Research plan for semantics-enabled technology forecasting: A case study on alternative energies

Stuart Madnick
Wei Lee Woon

Working Paper CISL# 2007-14

October 2007

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E53-320
Massachusetts Institute of Technology
Cambridge, MA 02142

Research plan for semantics-enabled technology forecasting: A case study on alternative energies

Stuart Madnick, Wei Lee Woon

October 2007

1 Scope of Work and Research Advances Anticipated

1.1 Automatic Analysis of Technology Futures from Diverse Data Sources

The planning and management of research and development activities is a challenging task and one which is further compounded by the ever-increasing amounts of information which researchers and decision-makers are required to sift through on a daily basis. One of the most difficult problems is the need to gain a good understanding of the current state of research, future scenarios and the identification of technologies with great potential and which hence need to be emphasized. Information regarding past and current research is available from a wide variety of channels (examples of which include publication and patent databases), providing both a difficult challenge as well as a rich source of possibilities; on the one hand, sifting through these databases is time consuming and subjective, while on the other, they provide a rich source of data with which a well-informed and comprehensive research strategy may be formed [Por05, Por07].

This sets the tone for the proposed research project, which will focus on novel methods for automatically mining science and technology information sources. The aim is to extract patterns and trends which can support the formulation of research strategies. Examples of the kinds of information to be extracted include growth forecasts for technologies of interest and intuitive representations of interrelationships between technology areas. The prescribed course of research will develop a suite of techniques which will significantly extend and improve existing methods for performing so-called "tech-mining". In the following paragraphs we list a number of aspects which will form the basis for the contributions of this research:

1. *Advanced feature extraction methods* – Pattern recognition and data mining methods handle large data sets by extracting *features* which summarize some aspect of the data being analyzed. For example, many existing studies in technology forecasting and visualization depend on so-called “bibliometric” approaches – where the prevalence of a certain topic in particular databases is defined by the number of hits obtained using text or keyword searches. Unfortunately, this approach may not accurately reflect the actual state of the research area being studied (more on this in a later section). To address this issue, we propose the development of novel features which are capable of higher levels of precision when identifying research topics. Examples of such features are techniques which exploit co-occurrences between terms and methods for resolving words with multiple senses.
2. *Novel clustering and visualization algorithms* – Information extracted during data mining activities are often complex or high dimensional. A common situation is where the data

instances or individuals being studied are embedded in some high dimensional (and frequently inaccessible) space; frequently only measures of *similarity* between each of these items are available. This can occur in the case where the individuals are fields of research, and the similarity values are co-citation frequencies between associated publications. We propose to develop novel clustering and visualization algorithms.

3. *Context awareness* – One of the important aims of the project is to identify and utilize as many appropriate sources of information as is possible. While the initial scope of the study will focus on publication and patent databases (in accordance with other examples in the literature), we expect to incorporate information obtained from a variety of other sources; examples of these could range from formal sources such as research funding databases and industrial data to distributed and informal resources such as the popular press and internet blogs. Since these sources will be diverse in content and emphasis, we will apply and advance methods for extracting the relevant data from these sources and reconciling the differences in implicit assumptions (called "contexts") amongst the different sources.
4. *Information Quality* – As the number of information sources increases, a further issue which needs to be considered is the relative quality of the data being extracted. As a simple example, an article in the journal *Nature* reviewing a particular aspect of renewable energy should be taken a lot more seriously than a random blog posting. Similarly, a posting found on an internal corporate blog is likely to be a lot more reliable than one found on the internet. Although there has been a recent increase in research on information quality[1], we propose to extend this base of research to address the need to quantify the quality of processed data.

1.2 Case Study

We propose to explore the study of new technologies for analyzing trends and promising directions for alternative energy research as well as environmental and sustainability impacts as an initial focal area for demonstration.

The continued use of fossil fuels at prevailing rates is not sustainable, leading to rapid progress in the development of sustainable energies and a growing body of related research. An additional complication is that technologies for renewable energy do not emerge in isolation but are closely associated with a large number of other fields, examples of which include nanotechnology, material sciences, information technologies, mechanical engineering, physics and chemistry. Accordingly, planners of energy research are required to not only be familiar with directly related research themes, but also to be aware of the broader group of supporting topics. Because these strategies could ultimately affect the distribution and potential impact of large investments and research funds, any method or technique which can support and inform this process clearly provides significant value. An example of such an inter-connected technology question might be: "What impact might advances in nanotechnology have on improving the effectiveness of solar cells."

These factors, along with the central position of alternative energy technologies in the context of

[1] MIT established the first research initiative – the Total Data Quality Management (TDQM) Program and hosts the annual International Conference on Information Quality (ICIQ). Prof Madnick is one of the founding co-Editors-in-Chief of the new *ACM Journal on Data and Information Quality*.

the Masdar Initiative, makes this the ideal choice for the focus of our case study.

1.3 Project objectives and relevance to the Masdar Initiative

Motivated by the factors described above, we propose a course of research aimed at devising improved methods for technology forecasting and visualization. As a demonstration of the techniques developed, a detailed case study focusing on research relevant to renewable energies will be conducted. Research activities conducted in the course of the project will include (but is not restricted to) the development of techniques for conducting semantically-enriched searches of relevant databases, the elucidation of key subject-area clusters, the identification of indicators tailored to the renewable energy field and the creation of an energy-specific ontology.

The contribution of the project to the research themes of the Masdar Initiative can broadly be organized into three main areas. These are listed below along with a description of the associated project outputs.

1. *Enhancing the prestige and visibility of the Masdar Institute*. As a totally new institution, it is important for the Masdar Institute to rapidly establish visibility in the academic world. The proposed project will contribute towards this end by publishing aggressively in top **scholarly publications** (journals and selected conferences).
2. *Immediate benefits to the research goals of the institute*. In addition to the benefit listed above, the project is also expected to produce a number of outputs which directly impact research into renewable energy. The main contribution of this type will be a **report** detailing the findings of the case studies; this will cover, amongst other issues, the projected growth potential of certain alternative energy technologies. A secondary output will be an **ontology**[2] for describing research in alternative energy. Components of an ontology include a taxonomy of terms and definitions specific to alternative energy as well as descriptions of the relationships which exist between instances of associated objects.
3. *Building internal capabilities in key enabling technologies*. While not directly related to renewable energy, IT is a key enabling technology which potentially facilitates all other research areas in the institute. The proposed project spans a broad range of IT-related topics with an emphasis on the analysis of large and complex data sets, an issue which will be encountered in a variety of research topics.

   The specific contributions of the project in this context will be a suite of **software** and **techniques** for analyzing science and technology databases and other unstructured information sources, as well as **research personnel** trained in a variety of high-leverage IT topics including data mining, information aggregation and semantic technology.

[2] In philosophy, **ontology** (from the Greek ὤν, genitive ὄντος: *of being* (part. of εἶναι: *to be*) and -λογία: *science*, *study*, *theory*) is the study of being or existence and forms the basic subject matter of metaphysics. It seeks to describe or posit the basic categories and relationships of being or existence to define entities and types of entities within its framework.

2 Description of Methodologies and Analytic Techniques to be employed

2.1 Introduction

Broadly speaking, technology mining researchers apply one of the following two methods:

1. Predictive analyses where models trained on existing data are used to generate forecasts of the future evolution of the data. In the case of technological trend analysis, this approach is used to estimate the growth potential of the respective areas of research.
2. Exploratory analysis where the aim is to visualize and understand the structure of the data. Activities include the identification of clusters of research topics which represent underlying themes or concepts, the detection of trends and correlations between seemingly disparate fields of research and the visualization of these structures in a visually intuitive format.

The proposed work will draw upon both of these approaches; in particular we are keen to select best-of-breed approaches from a number of different domains, which, in combination with the expertise of the researchers working at MIT and the Masdar Institute will result in substantial improvements over previous efforts. These methodologies will be strongly grounded in the following disciplines:

1. Text and data mining
2. Semantic technology and context mediation

The above two areas will be discussed in greater detail in the following two subsections. In addition, the following topics are important elements of the project, though perhaps not the main focus:

- Bibliometric analysis
- Domain knowledge regarding sources of alternative energy.
- AI and knowledge representation

2.2 Data Mining

Data mining activities span a variety of applications but share the common theme of dealing with large data sets. Common uses of data mining are the detection of patterns in high dimensional data (exploratory analysis), the prediction of future occurrences of a particular set of data, and the classification of newly presented data into one of a set of previously encountered categories. In this project the main problems will be predictive and exploratory data mining. The following subsections will review instances of existing technology-mining research where data mining principles are applied, and discuss aspects which this project hopes to enhance.

2.2.1 Bibliometrics

Bibliometrics is the statistical analysis of text documents, typically publications and patents, focusing on the production and consumption, rather than the content of the documents [LOK00, TGH*06]. Such analysis begins with the extraction of features or indicators which characterize the progress or state of science and technology activities. Commonly used indicators include:

1. *The number of papers* – this feature is an indicator of the degree of scientific interest, and the level of scientific output associated with a field. This indicator is widely used but is only suitable if publication numbers are high. Paper counts also do not provide any indication of the quality of the associated research. For usage examples, see [BeN06, TGH*06,SCS*06].
2. *Number of co-citations* – this can function both as an indicator of the quality of the research (for e.g. the impact of a journal is often defined in terms of the number of citations received normalized by the number of papers), as well a means for linking papers or patents to related work. This measure is useful for studying the structure of a research field or community. For examples of use, see [Sma06, KYT*07].
3. *Number of authors (within a group of publications)* – an indicator of the degree of interest in a field or topic of research (see [ChH07] for example usage of this indicator).
4. *Average year of publication* – An indicator of the currency of a publication [KYT07].
5. *Number of co-authors (for a particular paper)* – A good indicator of the levels of co-operation, and possibly a means of linking related research themes. See [ChH07] for an example of usage of this indicator.

The methods described above require the identification of an associated document-set, typically obtained via a keyword search. By applying appropriate techniques such as context awareness and semantic search techniques, the proposed research should help to improve the accuracy of all the indicators above. However, the main beneficiary will be indicator (1), which will be enhanced through both increasing the search precision (by discounting papers which are unrelated), and by retrieving a larger sample of papers through query enrichment using semantic search techniques.

2.2.2 Technology forecasting

An important motivation for conducting technology-mining activities is the possibility of technology *forecasting*, i.e. prediction of the future evolution of a particular field of research.

This can be done manually by studying charts or other graphical representations of the publication data (see [KYT*07, ChH07, SCS*06]), but of more interest is extrapolation using growth curves, common examples of which include the Fisher-Pry (logistic) and the Gompertz curves (equations (1) and (2) respectively):

$$Y_t = \frac{L}{1 + ae^{-bt}} \quad \ldots (1)$$

$$Y_t = Le^{-ae^{-bt}} \quad \ldots . \quad (2)$$

These curves may be easily fitted to time-indexed bibliometric indicators, whereupon the forecasts may be conducted directly by analyzing the estimated parameters [BeN05, Mar03], or by including these parameters in more complex technological evolution models (see [DRM05, DGM*06], for example).

However, as in the previous subsection, the accuracy of these predictions is only as good as the ability to estimate the quantities being studied, be it publication count or patent co-citation or any of the other bibliometric indicators mentioned previously. Many interesting findings have already been reported in the literature [BeN05, DGM*06, KYT*07] but there are possibilities for refining this method further. Certainly, we believe that the proposed enhancements using

semantics and context awareness are likely to result in significant improvements to the results obtained when conducting this kind of analysis.

2.2.3 Research profiling/visualization

One challenging aspect when entering a new field of research is familiarization with the "research landscape". This is particularly important when evaluating the merits of alternative proposals or strategies; the researcher is quickly confronted with the need to identify the position of new or potential projects relative to existing work within the field. Not only does she need to recognize elements (or absence thereof) of novelty within the project, it is also important to identify niches or promising opportunities within the field and even potential collaborators [Por07].

This is one of the areas where automated technology mining techniques can be useful. There are a variety of possibilities where data mining can inform this initial stage of research. One important example is the process of visualizing interrelationships between research elements [Sma06, ZhP02], where these visualizations can be created using data extracted from publication databases. For example, in [ZhP02], a topographic representation depicting the research interests of various organizations is generated based on similarities between topics cited by these organizations.

A related activity is the partitioning of large sets of data into a smaller number of self-similar groups. This process, known as *clustering*, can be used to help identify key themes and concepts in complex domains, as has been done in [KYT*07, Sma06]. The two operations of clustering and visualization are illustrated in Figure 1.

In this context, the contribution of the project is twofold. Firstly, as in the case of technology forecasting, the benefits of using enhanced feature extraction (via contextual extensions, for example) will carry over to the visualization process as well; i.e.: gains in precision and recall are also likely to be reflected in more accurate visualizations, or more representative clusters. Secondly, while a variety of clustering and visualization algorithms have already been deployed for technology profiling, these have generally been secondary to the analysis of the findings of the study. By leveraging existing expertise amongst the Masdar and MIT researchers working on this project it will be possible to experiment with a greater variety of techniques and methods for clustering and visualization than has been attempted hitherto in the literature.

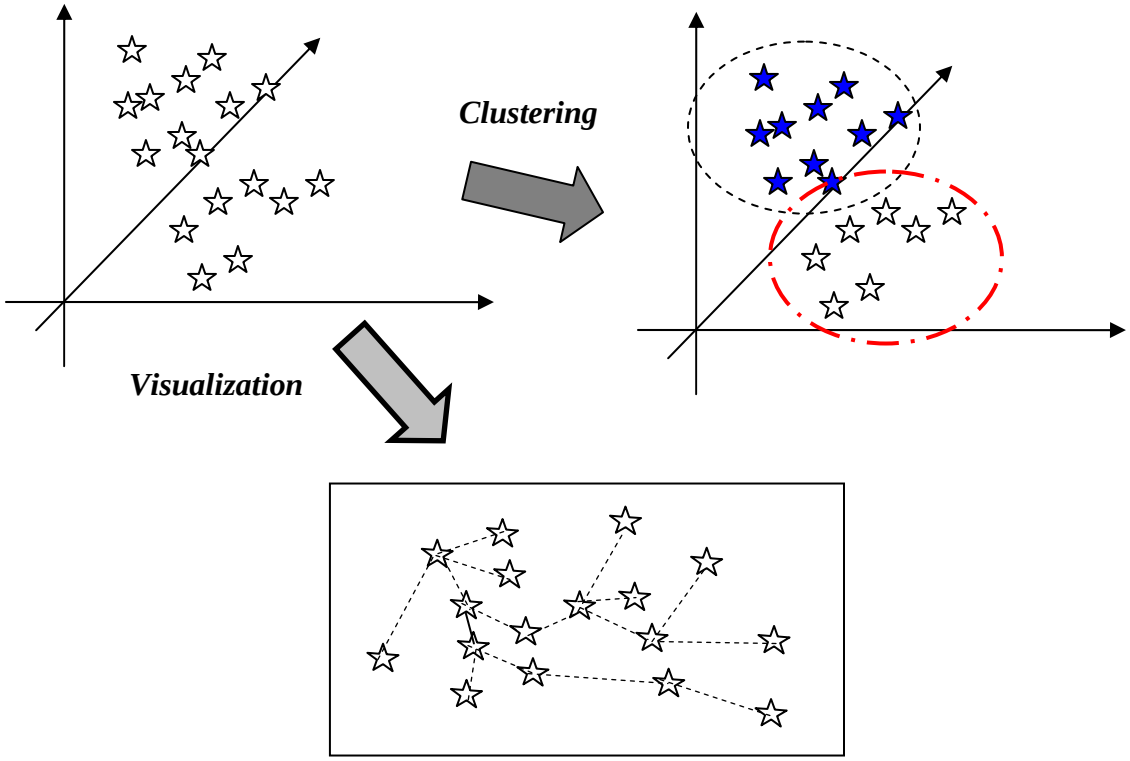

Figure 1 Clustering and visualizing operations

2.3 Data extraction and semantic reconciliation

As mentioned in the previous sections, we propose to use text and data mining to analyze diverse data sources to detect important trends and directions regarding new technologies (and combinations of technologies) as sources of alternative energy.

But, since the sources are diverse and heterogeneous, there are two immediate challenges: (1) Data Extraction and (2) Semantic Reconciliation.

Figure 2 gives a simple example of how these issues relate. The data sources are shown on the left side of the figure.

- Data Extraction: The data sources might be structured data bases, spreadsheets, or even semi-structured web pages. In order to do effective and efficient data mining, these must all be converted into a consistent format, usually structured data bases.

- Semantic Reconciliation: Since the data comes from diverse sources, there are usually many inconsistent implicit assumptions. Some simple examples are illustrated in Figure 2, such as the use of US style dates (mm/dd/yyyy) versus European style dates (dd/mm/yyyy) and costs in different currencies, such as US dollars versus Euros. These are trivial examples of much more complex cases that

are encountered in real data (a slightly more complex example is presented later.)

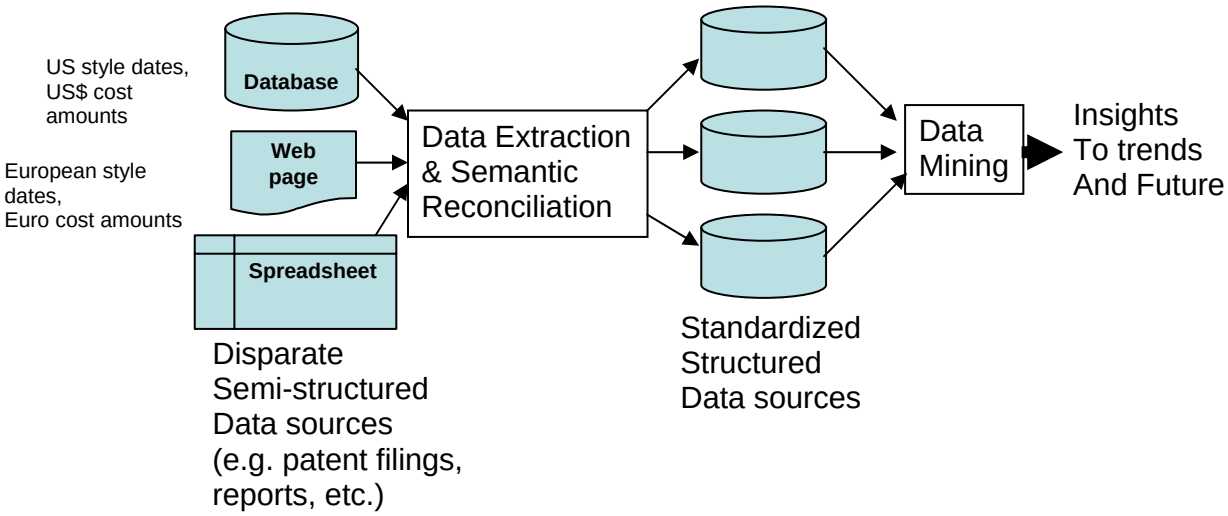

Figure 2 Data Extraction and Semantic Reconciliation

To address these challenges, we propose to use – and extent – technologies developed at MIT as part of its COntext INterchange (COIN) research effort:

- The COIN Web Wrapping technology to extract data from physically heterogeneous sources.
- The COIN Context Mediation technology to automate the process of semantic reconciliation.

2.3.1 Example of Semantic Reconciliation

For illustrative purposes only, let us consider an example of a question related to environmental analysis that draws on multiple diverse information sources – such as the types of information illustrated in Table 1. A specific question to be answered is: **to what extent have economic performance and environmental conditions in Yugoslavia been affected by the conflicts in the region?**

This is not an isolated case but one that illustrates concurrent challenges for information compilation, analysis, and interpretation – under changing and complex conditions.

For example, in determining the change of carbon dioxide (CO_2) emissions in the region, normalized against the change in GDP - before and after the outbreak of the hostilities – we need to take into account shifts in territorial and jurisdictional boundaries, changes in accounting and recording norms, and varying degrees of decision autonomy. User requirements add another layer of complexity. For example, what units of CO_2 emissions and GDP should be displayed, and what unit conversions need to be made from the information sources? Which Yugoslavia is of concern to the user: the country defined by its year 2000 borders, or the entire geographic area formerly known as Yugoslavia in 1990? One of the effects of war is that the region, which previously was one country consisting of six republics and two provinces, has been reconstituted into five legal international entities (countries), each having its own reporting formats, currency,

units of measure, and new socio-economic parameters. In other words, the meaning of the request for information will differ, depending on the *actors*, *actions, stakes* and *strategies* involved.

<table>
<tr><th>Domain and Sources Consulted</th><th>Sample Data Available</th><th>Basic Question, Information User Type & Usage</th></tr>
<tr><td>Economic Performance
• World Bank's World Development Indicators database
• UN Statistics Division's database
• Statistics Bureaus of individual counties</td><td>A. Annual GDP and Population Data:
<table><tr><td>Country</td><td>T0.GDP</td><td>T0.Pop</td><td>T1.GDP</td><td>T1.Pop</td></tr><tr><td>YUG</td><td>698.3</td><td>23.7</td><td>1627.8</td><td>10.6</td></tr><tr><td>BIH</td><td></td><td></td><td>13.6</td><td>3.9</td></tr><tr><td>HRV</td><td></td><td></td><td>266.9</td><td>4.5</td></tr><tr><td>MKD</td><td></td><td></td><td>608.7</td><td>2.0</td></tr><tr><td>SVN</td><td></td><td></td><td>7162</td><td>2.0</td></tr></table>- GDP in billions local currency per year
- Population in millions</td><td rowspan="4">Question:
How did economic output and environmental conditions change in YUG over time?

User 1: YUG as a geographic region bounded at T0:
<table><tr><td>Parameter</td><td>T0</td><td>T1</td></tr><tr><td>CO_2</td><td>35604</td><td>29523</td></tr><tr><td>CO_2/capita</td><td>1.50</td><td>1.28</td></tr><tr><td>GDP</td><td>66.5</td><td>104.8</td></tr><tr><td>GDP/capita</td><td>2.8</td><td>4.56</td></tr><tr><td>CO_2/GDP</td><td>535</td><td>282</td></tr></table>
User 2: YUG as a legal, autonomous state
<table><tr><td>Parameter</td><td>T0</td><td>T1</td></tr><tr><td>CO_2</td><td>35604</td><td>15480</td></tr><tr><td>CO_2/capita</td><td>1.50</td><td>1.46</td></tr><tr><td>GDP</td><td>66.5</td><td>24.2</td></tr><tr><td>GDP/capita</td><td>2.8</td><td>1.1</td></tr><tr><td>CO_2/GDP</td><td>535</td><td>640</td></tr></table>
Note (receiver' contexts):

<u>T0</u>: 1990 (prior to breakup)
<u>T1</u>: 2000 (after breakup)
<u>CO_2</u>: 1000's tons per year
<u>CO_2/capita</u>: tons per person
<u>GDP</u>: billions USD per year
<u>GDP/capita</u>: 1000's USD per person
<u>CO_2/GDP</u>: tons per million USD</td></tr>
<tr><td>Environmental Impacts
• Oak Ridge National Laboratory's CDIAC database
• WRI database
• GSSD
• EPA of individual countries</td><td>B. Emissions Data:
<table><tr><td>Country</td><td>T0</td><td>T1</td></tr><tr><td>YUG</td><td>35604</td><td>15480</td></tr><tr><td>BIH</td><td></td><td>1279</td></tr><tr><td>HRV</td><td></td><td>5405</td></tr><tr><td>MKD</td><td></td><td>3378</td></tr><tr><td>SVN</td><td></td><td>3981</td></tr></table>- Emissions in 1000s tons per year</td></tr>
<tr><td>Country History:
• CIA
• GSSD</td><td>T0.{YUG} = T1.{YUG, BIH, HRV, MKD, SVN}
(i.e., geographically, YUG at T0 is equivalent to YUG+BIH+HRV+MKD+SVN at T1)</td></tr>
<tr><td>Mappings Defined:
• Country code
• Currency code
• Historical exchange rates*

[As an interesting aside, the country last known as "Yugoslavia,"officially disappeared in 2003 and was replaced by the "Republics of Serbia and Montenegro." For simplicity, we will ignore this extra complexity.]

* Note: Hyperinflation in YUG resulted in establishment of a new currency unit in June 1993. Therefore, T1.YUN is completely different from T0.YUN.</td><td><table><tr><td>Country</td><td>Code</td><td>Currency</td><td>Currency Code</td></tr><tr><td>Yugoslavia</td><td>YUG</td><td>New Yugoslavian Dinar</td><td>YUN</td></tr><tr><td>Bosnia and Herzegovia</td><td>BIH</td><td>Marka</td><td>BAM</td></tr><tr><td>Croatia</td><td>HRV</td><td>Kuna</td><td>HRK</td></tr><tr><td>Macedonia</td><td>MKD</td><td>Denar</td><td>MKD</td></tr><tr><td>Slovenia</td><td>SVN</td><td>Tolar</td><td>SIT</td></tr></table>
<table><tr><td>C_From</td><td>C_To</td><td>T0</td><td>T1</td></tr><tr><td>USD</td><td>YUN</td><td>10.5</td><td>67.267</td></tr><tr><td>USD</td><td>BAM</td><td></td><td>2.086</td></tr><tr><td>USD</td><td>HRK</td><td></td><td>8.089</td></tr><tr><td>USD</td><td>MKD</td><td></td><td>64.757</td></tr><tr><td>USD</td><td>SIT</td><td></td><td>225.93</td></tr></table></td></tr>
</table>

Table 1. Operational Example: Information Available and Queries to be Answered

In this simple case, we suppose that the information request comes from a reconstruction agency interested in the following values: CO_2 emission amounts (in tons/yr), CO_2 per capita, annual GDP (in million USD/yr), GDP per capita, and the ratio CO_2/GDP (in tons CO_2/million

USD) for the entire region of the former Yugoslavia (see the alternative User 2 scenario in Table 1).

A restatement of the question would then become: **what is the change in CO_2 emissions and GDP in the region formerly known as Yugoslavia before and after the war?**

2.3.2 Diverse Sources and Contexts

By necessity, to answer this question, one needs to draw data from diverse types of sources (we call these differing *domains* of information) - such as, economic data (e.g., the World Bank, UN Statistics Division), environmental data (e.g., Oak Ridge National Laboratory, World Resources Institute), and country history data (e.g., the CIA Factbook), as illustrated in Table 1. Merely combining the numbers from the various sources is likely to produce serious errors due to different sets of assumptions driving the representation of the information in the sources. These assumptions are often not explicit but are an important representation of 'reality' (we call these the meaning or *context* of the information.)

The purpose of Table 1 is to illustrate some of the complexities in a seemingly simple question. In addition to variations in data sources and domains, there are significant differences in contexts and formats, critical temporality issues, and data conversions that all factor into a particular user's information needs. As specified in the table, time T0 refers to a date *before the war* (e.g., 1990), when the entire region was a single country (referred to as "YUG"). Time T1 refers to a date a*fter the war* (e.g., 2000), when the country "YUG" retains its name, but has lost four of its provinces, which are now independent countries. The first column of Table 1 lists some of the sources and domains covered by this question. The second column shows sample data that could be extracted from the sources. The bottom row of this table lists auxiliary mapping information that is needed to understand the meanings of symbols used in the other data sources. For example, when the GDP for Yugoslavia is written in YUN units, a currency code source is needed to understand that this symbol represents the Yugoslavian Dinar. The third column lists the outputs and units as requested by the user. Accordingly, for User 1, a simple calculation based on data from country "YUG" will invariably give a wrong answer. For example, deriving the CO_2/GDP ratio by simply summing up the CO_2 emissions and dividing it by the sum of GDP from sources A and B will not provide a correct answer.

2.3.3 Manual Approach

Given the types of data shown in Table 1, along with the appropriate context knowledge (some of which is shown in italics in Table 1), an analyst could determine the answer to our question – but through a time-consuming and error-prone process. The proper calculation involves numerous steps, including selecting the necessary sources, making the appropriate conversions, and using the correct calculations. For example:

<u>For time T0:</u>

1. Get CO_2 emissions data for "YUG" from source B;
2. Convert it to tons/year using scale factor 1000; call the result X;
3. Get GDP data from source A;
4. Convert to USD by looking up currency conversion table, an auxiliary source; call the result Y;

5. No need to convert the scale for GDP because the receiver uses the same scale, namely, 1,000,000;
6. Compute X/Y (equal to 535 tons/million USD in Table 1).

For time T1:

1. Consult source for country history and find all countries in the area of former YUG;
2. Get CO_2 emissions data for "YUG" from source B (or a new source);
3. Convert it to tons/year using scale factor 1000; call the result X1;
4. Get CO_2 emissions data for "BIH" from source B (or a new source);
5. Convert it to tons/year using scale factor 1000; call the result X2;
6. Continue this process for the rest of the sources to get the emissions data for the rest of the countries;
7. Sum X1, X2, X3, etc. and call it X;
8. Get GDP for "YUG" from source A (or alternative); Convert it to USD using the auxiliary sources;
9. No need to convert the scale factor; call the result Y1;
10. Get GDP for "BIH" from source E; Convert it to USD using the auxiliary sources; call the result Y2;
11. Continue this process for the rest of the sources to get the GDP data for the rest of the countries;
12. Sum Y1, Y2, Y3, etc. and call it Y;
13. Compute X/Y (equal to 282 tons/million USD in Table 1).

The complexity of this task would be easily magnified if, for example, the CO_2 emissions data from the various sources were all expressed in different metrics or, alternatively, if demographic variables were drawn from different institutional contexts (e.g., with or without counting refugees). This example shows some of the operational challenges if a user were to manually attempt to answer this question. This case highlights just some of the common data difficulties where information reconciliation continues to be made 'by hand'. It is easy to see why such analysis can be very labor intensive and error-prone. This makes it difficult under "normal" circumstances and possibly impossible under time-critical circumstances. This example may appear to be simple, but it includes major complexities such as reconciling spatial territoriality, currency, and atmospheric measures. Barriers to effective information access and reconciliation to prepare data for effective data mining and utilization usually involve complexities of this sort.

Although this particular example could be manually handled in a matter of a few hours, for data mining it is often necessary to gather and adjust hundreds, if not thousands, of such "data points." Thus, this example rapidly can grow to hundreds or thousands of hours of labor-intensive, time-consuming, and error-prone activities.

2.4 The Context Mediation Approach

Context Mediation technology, developed as part of MIT's COntext INterchange (COIN) Project, has developed a basic theory, architecture, and software prototype for supporting intelligent information integration employing context mediation technology [MAD99, GBM*99, GoBM96, Goh96, SM91a]. A fundamental concept underlying such a system is the representation of knowledge as Collaborative Domain Spaces (CDSs). A CDS is a grouping of

the knowledge including source schemas, data context, conversion functions, and source capabilities as related to a single domain ontology. The software components needed to provide harmonized information processing includes a context mediation engine [BGL*00, Goh96], one or more ontology library systems, a context domain and conversion function management system, and a query execution and planner [Fynn97]. In addition, support tools are required to allow for applications' (i.e. receivers', such as the data mining engine') context definition and source definitions to be added and removed easily (i.e., schemas, contexts, capabilities). We propose to utilize the current foundation of COIN as the starting point to develop new theories and methodologies for addressing the research challenges of this research effort.

The MIT COIN project has developed a platform including a theory, architecture, and basic prototype for such intelligent harmonized information processing. COIN is based on database theory and mediators [Wied92, Wied99]. Context Interchange is a mediation approach for semantic integration of disparate (heterogeneous and distributed) information sources as described in [BGL*00 and GBM*99]. The Context Interchange approach includes not only the mediation infrastructure and services, but also wrapping technology and middleware services for accessing the source information and facilitating the integration of the mediated results into end-users applications (see Figure 3).

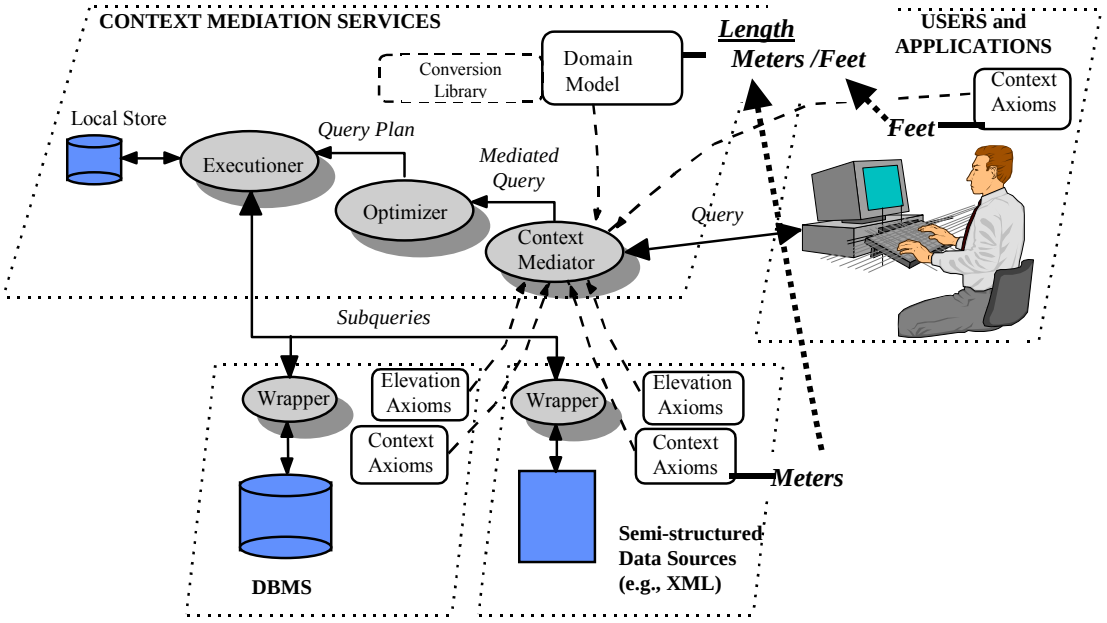

Figure 3. The Architecture of the Context Interchange System

The wrappers are physical and logical gateways providing uniform access to the disparate sources over the network [Chen99, FMS00a, FMS00b]. The set of Context Mediation Services, comprises a Context Mediator, a Query Optimizer and a Query Executioner. The Context Mediator is in charge of the identification and resolution of potential semantic conflicts induced by a query. This automatic detection and reconciliation of conflicts present in different information sources is made possible by ontological knowledge of the underlying application domain, as well as informational content and implicit assumptions associated with the receivers and sources.

The result of the mediation is a mediated query. To retrieve the data from the disparate

information sources, the mediated query is then transformed into a query execution plan, which is optimized, taking into account the topology of the network of sources and their capabilities. The plan is then executed to retrieve the data from the various sources, then results are composed and sent to the receiver.

The knowledge needed for harmonization is formally modeled in a COIN framework [Goh96], The COIN framework is a mathematical structure offering a robust foundation for the realization of the Context Interchange strategy. The COIN framework comprises a data model and a language, called COINL, of the Frame-Logic (F-Logic) family [KLW95, DT95]. The framework is used to define the different elements needed to implement the strategy in a given application:

- The Domain Model is a collection of rich types (semantic types) defining the domain of discourse for the integration strategy;
- Elevation Axioms for each source identify the semantic objects (instances of semantic types) corresponding to source data elements and define integrity constraints specifying general properties of the sources;
- Context Definitions define the different interpretations of the semantic objects in the different sources and/or from a receiver's point of view.

The comparison and conversion procedure itself is inspired by and takes advantage of a formal logical framework of Abductive Logic Programming [viz., KKT93]. One of the main advantages of the COIN abductive logic programming approach is the simplicity with which it can be used to formally combine and implement features of query processing, semantic query optimization and constraint programming.

Further technical details of the COIN theories is beyond the scope of this proposal but can be found in the References cited.

3 Description of Project Tasks and Individual Responsibilities

The division of labor shall be as follows:

Joint responsibilities

- Survey and selection of suitable databases and information resources
- Tools development
- Development of renewable energy specific ontology
- Choice of example applications

Masdar responsibilities

- Identification and implementation of appropriate visualization techniques
- Selection of suitable clustering methodologies, cluster evaluation and analysis
- Feature extraction using term frequency information, co-occurrence of terms and other similar information.
- Novel research in data mining and visualization

MIT responsibilities

- Adapting the context mediation framework for use with the project
- Identification of necessary “context modifiers: to augment the renewable energy ontology (see above)
- Development of necessary semantic reconciliation conversion algorithms

- Novel research in extending the "context interchange" approach

4 Description of Needed Facilities (equipments and computers)

- Two high performance workstations for computation and storage of project related materials
- Subscriptions to relevant Databases and Information Services
- Books and publications

5 Work Schedule and Milestone Chart, including any international or local travel anticipated

The execution of the project is envisioned to run in two main phases; the first phase will correspond to the initial period during which Masdar faculty members are stationed at MIT, while the second phase will correspond to all subsequent research. The division of tasks between these two phases is not strict but will approximately be as follows:

- *Phase 1*: the main activities during this phase will be the development and testing of requisite tools and utilities. Some data mining and analysis of various data sources will be conducted but primarily as proof-of-concepts and tests of the capabilities of the tools

- *Phase 2*: the tools and techniques developed during the first year will be applied and fine-tuned on the actual target data sources and specific case study (i.e.: energy-related research).

This division is a logical one because phase 1 activities will leverage the expertise of MIT researchers in developing the required software tools, as well as in the adoption and extension of existing MIT-developed technology. Phase 2 activities, on the other hand, are domain-specific and will require interaction with domain experts and stakeholders in Abu Dhabi.

Below is a Gantt chart depicting the approximate schedules for the key activities of the project. Note: The following chart assumes that work will commence on the 1st of October 2007. To be adjusted accordingly for other start dates.

Milestones

M1: 03/31/08 Interim Y1 report due
M2: 09/30/08 Full Y1 report due
M3: 03/31/09 Interim Y2 report due
M4: 09/30/08 Final report due.

Tasks \ Time period	10/07-12/07	01/08-03/08	04/08-06/08	07/08-09/08	10/08-12/08	01/09-03/09	04/09-06/09	07/09-09/09
Survey of databases								
Tool development								
Base indicators								
Base visualizations								
Contextual/semantic extensions								
Enhanced visualizations								
Data analysis								
Milestones		**M1**		**M2**		**M3**		**M4**

6 Deliverables

The expected outputs of the project and the related contributions to the Masdar Institute have been described in section 1.3, but we provide here a summarized listing of these contributions.

1. Software tools and techniques for conducting "tech-mining".
2. An ontology tailored to renewable energy technologies.
3. A detailed report describing the key findings of the case-study.
4. Scholarly publications in respected and peer-reviewed journals and conferences.

It is also helpful to distinguish between two classes of deliverables:

The first class consists of contributions which are methodological in nature, *viz* which involve the creation of novel techniques; items 1 and 2 in the list above fall into this category.

The second class, consisting of items 3 and 4, describe novel findings obtained via applications of the methods developed in this project.

7 References

[BeN06] Bengisu M and Nekhili R, "Forecasting emerging technologies with the aid of science and technology databases" Technological Forecasting and Social Change, **73(** 7) 835-844., 2006.

[BGL*00] Bressan, S.**,** Goh**,** C., Levina, N., Madnick, S., Shah, S., Siegel, S. (2000) "Context Knowledge Representation and Reasoning in the Context Interchange System", *Applied Intelligence: The International Journal of Artificial Intelligence, Neutral Networks, and Complex Problem-Solving Technologies,* **12**(2): 165-179**.**

[BGLM*00] Bresson, S., C. Goh, N. Levina, S. Madnick, A. Shah, and M. Siegel, "Context Knowledge Representation and Reasoning in the Context Interchange System," *The International Journal of Artificial Intelligence, Neutral Networks, and Complex Problem-Solving Technologies*, Volume 12, Number 2, September 2000, pp. 165-180.

[ChH07] Chiu H and Ho Y, "Bibliometric analysis of tsunami research", *Scientometrics*, **73(**1):3-17, 2007.

[DGM*06] Daim T, Guillermo R, Martin H. and Gerdsri P, "Forecasting emerging technologies: Use of bibliometrics and patent analysis", *Technological Forecasting and Social Change*, **73**(8):981-1012, 2006.

[DRM05] Daim TU, Rueda GR and Martin HT, "Technology forecasting using bibliometric analysis and system dynamics", *Technology Management: A Unifying Discipline for Melting the Boundaries*, pp. 112-122, 2005.

[FGM02] Firat, A., Grosof, B., Madnick, S. (2002) "Financial Information Integration In the Presence of Equational Ontological Conflicts," *Proceedings of the Workshop on Information Technology and Systems*, Barcelona, Spain, December 14-15: 211-216

[FMG02] Firat, A., S. Madnick, and Grosof, B., "Financial Information Integration In the Presence of Equational Ontological Conflicts," *Proceedings of the Workshop on Information Technology and Systems*, Barcelona, Spain, December 14-15, 2002, pp. 211-216 [Best Paper Award]

[FMG06] Firat, A. S. Madnick and B. Grosof, "Contextual Alignment of Ontologies in the eCoin Semantic Interoperability Framework",, to be published in *Information Technology and Management Journal*, Springer US. [SWP # 4567-06; CISL 2006-01, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=874097]

[FMM05] Firat, A. S. Madnick and F. Manola, "Multi-dimensional Ontology Views via Contexts in the ECOIN Semantic Interoperability Framework", *Proceedings of the International Workshop on Contexts and Ontologies: Theory, Practice, and Applications*, Pittsburgh, Pennsylvania, July 9, 2005 [SWP # 4543-05, CISL 2005-05, ESD 2005-02, http://esd.mit.edu/WPS/esd-wp-2005-02.html, http://ssrn.com/abstract=729383].

[FMS00a] Firat, A., Madnick, S., Siegel, S. (2000) "The Caméléon Web Wrapper Engine", *Proceedings of the VLDB2000 Workshop on Technologies for E-Services,* September 14-15.

[FMS00b] Firat, A., Madnick, S., Siegel, S. (2000) "The Caméléon Approach to the Interoperability of Web Sources and Traditional Relational Databases," *Proceedings of the Workshop on Information Technology and Systems*, December.

[Früh94] Frühwirth, T., "Temporal Reasoning with Con-straint Handling Rules," ECRC-94-5, 1994.

[Früh98] Frühwirth, T., "Theory and Practice of Constraint Handling Rules", Special Issue on Constraint Logic Programming (P. Stuckey and K. Marriot, Eds.), Journal of Logic Programming, Vol 37(1-3): 95-138, October 1998

[FYKMB05] Firat, A. N. A. Yahaya, C.W. Kuan, S, Madnick, and S. Bressan, "Information Aggregation using the Caméléon# Web Wrapper", *Proceedings of the 6th International Conference on Electronic Commerce and Web Technologies* (EC-Web 2005)," Copenhagen, Denmark, August 23 - August 26, 2005, also published in Springer Lecture Notes in Computer Science (LNCS) 3590, K. Bauknecht et al (Eds.), pp.76-86, Springer-Verlag Berlin, 2005 [SWP # 4562-05, CISL 2005-06, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=771492].

[Fynn97] Fynn, K.D. (1997) *A Planner/Optimizer/Executioner for Context Mediated Queries*, MS Thesis, MIT.

[GBM*99] Goh, C.H., Bressan, S., Madnick, S., Siegel, S. (1999) "Context Interchange: New Features and Formalisms for the Intelligent Integration of Information", *ACM Transactions on Information Systems*, 17(3): 270-293.

[GBMS99] Goh, C., S. Bresson, S. Madnick, and M. Siegel , "Context Interchange: New Features and Formalisms for the Intelligent Integration of Information," *ACM Transactions on Information Systems*, July 1999.

[GoBM96] Goh, C.H., Bressan, S., Madnick, S.E., and Siegel, M.D. (1996). "Context Interchange: Representing and reasoning about data semantics in heterogeneous systems," *Sloan School Working Paper #3928*, Sloan School of Management, MIT, 50 Memorial Drive, Cambridge MA 02139.

[Goh96] Goh, C. (1996). *Representing and Reasoning about Semantic Conflicts* In *Heterogeneous Information System*, PhD Thesis, MIT.

[Kal03] Kaleem, M. B., "CLAMP: Application Merging in the ECOIN Context Mediation System using the Context Linking Approach," CISL Working Paper 2003-05 and MIT Thesis, 2003.

[KKT93] Kakas, A.C., Kowalski, R.A., and Toni, F. (1993). "Abductive logic programming," *Journal of Logic and Computation*, 2(6):719--770.

[KLW95] Kifer, M., Lausen, G., and Wu, J. (1995). "Logical foundations of object-oriented and frame-based languages," *JACM*, 4:741--843.

[KYT*07] Kajikawa Y, Yoshikawa J, Takeda Y and Matsushima K, "Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy", *Technological Forecasting and Social Change*, (In Press) 2007.

[LCN*99] Lee, T., Chams, M., Nado, R., Madnick, S., Siegel, M. (1999) "Information Integration with Attribution Support for Corporate Profiles", *Proceedings of the International Conference on Information and Knowledge Management,* November: 423-429.

[Lee02] Lee, T. "Attribution Principles for Data Integration: Technology and Policy Perspectives - Part 1: Focus on Technology," CISL Working Paper 2002-03 and MIT Thesis, 2002.

[Lee03] Lee, P. "Metadata Representation and Management for Context Mediation," CISL Working Paper 2003-01 and MIT Thesis, 2003.

[LMB98] Lee, T., Madnick, S., and Bressan, S. (1998) "Source Attribution for Querying Against Semi-Structured Documents", *Proceedings of the ACM Workshop on Web Information and Data Management* (WIDM'98), Washington, DC, November 6: 33-39.

[LMS96b] Lee, J., Madnick, S., Siegel, M. (1996) "Conceptualizing Semantic Interoperability: A Perspective from the Knowledge Level", *International Journal of Cooperative Information Systems*: [Special Issue on Formal Methods in Cooperative Information Systems], 5(4), December.

[LOK00] Losiewicz P, Oard D and Kostoff R, "Textual data mining to support science and technology management", *Journal of Intelligent Information Systems*, **15(**2):99-119, 2000.

[Mad01] Madnick, S., "The Misguided Silver Bullet: What XML will and will NOT do to help Information Integration," *Proceedings of the Third International Conference on Information Integration and Web-based Applications and Services* (IIWAS2001; Linz, Austria), published by Osterreichische Computer Gesellschaft (ISBN 3-85403-157-2), September 2001, pp. 61-72.

[Mad03] Madnick, S., "Oh, So That is What you Meant! The Interplay of Data Quality and Data Semantics," Proceedings of the 22nd International Conference on Conceptual Modeling (ER'03), Chicago, October 2003; in Conceptual Modeling – ER 2003, (ISBN 3-540-20299-4) Springer-Verlag, 2003, pp. 3-13.

[Mad99] Madnick, S. (1999) "Metadata Jones and the Tower of Babel: The Challenge of Large-Scale Heterogeneity," *Proceedings of the IEEE Meta-data Conference*, April.

[Mar03] Martino JP "A review of selected recent advances in technological forecasting", Technological Forecasting and Social Change, **70**(8):719-733.

[MBM*98] Moulton, A., Bressan, S., Madnick, S., Siegel, M. (1998) "Using an Active Conceptual Model for Mediating Analytic Information Exchange in the Fixed Income Securities Industry", *Proceedings of the 17th International Conference on Conceptual Modeling* (ER'98), Singapore, November.

[MHR00] March, S., Hevner, A., Ram, S. (2000) "Research Commentary: An Agenda for Information Technology Research in Heterogeneous and Distributed Environments", Information Systems Research, **11**(4): 327-341.

[MS02] Madnick, S., M. Siegel. "Seizing the Opportunity: Exploiting Web Aggregation", *MISQ Executive*, Vol 1, No. 1, March 2002, pp. 35-46.

[MWX03] Madnick, S., Wang, R., and Xian, X, "The Design and Implementation of a Corporate Householding Knowledge Processor to Improve Data Quality," Journal of Management Information Systems, Vol. 20, No. 3, Winter 2003-04.

[MWZ02] Madnick, S., Wang, R., and Zhang, E., "A Framework for Corporate Householding,",

Proceedings of the International Conference on Information Quality), Cambridge, November 8-10, 2002, pp. 36-46.

[Oku97] Okubo Y, "Bibliometric indicators and analysis of research systems: methods and examples", *OECD Science,Technology and Industry Working Papers,* 1997.

[Por05] Porter, A, "Tech Mining", *Competitive Intelligence Magazine*, **8**(1):30-36, 2005.

[Por07] Porter A, "How 'tech mining' can enhance R&D management", *Research Technology Management*, **50**(2):15-20, 2007.

[Sma01] Smalheiser NR, "Predicting emerging technologies with the aid of text-based data mining: the micro approach", *Technovation*, **21:** 689-693, 2001.

[SCS*06] Santo MM, Coelho GM, Santos DM, Filho LF, "Text mining as a valuable tools in foresight exercises: a study on nanotechnology", *Technological Forecasting and Social Change*, **73**(8):1013-1027, 2006.

[SM91a] Siegel, M. and Madnick, S. (1991). "Context Interchange: Sharing the Meaning of Data," *SIGMOD RECORD*, **20**(4), December: 77-8.

[Sma06] Small H, "Tracking and predicting growth areas in science", *Scientometrics*, (**68**):595-610, 2006.

[Tar02] Tarik A, "Capabilities Aware Planner/Optimizer/Executioner for COntext INterchange Project," CISL Working Paper 2002-01 and MIT Thesis, 2002.

[TTM04] Tan, P. K-L Tan S. Madnick,, "Context Mediation in the Semantic Web: Handling OWL Ontology and Data Disparity through Context Interchange,", *Proceedings of the International Workshop on Semantic Web and Databases* (SWDB), Toronto, Canada, 30 August - 3 September 2004; also published in Springer Lecture Notes in Computer Science (LNCS) 3372, C. Bussler et al (Eds.), pp.140-154, Springer-Verlag Berlin, 2005 [SWP #4496-04, CISL 2004-13, Center for eBusiness Working Paper #209, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=577225].

[WFMA06] Wu, L. A. Firat, S. Madnick and T. Alatovic, "Querying Web-Sources within a Data Federation", (with), *Proceedings of the International Conference on Information Systems* (ICIS), Milwaukee, Minnesota, December 2006, [SWP #4624-06, CISL 2006-09, http://ssrn.com/abstract=926628]

[Wied92] Wiederhold, G. (1992). "Mediation in the Architecture of Future Information Systems", IEEE *Computer*, **25**(3): 38-49.

[Wied99] Wiederhold, G. (1999). "Mediation to Deal with heterogeneous Data Sources", Proceedings of Interop'99, Zurich, March: 1-16.

[WKM93] Wang, R., Kon, H., and Madnick, S (1993). "Data Quality Requirements Analysis and Modeling", *International Conference on Data Engineering*, 670-677

[ZhP02] Zhu D and Porter D, "Automated extraction and visualization of information for technological intelligence and forecasting*", Technological Forecasting and Social Change*, **69(**5):495-506, 2002.

[ZM05] Zhu, H. and S. Madnick, "Structured Contexts with Lightweight Ontology", *Proceedings of the International AAAI Workshop on Modeling and Retrieval of Context* (MRC2006), Boston, July 16-17, 2006 [SWP #4620-06, CISL 2006-05, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=926605]

[ZM06] Zhu, H. and S. Madnick, "A Lightweight Ontology Approach to Scalable Interoperability", *Proceedings of VLDB Workshop on Ontologies-based techniques for DataBases and Information Systems* (ODBIS 2006), Seoul, Korea, September 11, 2006, pp. 45-54, [SWP #4621-06, CISL 2006-06, http://ssrn.com/abstract=926605]

[ZM06] Zhu, H. and S. Madnick, "Improving Data Quality Through Effective Use of Data Semantics", *Data & Knowledge Engineering* (DKE), Vol. 59, 2006, pp. 460-476. [SWP # 4558-05, CISL 2005-08, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=825650]

[ZMS02] Zhu, H., Madnick, S. and Siegel, M., "The Interplay of Web Aggregation and Regulations" (with H. Zhu and M. Siegel), Proceedings of the IASTED International Conference on Law

and Technology (LAWTECH 2002), Cambridge, MA, November 6-8, 2002.

[ZMS04] Zhu, H. S. Madnick, M. Siegel, "Representation and Reasoning about Changing Semantics in Heterogeneous Data Sources,", *Proceedings of the International Workshop on Semantic Web and Databases* (SWDB), Toronto, Canada, 30 August - 3 September 2004; also published in Springer Lecture Notes in Computer Science (LNCS) 3372, C. Bussler et al (Eds.), pp.127-139, Springer-Verlag Berlin, 2005 [SWP #4497-04, CISL 2004-14, Center for eBusiness Working Paper #210, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=577242].

[ZMS04] Zhu, H., Madnick, S., and Siegel, M., "Effective Data Integration in the Presence of Temporal Semantic Conflicts," CISL Working Paper, 2004.