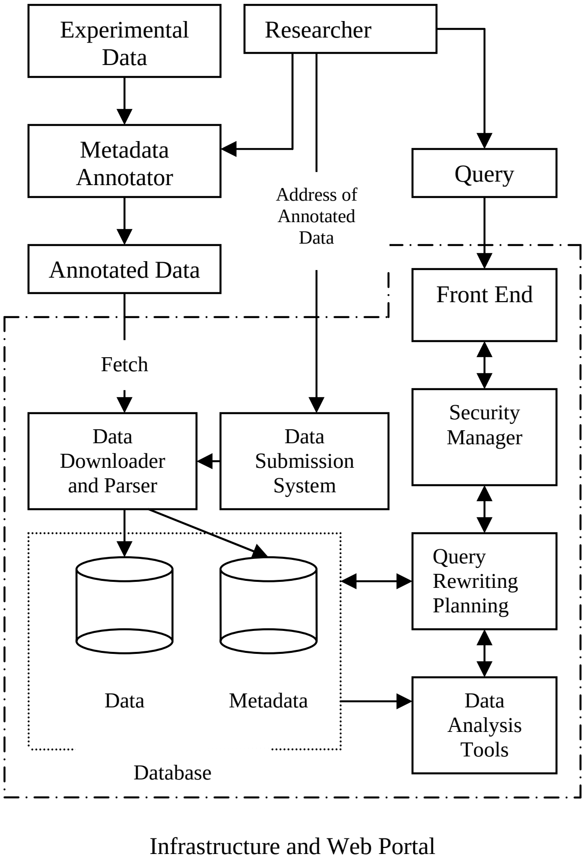


Figure 1: The architecture of our collaborative semantically capable infrastructure for data aggregation and sharing.

1.3 Contributions

In this paper

1. We propose an architecture for a semantically capable collaborative infrastructure for data collection and sharing.
2. Our system architecture utilizes a two level metadata scheme that provides metadata description for documents (experimental datasets) and semantic description of dataset attributes.
3. We then describe our current system implementation and show that such an architecture enables greater semantic search capabilities as well as the automatic generation of dynamic collaborative ontologies which will allow for greater inter-operation.

1.4 Outline of the Paper

The rest of this paper is organized as follows: Section 2 discusses our design issues, goals and solutions. Section 3 describes our current infrastructure implementation. We then conclude and suggest future research directions in Section 4.

2. DESIGN ISSUES AND GOALS

2.1 Metadata

The metadata description of datasets is crucial for effective utilization of the data repository and for better inter-operability. However, providing metadata is the responsibility of each data publisher (researcher), since each dataset is known correctly only to the data generator. To allow interoperability and effective sharing of datasets within the infrastructure, our proposed architecture utilizes a metadata annotation module that allows individual data publishers to annotate their data with metadata. Furthermore, our system utilizes a two level metadata scheme: 1) Dataset Level: metadata that describes the experimental dataset 2) Dataset Attribute Level: metadata that describes the variables (attributes) within each dataset. The following sections discuss the two level metadata scheme utilized within our infrastructure.

2.1.1 Dataset/ Document Metadata

The Dublin Core [9] is a widely used metadata standard for digital libraries and defines 15 elements for resource description: *Title, creator, subject, description, contributor, publisher, date, type, format, identifier, source, relation, references, is referenced by, language, rights* and *coverage*. This basic set of metadata elements is used by the Open Archives Initiative Protocol for Metadata Harvesting [10] (OAI-PMH) for minimal interoperability. The metadata annotation module utilizes the Dublin core metadata standard to provide dataset metadata. Our infrastructure allows users to locate datasets by querying the document metadata. For example researcher "John Doe" believes datasets from researcher "Jane Doe" satisfies his research requirements since she either works on similar research topics or is believed to provide good quality datasets. The dataset-level metadata allows researcher "John Doe" to locate datasets from a specific researcher, conference, description, date etc.

2.1.2 Attribute Metadata

Research datasets consist of data collected over a number of variables (attributes, experimental condition) that define the experiment and/or simulation program. Typically, similar datasets contain a subset of variables that are equivalent (i.e. they refer to the same variable or experimental condition). Quite often, researchers name their variables differently which results in semantic heterogeneity on similar lexical objects. As an example consider the three tables shown in Figure 2 which correspond to three different experimental datasets which are obtained by three different researchers performing similar experiments. Consider the variables "Temperature", "Temp" and "C" in the three tables respectively. In the above example, the three variables refer to the same attribute "Temperature". This results in semantic heterogeneity when defining the variable "Temperature". Though an intelligent lexical matching algorithm might be able to deduce correspondences between "Temp" and "Temperature" since "Temp" is a commonly used shortened version of "Temperature" in research datasets. However, no correspondence between "Temp" and "C" can be made. Further similar semantics might be used to refer to different objects. Consider the attribute "Rate" which is common in the three dataset fragments in Figure 2. It is quite likely that each table might refer to a different variable name for example "Chemical Rate", "Reaction Rate", "Dissolution Rate" or "Evaporation Rate". Assuming these three different variables "Rate" to be equivalent will cause incorrect attributes to be treated as semantically equivalent, which will result in incorrect interpretations and decisions.

Chemical Name	Rate	Temperature (K)
NaOH	0.1	278
H_2SO_4	0.2	279

Table 1: Researcher A Dataset

Chemical Name	Rate	Temp (F)
Sodium Hydroxide	0.1	300
Sulfuric Acid	0.2	301

Table 2: Researcher B Dataset

A	Rate	C (C)
Sodium Hydroxide	0.3	297
Sulfuric Acid	0.4	298

Table 3: Researcher C Dataset

Figure 2: Fragments of Three Datasets from different researchers performing the same experiment.

Furthermore, most variables associated with datasets produced by scientific research communities contain additional information, which if not captured will cause a significant loss in information. Consider the variable “Temperature” in a chemical-kinetics dataset. The variable might have additional information such as “Units”, “Measurement Type”, etc., which must be captured since they reflects the true nature (or scientific semantics) of data elements (values). For example, the variable “Temperature”, “Temp” and “C” in the three tables in Figure 2 respectively, represent the variable “Temperature”. However, the variable in Table 1 has units “Kelvin”, in Table 2 has units “Fahrenheit” and in Table 3 has units “Celsius”. If the information regarding units is not captured, and the data values in the three tables are treated equivalently, the data will get misrepresented.

To make progress in such difficult scenarios, our architecture utilizes attribute metadata that is annotated on to the data attributes by use of the metadata annotation module. Attribute level metadata describes each attribute within a research dataset more descriptively. Current attribute metadata tags that we utilize consist of the following: “Fully Qualified Name”, “Data Type”, “Units”, “Examples”, “Other Information”, “Equivalent To”, “Different From”, “Superset Of”, “Subset Of” and “Type of” tags. The explanation on each of these tags is shown in Table 1.

In addition, to resolving semantic heterogeneity resulting due to various syntaxes followed by different researchers, these attribute metadata tags allow for the generation of a dynamic collaboration ontology that defines the dataset attributes that are contained within the infrastructure. Such ontologies can be derived by utilizing a collaborative probabilistic score based on community-wide descriptions of attributes. For example, if 90% of the users within the infrastructure define that attribute “A” is different from attribute “B” and equivalent to attribute “C”, then we could derive an ontology which contains the reference to this community wide description of attribute “A” and its semantic mapping to other attributes. Such ontologies will further allow for better dataset inter-operation, search-ability and query rewrite capabilities.

2.2 Dataset Submission

Datasets can be entered into the infrastructure using either a “push” or “pull” based method. In the “push” or “put” technique, the datasets are directly submitted to the infrastructure while in the “pull” or “get” technique, the infrastructure gathers the datasets from a web accessible location. Our infrastructure architecture is designed using the “pull” based method because we believe that this is more suited to our infrastructure needs. The rationale is as follows:

- A pull based method provides greater security and because the malicious user cannot upload large sets of junk data onto the repository. They will submit the URL or the ftp site of where the data resides. Our tool fetches the data from the submitted location. For non-authorized users, the submitted data is checked by a moderator to ascertain the appropriateness of hosting it in our cyber-infrastructure.
- Datasets can then be tagged with the provenance information automatically, e.g., the source URI, the time, and the authenticated user. This information can be useful in the future to determine the quality and the reliability of the data or to detect malicious users.
- A push based infrastructure is less robust to malicious DOS attacks because we can implement a fair round-robin policy of fetching datasets across different users. Malicious users are banned.

Table 1: Metadata tags for defining attribute metadata for scientific datasets.

Metadata tag	Description
Fully Qualified Name	A descriptive name that describes this variable with minimal (scientific) semantic confusion.
Data Type	Describes the format the attribute data values follow. E.g., Text, Numeric
Units	Scientific units of this variable if present.
Examples	More examples and notes.
Equivalent To	Variable names that are equivalent.
Different From	Variable names that are different from this variable. E.g., Heat, Temperature.
Superset Of	Variables which are superset to this variable. E.g., “Rate” is superset of “Chemical Rate”.
Subset Of	Inverse of above.
Type Of	Type of definition (similar to that in RDF). E.g., “Temperature in Celsius” is a type of “Temperature”.
Comments	User comments.

3. CURRENT IMPLEMENTATION

This section describes our current implementation for creating a semantically capable collaborative data sharing infrastructure for the chemical-kinetics domain. We describe the metadata annotation and validation module and the basic infrastructure implementation.

3.1 Metadata Annotation and Validation Module

Most scientists in the chemical kinetics domain utilize Microsoft Excel to store their experimental datasets. Our current solution utilizes the use of an Excel Addin to provide client-side metadata annotation and validation. Once installed, the Addin provides an Excel Toolbar containing multiple buttons visible within Excel as shown in Figure 3. The Document Metadata button in Figure 3 allows researchers to provide Dublin Core metadata, which follows the OAI-PMH format via a user form (also shown) for the dataset. Similarly, each attribute (variable name) within the datasets can be annotated with attribute metadata tags as shown in Table 1 using the user form as shown in Figure 4. In addition, we validate the dataset by utilizing the attribute metadata tag "Data Type". This attribute metadata tag defines the syntax (encodings) that the data values follow in a variable or attribute. The validator checks all data values contained within all variables in the dataset and ensures that they strictly follow the attribute metadata tags that have been set. If any data values are incorrectly input, the user is informed and requested to make the necessary changes. The validation step reduces the data value errors that may occur when the dataset is sent to the infrastructure. For example, these errors occur for when a user inputs a "text" data value mistakenly into a variable whose data type has been set to "Numeric".

3.2 System Implementation

The infrastructure consists of the following subsystems: 1) Web Portal and Front end 2) Data downloader and Parser and 3) Data Analysis Toolkit.

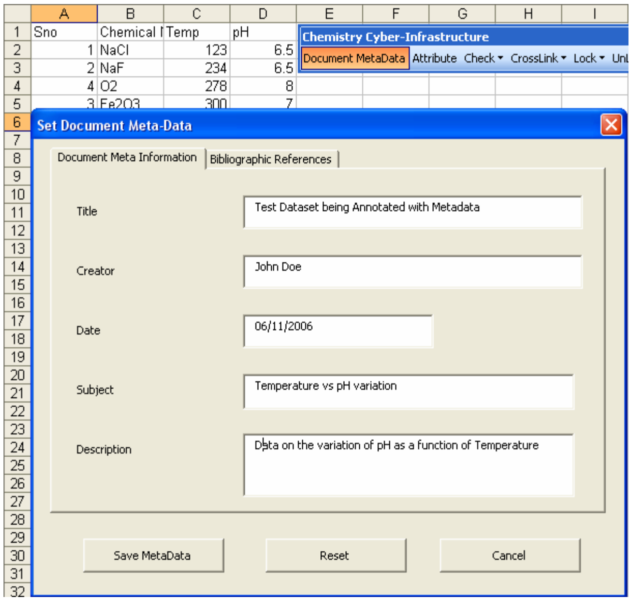
3.2.1 Web Portal and Front End

Users can access the infrastructure by use of a web portal. The infrastructure is currently deployed on an AMD Opteron-based server running Red Hat Enterprise Linux 4 Advanced Server with Apache Web Server, Apache Jakarta Tomcat 5 and MySQL database server. The web portal consists of 1) Content Management System 2) Dataset Viewer and 3) Data Submission System.

The content-management system is used manage user registration, moderation, set up security manager and to display dynamic web content to users. Registered users have access to complete functionality of the system including the data submission system. We have currently deployed a Mambo Server, an open-source PHP-based, content-management system with modifications, to implement the Web Portal. The Data submission system is deployed using Java Server Pages (JSP) and Servlet deployed on the Apache Jakarta Tomcat 5 Servlet container. Registered users can submit the URL/address of web accessible annotated datasets. Errors that are be generated while downloading and parsing the data are then reflected to the user, or else the dataset is uploaded and made available via the online database within the infrastructure. A data viewer allows users to query, download and view all available datasets. Our datasets in chemical kinetics do not require any read-access control. However, using standard database technology it is easy to provide access control to the data on a per-user basis if required.

3.2.2 Data Downloader and Parser

Once users have input the annotated dataset address into the dataset download queue, the data downloader and parser subsystem attempts to download and integrate the dataset into the infrastructure. Our data downloader and parser sub-system is currently implemented using C# .NET using the Visual Studio Professional 2005 Development Environment. The data downloader is implemented on a Pentium 4 based machine running Microsoft Windows XP Professional. The architecture of our data downloader and parser is shown in Figure 5.



Sno	Chemical I	Temp	pH
1	NaCl	123	6.5
2	NaF	234	6.5
4	O2	278	8
3	Fe2O3	300	7

Set Document Meta-Data

Title: Test Dataset being Annotated with Metadata

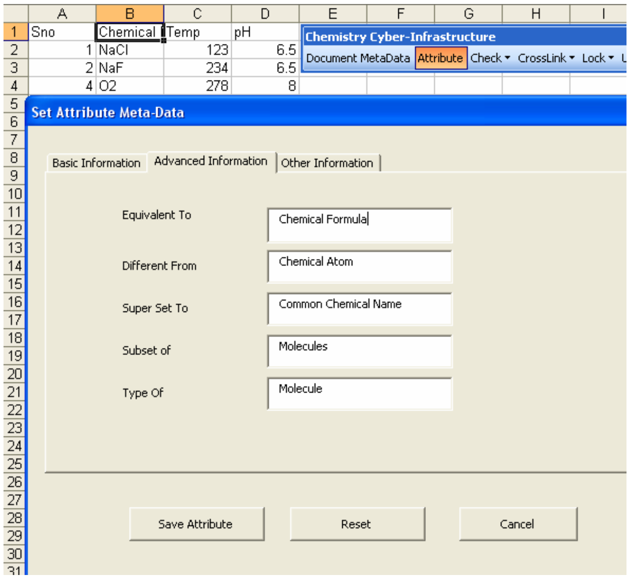
Creator: John Doe

Date: 06/11/2006

Subject: Temperature vs pH variation

Description: Data on the variation of pH as a function of Temperature

Figure 3: The Excel Toolbar and User Form to input document metadata.



Sno	Chemical I	Temp	pH
1	NaCl	123	6.5
2	NaF	234	6.5
4	O2	278	8

Set Attribute Meta-Data

Equivalent To: Chemical Formula

Different From: Chemical Atom

Super Set To: Common Chemical Name

Subset of: Molecules

Type Of: Molecule

Figure 4: The Excel User Form to input attribute semantics and metadata.

The data downloader and parser subsystem consists of 1) Scheduler 2) Data Downloader and 3) Dataset Parser. The Scheduler system retrieves an address (URL) of a new dataset to be integrated, into the infrastructure, from the data set queue. It assigns a unique ID to this dataset and accumulates errors that maybe encountered during data downloading, verification, parsing, metadata creation, table creation and/or dataset index update and logs the error information. The user who submitted the dataset can see the submission entry, its status, and its associated log of errors. The dataset downloader downloads the dataset from the URL/ address, caches the dataset on disk, performs dataset verification, virus scanning and renames the dataset to an internally consistent dataset identifier. The Dataset Parser then parses this cached dataset and automatically creates the metadata and the semantic dataset attribute metadata as XML files. Note that the metadata describing the dataset was created by the user as discussed in Section 3.1. An example dataset metadata XML file that is created by the dataset downloader and parser is shown in Figure 6. The dataset metadata file contains Dublin Core metadata as well as extra metadata that reference the Attribute metadata files as well as the data tables that are created for this dataset. An example of an attribute metadata (based on the data annotation input by a user) XML file that is created is shown in Figure 7, for the dataset attribute “Chemical Name”. All Data sheets within an excel file are converted to data tables in a MySQL database. The parser then creates a Dataset Index, which ties the dataset with dataset metadata; attribute metadata and data tables and transfers the corresponding files to disk on the infrastructure server.

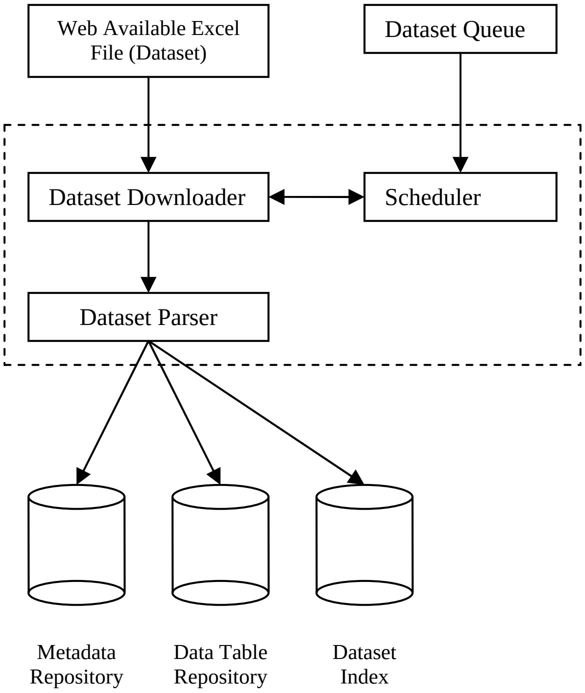


Figure 5: Dataset Downloader and Parser Architecture.

Capturing metadata automatically from datasets is problematic, since similar column headers (attributes) describing datasets can be expressed differently by individual researchers. In addition, multiple-rows of column headers, as shown in Figure 8, are used quite often by researchers. Deciding that row 2 or row 3 describes the dataset is the first problem. Capturing the semantics expressed in row 3 (which describes row 2) into metadata is another issue.

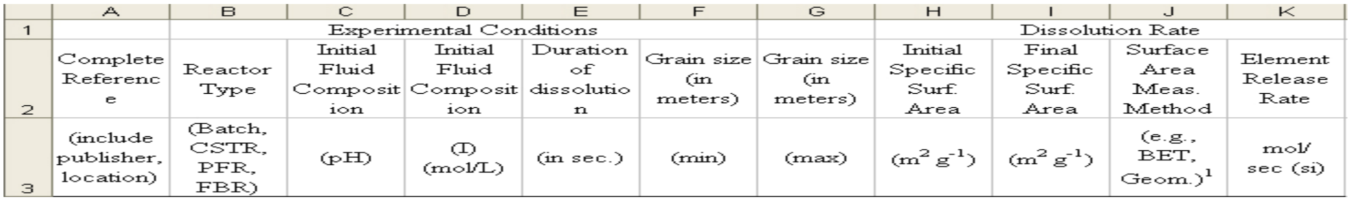
```
<document>
 <metadata>
   <fileLink>
     <file>38913_0298006366AttrMetaDataSheet1.xml</file>
     <file>38913_0298006366AttrMetaDataSheet2.xml</file>
   </fileLink>
   <tableLink>
     <table>Sheet138913_0298006366</table>
     <table>Sheet238913_0298006366</table>
   </tableLink>
   <oai_dc xmlns="http://www.w3.org/2001/XMLSchema"
xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_
dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">

<dc:source>http://www.johndoe.com/dataset1.xls</dc:source>
    <dc:title>Dataset with Annotated Data</dc:title>
    <dc:creator>John Doe</dc:creator>
    <dc:date>06/11/2006</dc:date>
    <dc:subject>Temperature vs pH variation</dc:subject>
    <dc:description>Data on the variation of pH as a function of
temperature for different chemicals</dc:description>
    <dc:references>None</dc:citation>
    <dc:isReferencedby>None</dc:isReferenceby>
    <dc:publisher>None</dc:publisher>
    <dc:venue>None</dc:venue>
   </oai_dc>
 </metadata>
</document>
```

Figure 6: Example Dataset Metadata created.

```
<document>
  <metadata>
    <attribute>
      <name>Chemical Name</name>
      <fullyQualifiedName>Chemical Name or
Formula</fullyQualifiedName>
      <dataType>String</dataType>
      <units>None</units>
      <examples>Sodium Chloride</examples>
       <equivalentTo>Chemical Formula</equivalentTo>
      <differentFrom>Chemical Atom</differentFrom>
      <supersetTo>Common Chemical Name</supersetTo>
      <subsetTo>Molecules</subsetTo>
      <typeOf>Molecule</typeOf>
      <Comments />
      <colIndex>2</colIndex>
      <rowIndex>1</rowIndex>
      <set>YES</set>
    </attribute>
  </metadata>
</document>
```

Figure 7: Example Attribute Metadata generated shown for a single variable “Chemical Name” within a dataset.

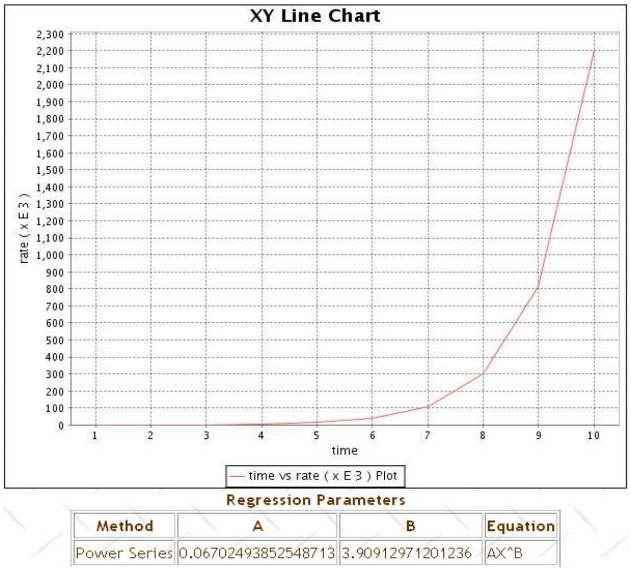


	A	B	C	D	E	F	G	H	I	J	K
1		Experimental Conditions						Dissolution Rate			
2	Complete Reference	Reactor Type	Initial Fluid Composition	Initial Fluid Composition	Duration of dissolution	Grain size (in meters)	Grain size (in meters)	Initial Specific Surf. Area	Final Specific Surf. Area	Surface Area Meas. Method	Element Release Rate
3	(include publisher, location)	(Batch, CSTR, PFR, FBR)	(pH)	(I) (mol/L)	(in sec.)	(min)	(max)	($m^2 g^{-1}$)	($m^2 g^{-1}$)	(e.g., BET, Geom.)[1]	mol/ sec (si)

Figure 8: Multiple-rows of column headers.

3.2.3 Data Analysis Toolkit

In addition the infrastructure provides chemical kinetic researchers with a data analysis toolkit. The toolkit currently provides an online plotting system and statistical data analysis system. The online plotting system allows users to query the datasets and plot data variables. It is implemented using JSP, Servlets and JDBC. An example X-Y-Line plot generated using a dataset present in the infrastructure is shown in Figure 9. In addition, a statistical toolkit is currently available which allows researchers to run statistical analysis such as regression, correlation and mutual information on datasets. Figure 10 shows the correlation results obtained after using the statistical correlation toolkit on an example dataset.



Regression Parameters

Method	A	B	Equation
Power Series	0.06702493852548713	3.90912971201236	AX^B

Figure 9: An example online plot and regression analysis generated by the infrastructure for a dataset.

4. CONCLUSIONS AND FUTURE WORK

We have proposed an architecture for creating a collaborative centralize infrastructure for sharing data with a client-side metadata-annotation module. Our current architecture utilizes a two-level metadata scheme, which provides document/dataset-level and dataset-variable/attribute-level metadata. Users can locate datasets by searching using the dataset metadata information or by querying datasets for specific variables names. In addition, online analysis tools such as plotting and statistical analysis allows the infrastructure to serve most needs of a researcher.

A great deal of room for future work exists. Developing algorithms to derive dynamic collaboration ontologies by using the attribute level semantics (i.e., metadata) is the first direction. Integrating query rewriting and semantic searching using attribute-level semantics and/ or community collaborative ontologies might provide for better data location and is an important direction. Another potentially important direction is to study the variable naming protocols that users within the infrastructure will follow. Using this information to automatically generate variable-name metadata in datasets that lack attribute-level metadata will be a significant direction. Using the user's past dataset submissions to automatically generate metadata for datasets is another important direction. Further, saving the metadata that a user enters and providing the user the ability to reuse this metadata is important, as this will significantly reduce the time required to generate metadata for new datasets.

Providing group, trust and privacy mechanisms for sharing datasets between users is another important direction. This will allow collaborating researchers to form groups and share datasets within the group. Currently, we have implemented a small subset of analysis tools that are used by chemical kinetics researchers. Providing complex curve fitting and graphing toolkits are important for better utilization of our infrastructure. Further, we intend to extend the infrastructure to be able to integrate Gaussian output and VASP datasets generated by chemical-kinetic experiments and simulation programs.

5. ACKNOWLEDGMENTS

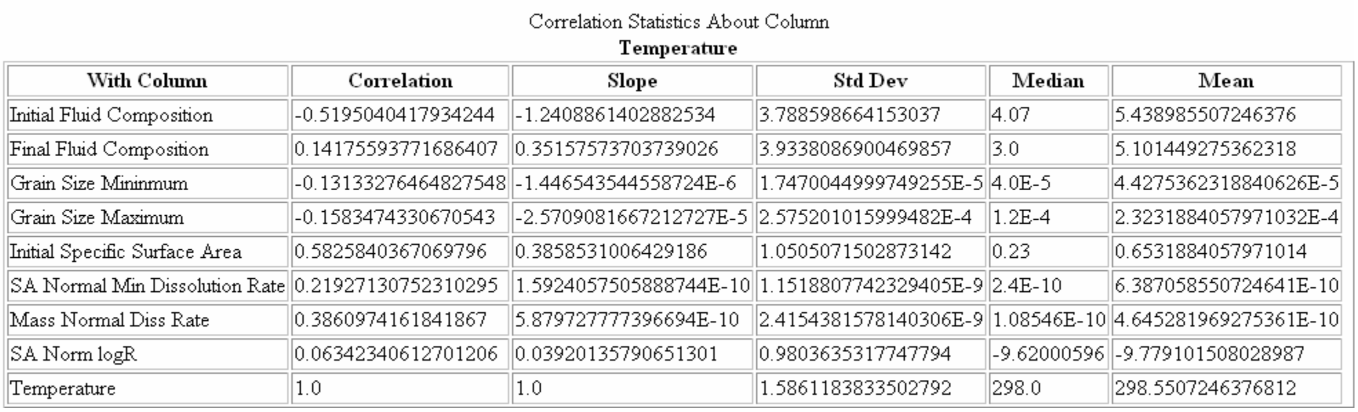
This work was supported in part by the US National Science Foundation under grants 0535656, 0347148, 0454052, and 0202007. The techniques we have developed are intended for practical use in scientific research dataset repositories.

The authors would like to acknowledge Joel Bandstra, from the Center of Environmental Kinetics Analysis at the Penn State University, for helpful discussions during the development of the infrastructure.

6. REFERENCES

[1] Bairoch, A., Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, vol. 28, no. 1, pp. 45-48, 2000.

[2] Bateman, A., Coin L., et al. The Pfam protein families database. *Nucleic Acids Research*, vol. 30, no. 1, pp. 276-280, 2002.



Correlation Statistics About Column
Temperature

With Column	Correlation	Slope	Std Dev	Median	Mean
Initial Fluid Composition	-0.5195040417934244	-1.2408861402882534	3.788598664153037	4.07	5.438985507246376
Final Fluid Composition	0.14175593771686407	0.35157573703739026	3.9338086900469857	3.0	5.101449275362318
Grain Size Mininmum	-0.13133276464827548	-1.446543544558724E-6	1.7470044999749255E-5	4.0E-5	4.4275362318840626E-5
Grain Size Maximum	-0.1583474330670543	-2.5709081667212727E-5	2.575201015999482E-4	1.2E-4	2.3231884057971032E-4
Initial Specific Surface Area	0.5825840367069796	0.3858531006429186	1.0505071502873142	0.23	0.6531884057971014
SA Normal Min Dissolution Rate	0.21927130752310295	1.5924057505888744E-10	1.1518807742329405E-9	2.4E-10	6.387058550724641E-10
Mass Normal Diss Rate	0.3860974161841867	5.879727777396694E-10	2.4154381578140306E-9	1.08546E-10	4.645281969275361E-10
SA Norm logR	0.06342340612701206	0.03920135790651301	0.9803635317747794	-9.62000596	-9.779101508028987
Temperature	1.0	1.0	1.5861183833502792	298.0	298.5507246376812

Figure 10: Correlation statistics toolkit run on an Example Dataset.

[3] Bouganim, L. et al. The Ecobase Project: Database and Web Technologies for Environmental Information Systems. *ACM SIGMOD Record*, vol. 30, no. 3, pp. 70 – 75, 2001.

[4] Buneman, P., Khanna, S., Tajima, K., Tan, W.C. Archiving scientific data. *ACM Transactions on Database Systems*, vol. 29, no. 1, pp. 2-42, 2004.

[5] Cavalcanti, M.C., Mattoso, M., Campos, M.L., Simon, E., Llirbat, F. An architecture for managing distributed scientific resources. In *IEEE International Conference on Scientific and Statistical Database Management*, 2002.

[6] Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., Tuecke, S., The Data Grid: Towards an architecture for the distributed management and analysis of large scientific datasets. *Journal of Network and Computer Applications*, vol. 23, No. 3, pp. 187-200, July 2000.

[7] Deutsch, A., Fernadez, M., Suciu, D. Storing Semistructured data with STORED. In *Proceedings of the ACM SIGMOD international conference on Management of data*. 1999.

[8] Dongilli, P., Franconi, E., Tessaris, S. Semantics driven support for query formulation. In *Proceedings of the International Workshop on Description Logics (DL)*, vol. 104, Whistler, BC, Canada, June 2004.

[9] Dublin Core Qualifiers. Dublin Core Metadata Initiative, 2000.

[10] Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., McKusick, V. A., Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, vol. 30, no. 1, pp. 52-55. 2002.

[11] OAI – Protocol for Metadata Harvesting. Open Archives Initiative. http://www.openarchives.org/. 2001.

[12] Shosani, A., Bernado, L. M., Nordberg, H., Rotem, D., Sim, A. Storage management for high energy physics applications. In *Computing in High Energy Physics (CHEP)*. 1998.