Improving National and Homeland Security through Context Knowledge Representation & Reasoning Technologies

Nazli Choucri
Stuart E. Madnick
Michael D. Siegel

Working Paper CISL# 2006-03

April 2006

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

Improving National and Homeland Security through Context Knowledge Representation & Reasoning Technologies

Nazli Choucri
Stuart E. Madnick
Michael D. Siegel

MIT Sloan School of Management
30 Wadsworth Street, Cambridge, MA 02142, USA
{nchoucri, smadnick, msiegel}@mit.edu

ABSTRACT

In the aftermath of the 9/11 tragedy it has became clear that the lack of effective information exchange among government agencies hindered the capability of identifying potential threats and preventing terrorism actions. It has been noted by the National Research Council that "*Although there are many private and public databases that contain information potentially relevant to counterterrorism programs, they lack the necessary context definitions (i.e., metadata) and access tools to enable interoperation with other databases and the extraction of meaningful and timely information*[1]". This report clearly recognized the important problem that the semantic data integration research community has been studying.

In this chapter, we describe the Laboratory for Information Globalization and Harmonization Technologies (LIGHT) developed at MIT. LIGHT arises from previous research, most notably the COntext INterchange (COIN) context mediation technology and the Global System for Sustainable Development (GSSD).

Context Mediation technology addresses the above problem and deals directly with the integration of heterogeneous contexts (i.e. data meaning) in a flexible, scalable and extensible environment. This approach makes it easier and more transparent for receivers (e.g., applications, sensors, users) to exploit distributed sources (e.g., databases, web, information repositories, sensors). In this paper we define context as the assumptions of the source and receiver that affect correct interpretation of the meaning of the information. Receivers are able to specify their desired context so that there will be no uncertainty in the interpretation of the information coming from heterogeneous sources. The COIN context knowledge representation approach and associated reasoning tools significantly reduce the overhead involved in the integration of multiple sources and simplifies maintenance in an environment of changing source and receiver context.

This technology is essential in the counter-terrorism environment in a number of areas including: (1) allowing for receivers (i.e., applications, analysts) to have multiple views of the same data (e.g., different semantic assumptions - two analysts may have a different meaning for Soviet Union depending on the application), (2) allowing for the collection of information into a single data warehouse, and (3) use in a dynamic federated environment where applications may have changing contexts and sources are added and removed from the grid. This approach is essential to the agile integration of information to support counter terrorism.

[1] Emphasis added

1. Introduction

1.1 Emergent Challenges to Effective Use of Information

The convergence of three distinct but interconnected trends – unrelenting globalization, rapidly changing global and regional strategic balances, and increasing knowledge intensity of economic activity – is creating critical new challenges to current modes of information access and understanding. First, the discovery and retrieval of relevant information has become a daunting task due to the sheer volume, scale, and scope of information on the Internet, its geographical dispersion, varying context, heterogeneous sources, and variable quality. Second, the opportunities presented by this transformation are shaping new demands for improved information generation, management, and analysis. Third, more specifically, the increasing diversity of Internet uses and users points to the importance of cultural and contextual dimensions of information and communication. There are significant opportunity costs associated with overlooking these challenges, potentially hindering both empirical analysis and theoretical inquiry so central to many scholarly disciplines, and their contributions to national policy. In this chapter, we identify new ways to address these challenges by significantly improving access to diverse, distributed, and disconnected sources of information.

1.2 National and Homeland Security

The information needs in the realm of national and homeland security involve emergent risks, threats of varying intensity, and uncertainties of potentially global scale and scope. Specifically, there is need to focus on: (a) crisis situations; (b) conflicts and war; and (c) anticipation, monitoring, and early warning. Information needs in these domains are extensive and vary depending on: (1) the *salience* of information (i.e. the criticality of the issue), (2) the *extent of customization*, and (3) the *complexity* at hand. More specifically, in:

- **Crisis situations:** the needs are characteristically immediate, usually highly customized, and generally require complex analysis, integration, and manipulation of information. International crises are now impinging more directly than ever before on national and homeland security, thus rendering the information needs and requirements even more pressing.
- **Conflicts and War**: the needs are not necessarily time-critical, are customized to a certain relevant extent, and involve a multifaceted examination of information. Increasingly, it appears that coordination of information access and analysis across a diverse set of players (or institutions) with differing needs and requirements (perhaps even mandates) is more the rule rather than the exception in cases of conflict and war.
- **Anticipation, Monitoring and Early Warning**: the needs tend to be gradual, but may involve extensive though routine searches and may require extraction of information from sources that may evolve and change over time. Furthermore, in today's global context, 'preventative action' take on new urgency, and create new demands for information services.

Illustrative Cases	Information Needs	Intended Use of Information
1. Strategic Requirements for Managing Cross-Border Pressures in a *Crisis* UNHCR needs to respond to the internal dislocation and external flows of large numbers of Afghans into neighboring countries, triggered by waves of post Soviet violence in Afghanistan.	Logistical and infrastructure information for setting up refugee camps, such as potential sites, sanitation, and potable water supplies. Also streamlined information on sabotage.	Facilitate coordination of relief agencies with up-to-date information during a crisis for more rapid response (as close to real time as possible). Reduce vulnerability to disruption.
2. Capabilities for Management during an Ongoing *Conflict & War* The UNEP-Balkans group needs to assess whether the Balkan conflicts have had significant environmental and economic impacts. Existing data is extensive, but highly dispersed, presented in different formats and prepared for different purposes.	Environmental and economic data on the region prior to the initiation/ escalation of the conflict. Comparison of this data with newly collected data to assess the impacts to environ-mental and economic viability.	Improved decision making during conflicts -- taking into account contending views and changing strategic conditions -- to prepare for and manage future develop-ments and anticipate the need for different modes of action.
3. Strategic Response to Security Threats for *Anticipation, Prevention, and Early Warning* The Department of Homeland Security needs to coordinate efforts with local government, private businesses and foreign governments using information from different regions of the world.	Intelligence data from foreign governments, non-governmental agencies, US agencies, and leading institutions on international strategy and security here and overseas .	Streamline potentially conflicting information content and sources in order to facilitate coherent interpretation, anticipation, preventive monitoring, and early warning.

Table 1. Illustrating Information Needs in Three Contexts

Table 1 illustrates the types of information needs required for effective research, education, decision-making, and policy analysis on a range of conflict issues. Indeed, "Critical central decisions should flow smoothly downward. Similarly, low-level urgent requests for communication, assistance, or information should flow upward to the appropriate agency and then back to the appropriate operatives." [NRC02 p.160] These issues remain central to matters of security in this increasingly globalized world.

1.3 Addressing Information Needs

1.3.1 Examples of Information Challenges

There are many important data elements critical to effective national and homeland security, such as place names, geographic locations, people names, and many others. All of these are subject to possibly confusion, especially when the information is gathered by many different agencies (possibly from different countries) using different procedures and different standards. Some examples are briefly illustrated below.

Airport naming: In addition to airport names themselves which are often written in different ways (e.g., "London airport," "London Heathrow Airport," "Heathrow Airport"), there are two major standards for codes designating airports: IATA and ICAO. An example of these differences is:

IATA	ICAO	LOCNAME	AIRPORTNAME	COUNTRY
LHR	EGLL	London	Heathrow	United Kingdom

City and country names: Is the city "Brussels" or "Bruxelle" or "Brussel"? It depends on whether it is being identified by a USA, French, or German source**.**

Geographic Co-ordinate Systems: Not only are there over 40 different geographic coordinate systems used around the world – there are even differences within the same governmental departments, such as within the US Department of Defense. The Army and Marine Corps use the Universal Transverse Mercator (UTM) Grid and Military Grid Reference System

(MGRS), while the US Navy uses latitude and longitude expressed in degrees, minutes and seconds. The Air Force uses latitude and longitude expressed in degrees and decimal degrees.[2]

People naming: Many problems exist in the identification of person by names in a database. For example, the name

قذافي

has been shown to have over 60 romanizations including: Gadaffi, Gaddafi, Gathafi, Kadafi, Kaddafi, Khadafy, Qadhafi, and Qathafi. There are numerous Romanization from Transliteration Standards. But different agencies may choose different standards. For example from Arabic to English, some examples of romanization standards are:

ALA-LC (library of Congress) 1972[3]
DIN 31636 – 198 (Germany)
EI (encyclopedia of Islam) 1960
ISO 233 –1984
UN 1972
USC – Transliteration of the Quran[4]

Many more: The above examples illustrate just a few of the challenges to using data effectively for national and homeland security.

1.3.2 Operational Example

For illustrative and simplification purposes only, let us consider the types of information illustrated by Example 2 in Table 1. A specific question is: **to what extent have economic performance and environmental conditions in Yugoslavia been affected by the conflicts in the region?** The answer could shape policy priorities for different national and international institutions, influence reconstruction strategies, and may even determine which agencies will be the leading players. Moreover, there are potentials for resumed violence and the region's relevance to overall European stability remains central to the US national interest. This is not an isolated case but one that illustrates concurrent challenges for information compilation, analysis, and interpretation – under changing strategic 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 consider 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.

---

[2] From http://www.findarticles.com/p/articles/mi_m0IAU/is_1_8/ai_98123571
[3] See http://www.loc.gov/catdir/cpso/romanization/arabic.pdf
[4] See http://www.usc.edu/dept/MSA/quran/transliteration/table.html

In this simple case, we suppose that the 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 2). 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?**

1.3.2.1 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 2. 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, to be explained in more detail later.)

The purpose of Table 2 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 2 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.

1.3.2.2 Manual Approach

Given the types of data shown in Table 2, along with the appropriate context knowledge (some of which is shown in italics) an analyst could determine the answer to our question. The proper calculation involves numerous steps, including selecting the necessary sources, making the appropriate conversions, and using the correct calculations. For example:

For time T0:

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 2).

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 2).

<table>
<tr><th>Domain and Sources Consulted</th><th>Sample Data Available</th><th>Basic Question, Information User Type & Usage</th></tr>
<tr><td><u>Economic Performance</u>
• World Bank's World Development Indicators database
• UN Statistics Division's database
• Statistics Bureaus of individual counties</td>
<td><u>A. Annual GDP and Population Data:</u>
<table>
<tr><th>Country</th><th>T0.GDP</th><th>T0.Pop</th><th>T1.GDP</th><th>T1.Pop</th></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"><u>Question</u>:
How did economic output and environmental conditions change in YUG over time?

User 1: YUG as a geographic region bounded at T0:
<table>
<tr><th>Parameter</th><th>T0</th><th>T1</th></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><th>Parameter</th><th>T0</th><th>T1</th></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><u>Environmental Impacts</u>
• Oak Ridge National Laboratory's CDIAC database
• WRI database
• GSSD
• EPA of individual countries</td>
<td><u>B. Emissions Data:</u>
<table>
<tr><th>Country</th><th>T0</th><th>T1</th></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><u>Country History:</u>
• 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><u>Mappings Defined:</u>
• 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><th>Country</th><th>Code</th><th>Currency</th><th>Currency Code</th></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><th>C_From</th><th>C_To</th><th>T0</th><th>T1</th></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 2. Operational Example: Information Needs in Cases of Conflict

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 utilization usually involve complexities of this sort.

1.3.3 LIGHT: A Better Way

With reference to national and homeland security concerns, a NRC study states: "*Different emergency responders must be able to communicate with each other, but poor interoperability among responding agencies is a well-known problem . . . The fundamental technical issue is that different agencies have different systems, different frequencies and waveforms, different protocols, different databases, and different equipment.*" [NRC02, p.159]. A key goal of the MIT Laboratory for Information Globalization and Harmonization Technologies (LIGHT) is to automatically determine and reliably perform the steps shown above in response to each user's request. Every user is distinct. LIGHT will be capable of storing the necessary context information about the sources and users – and have a reasoning engine capable of determining the sources, conversions, and calculations necessary to meet each user's needs. The COIN and GSSD systems, to be described briefly below, have proven the feasibility of this approach in more limited situations. LIGHT is the next generation: it will combine context and content.

1.4 Existing Foundations – COIN and GSSD

Important research in two areas has already been completed that provides essential foundations for addressing the emergent and pressing challenges discussed above: the *COntext INterchange* Project (COIN) and the *Global System for Sustainable Development* (GSSD).

1.4.1 COIN

The *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]. We utilize the foundation of COIN to develop theories and methodologies for the new System for Harmonized Information Processing (SHIP). 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 (i.e. through the use of a CDS or collections of linked CDSs) include 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') context definition and source definitions to be added and removed easily (i.e., schemas, contexts, capabilities).

1.4.2 GSSD

The *Global System for Sustainable Development* serves as an Internet-based platform for exploring the contents transmitted through different forms of information access, provision, and integration across multiple information sources, languages, cultural contexts, and ontologies. GSSD has an extensive, quality-controlled set of ontologies related to system sustainability (specifically, to sources of instability and alternative responses and actions), with reference to a large set of specific domains related to the field of international relations. In addition, GSSD has made considerable gains into understanding and undertaking the organization and management of large scale, distributed, and diverse research teams, including cross-national (China and Japan, and countries in the Middle East and Europe) and institutional partners (private, public, and international agencies). Designed and implemented by social scientists, GSSD is seen as demonstrating 'opportunities for collaboration and new technologies,' according to the National Academy of Engineering [RAC01, p. viii]. GSSD databases cover issues related to dynamics of conflict, as well as other domains relevant to our proposed research, such as population, migration, refugees, unmet human needs, as well as evolving efforts at strategic and coordinated international actions. (As an example, for 'population' see [Cho99:280-282].) GSSD provides a rich ground for the technologies, including automated methods for information aggregation from various sources, context mediation capabilities, customized information retrieval capabilities, and ontology representations.

2. IT Theory and Technology Research

2.1 Needs for Harmonized Information Processing and Collaborative Domain Spaces

Advances in computing and networking technologies now allow extensive volumes of data to be gathered, organized, and shared on an unprecedented scale and scope. Unfortunately, these newfound capabilities by themselves are only marginally useful if the information cannot be easily **extracted** and **gathered** from **disparate sources**, if the information is represented with **different interpretations**, and if it must satisfy **differing user needs** [MHR00, MAD99, CFM*01]. The data requirements (e.g., scope, timing) and the sources of the data (e.g., government, industry, global organizations) are extremely diverse. National and homeland security, by definition, take into account internal as well as external dimensions of relations among actors in both the public and the private domains.

It is necessary to:

1. Analyze the data and technology requirements for the categories of problems described in Section 1;
2. Research, design, develop and test extensions and improvements to the underlying COIN and GSSD theory and components;
3. Provide a scalable, flexible platform for servicing the range of applications described in Section 1; and
4. Demonstrate the effectiveness of the theories, tools, and methodologies through technology transfer to other collaborating organizations.

2.2 Illustrative Example of Information Extraction, Dissemination, and Interpretation Challenges

As an illustration of the problems created by information disparities, let us refer back to the example introduced in Section 1.3. The question was**: what are the impacts of CO_2 emissions on economic performance in Yugoslavia.** It is necessary to draw data from diverse sources such as CIA Worldbook (for current boundaries), World Resources Institute (for CO_2 emissions), and the World Bank (for economic data). There are many additional information challenges that had not been explicitly noted earlier, such as:

Information Extraction: Some of the sources may be full relational databases, in which case there is the issue of remote access. In many other cases, the sources may be traditional HTML web sites, which are fine for viewing from a browser but not effective for combining data or performing calculations (other than manually "cut & paste"). Other sources might be tables in a text file, Word document, or even a spreadsheet. Although the increasing use of eXtensible Markup Language (XML) will reduce some of these interchange problems [MAD01], we will continue to live in a very heterogeneous world for quite a while to come. So we must be able to extract information from all types of sources.

Information Dissemination: Different users want the resulting "answers" expressed in different ways. Some will want to see the desired information displayed in their web browser but others might want the answers to be deposited into a database, spreadsheet, XML document, or application program for further processing.

Information Interpretation: Although the problems of information extraction and dissemination will be addressed in this research, the most difficult challenges involve information interpretation. Specifically, an example question is: "What is the change of CO_2 emissions per GDP in Yugoslavia before and after the Balkans war?"

Before the war (time T0), the entire region was one country. Data for CO_2 emissions was in thousands of tons/year, and GDP was in billions of Yugoslavian Dinars. *After the war* (time T1), Yugoslavia only has two of its original five provinces; the other three provinces are now four independent countries, each with its own currency. The size and population of the country, now known as Yugoslavia, has changed. Even Yugoslavia has introduced a new currency to combat hyperinflation.

From the perspective of any one agency, UNEP for example, the question: "How have CO_2 emissions per GDP changed in Yugoslavia after the war?" may have multiple interpretations. Not only does each source have its own context, but so does each user (also referred to as a receiver). For example, does the user mean Yugoslavia as the original geographic area (depicted as *user 1* in Table 2) or as the legal entity, which has changed size (*user 2*). To answer the question correctly, we have to use the changing context information. A simple calculation based on the "raw" data will not give the right answer. As seen earlier, the calculation will involve many steps, including selecting necessary sources, making appropriate conversions, and using correct calculations. Furthermore, each receiver context may require data expressed in different ways, such as: tons/million USD or kilograms/billion Euro.

Although seemingly simple, this example addresses some of the most complex issues: namely the impact of changing legal jurisdictions and sovereignties on (a) state performance, (b) salience of socio-political stress, (c) demographic shifts and (d) estimates of economic activity, as critical variables of note. Extending this example to the case of the former Soviet Republics, before and after independence, is conceptually the same type of challenge – with greater complexity. For example, the US Department of Defense may be interested in demographic distributions (by ethnic group) around oil fields and before and after independence. Alternatively, UNEP may be interested in CO_2 emissions per capita from oil-producing regions. Foreign investors, however, may be interested in insurance rates before and after independence. The fact that the demise of the Soviet Union led to the creation of a large number of independent and highly diverse states is a reminder that the Yugoslavia example is far from unique. It highlights a class of increasingly complex information reconciliation problems. Many of the new states in Central Asia may also rank high as potential targets and bases for global terrorism.

The information shown in italics in Table 2 (e.g., "population in millions") illustrates **context knowledge**. Sometimes this context knowledge is explicitly provided with the source data (but still must be accessed and processed), but often it must be found from other sources. The good news is that such context knowledge almost always exists, though widely **distributed** within and across organizations. Thus, a central requirement is **the acquisition, organization, and effective intelligent usage of distributed context knowledge to support information harmonization and collaborative domains**[5].

2.3 Research Platform

The MIT COntext INterchange (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 1).

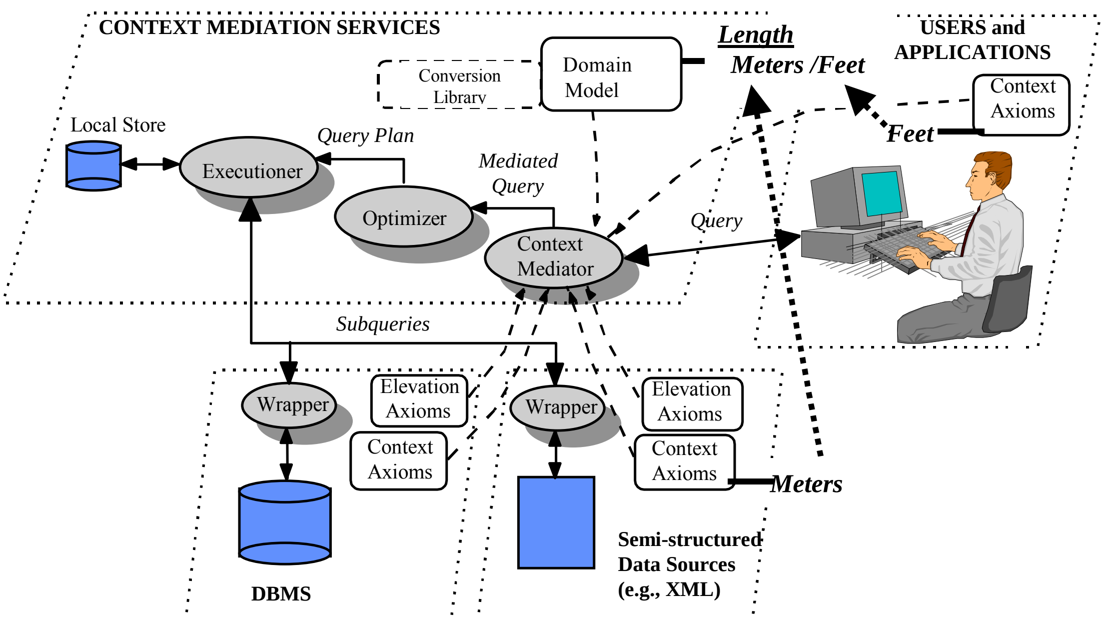

Figure 1. The Architecture of the Context Interchange System

[5] See http://www.gao.gov/new.items/d03322.pdf for types of information central to national and homeland security; and the functionalities listed in http://www.dhs.gov/dhspublic/ for the range of some domain-specific information needs.

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 reasoning procedure 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.

2.4. Advances in Integrating Systems and Data Involving Complex and Interdependent Social Systems

There are a number of important advances demonstrated by the COIN and GSSD efforts and the emerging LIGHT that builds on them. Several of these key advances are described below.

1. Extended Domain of Knowledge – Equational Context. In addition to the representational context knowledge currently handled by the original COIN framework, there was need to add capabilities for both the representation and reasoning to provide support for equational [FGM02] context. Equational context refers to the knowledge such as "average GDP per person (AGDP)" means "total GDP" divided by "population." In some data sources, AGDP explicitly exists (possibly with differing names and in differing units), but in other cases it may not explicitly exist but could be calculated by using "total GDP" and "population" from one or more sources – if that knowledge existed and was used effectively. The original COIN design has been extended to exploit simultaneous symbolic equation solving techniques through the use of Constraint Handling Rules (CHR) [Früh98], a high-level language extension of constraint logic programming (CLP). This extension, coupled with our context based reasoning approach to detecting and reconciling data semantics, provides an elegant and powerful solution to the problem of detecting and resolving equational conflicts. This combines the advantages of logic programming and constraint solving by providing a declarative approach to solving problems, while at the same allowing users to employ special purpose algorithms in the sub problems [FMG02].

2. Extended Domain of Knowledge – Temporal Context. Temporal context refers to variations in context not only across sources but also over time. Thus, the implied currency for France's GDP prior to 2002 might be French Francs, but after 2002 it is Euros. If one were performing a longitudinal study over multiple years from multiple sources, it is essential that variation in context over time be understood and processed appropriately. A seemingly straightforward variable, like the size of 'military expenditures' across countries, is defined differently depending on the rules of inclusion or exclusion (for example, military pensions) used in different jurisdictions. Changes in territorial boundaries signal changes in jurisdiction, and often changes in modes of information provision and formatting. This is a common problem facing a new government after a revolution. The COIN context knowledge representation has been augmented to include a specification of the history of all contextual attributes in the ontology [ZMS04]. Mathematically, it is set of <*contextual_attribute, history*> pairs, where history is a set of <*value, valid_interval*> pairs. Then temporal reasoning can be treated as a constraint solving problem, using constraint handling rules similar to [Früh94].

3. Extended Domain of Knowledge – Entity Aggregation Context. Entity aggregation addresses the reality that we often have multiple interpretations of what constitutes an entity. We have already seen that example in the multiple interpretations of what is meant by "Yugoslavia." This situation occurs even in domestic cases, such as does "IBM" include "Lotus Development Corp" (a wholly-owned subsidiary)? The frequent answer is "depends on the context." We have defined this problem as "corporate householding"[MWZ02]. This is also a common occurrence and challenge in many aspects of national and homeland security. Corporate householding entity aggregation problems are very similar to traditional COIN applications in the sense that entity aggregation also involves different source and receiver contexts. Under different contexts, an entity may or may not need to be aggregated. The semantic types in the ontology can be divided into two categories – structure related and task related. Structure related semantic types represent common concepts in organizational structure and entity aggregation, and thus are useful in any entity aggregation problems; the task related semantic types are specific to particular applications. The COIN reasoning process has been extended to comprehend the general semantics of the organization hierarchies that must be navigated [MWX03].

4. Linked Collaborative Domain Spaces. The original COIN framework provided representation and reasoning capabilities for a single domain. Although there are a number of ontology library systems that allow for management of multiple ontologies [DSW*99, DFen01 Fensel01, HelfH00], they have limitations in scalability and dynamically incorporating new ontological knowledge. Especially, they lack the capability of representing rich context knowledge needed for reconciling differences among sources. A primary need is the ability to operate in a multi-disciplinary environment across multiple linked collaborative domain spaces. The representational capabilities to relate concepts across domains, and efficiently maintain the effectiveness of these collaborative domain spaces is critically important – especially in an environment where we believe the underlying domains themselves will continually undergo evolution. For some users, the reality of domain shifts itself is the defining feature of interest [Nuna01] [Kal03].

5. Advanced Mediation Reasoning and Services. The COIN abductive logic framework can also be extrapolated to problem areas such as integrity management, view updates and intentional updates for databases [Chu00]. Because of the clear separation between the generic abductive procedure for query mediation and the declarative logical definition of domain models and source and receiver contexts, we are able to adapt our mediation procedure to new situations such as mediated consistency management across disparate sources, mediated update management of one or more database using heterogeneous external auxiliary information, or mediated monitoring of changes. An update asserts that certain data objects must be made to have certain values in the updater's context. By combining the update assertions with the COIN logical formulation of context semantics, we can determine whether the update is unambiguous and feasible in the target context, and if so, what source data updates must be made to achieve the intended results. If ambiguous or otherwise infeasible, the logical representation may be able to indicate what additional constraints would clarify the updater's intention sufficiently for the update to proceed. We build upon the formal system underlying our framework, abductive reasoning, and extend the expressiveness and the reasoning capabilities leveraging ideas developed in different yet similar frameworks such as Description Logic and classification, as well as ongoing in Semantic Web research. These national security applications, where there are fundamental shifts in relationships, systems, and pressures, is a 'tough test' since the underlying domain is highly dynamic even volatile.

6. Automatic Source Selection. A natural extension is to leverage context knowledge to achieve context-based automatic source selection. One particular kind of context knowledge useful to enable automatic source selection is the content scope of data sources. Data sources differ either significantly or subtly in their coverage scopes. In a highly diverse environment with hundreds and thousands of data sources, differences of content scopes can be valuably used to facilitate effective and efficient data source selection. Integrity constraints in COINL and the consistency checking component of the abductive procedure provide the basic ingredients to characterize the scope of information available from each source, to efficiently rule out irrelevant data sources and thereby speed up the selection process [TM98]. For example, a query requesting information about *companies with assets lower than $2 million* can avoid accessing a particular source based on knowledge of integrity constraints stating that *the source only reports information about companies listed in the New York Stock Exchange* (NYSE), and that *companies must have assets larger than $10 million to be listed in the NYSE*. In general, integrity constraints express necessary conditions imposed on data. However, more generally, a notion of completeness degree of the domain of the source with respect to the constraint captures a richer semantic information and allows more powerful source selection. For instance, a source could contain exactly or at least all the data verifying the constraint (e.g., all the companies listed in the NYSE are reported in the source). The source may be influenced by institutional objectives, resulting in major differences in metrics (for concepts like 'terrorism') due to differences in definitions of the concept itself. In cases of violent conflict, casualty reports vary significantly largely because of differences in definitions of the variable (i.e., who is being counted).

7. Source Quality. Not only do the sources vary in semantic meaning, they also vary in quality, and they do so in various ways. We must be able to represent and reason about the quality attributes of the sources [WKM93], [Mad03].

8. Attribution Knowledge Processing. For quality assessment and other reasons, it is important to know the attribution of the sources [LCN*99, LMB98]. For example, it can be important to know that although three different sources agree on a controversial piece of the information (e.g., casualties in the Afghanistan war), all three sources acquired that information from the same, maybe questionable, origin source [Lee02].

9. Domain Knowledge Processing – Improving Computer Performance. While domain and context knowledge processing has been shown to have considerable conceptual value [CZ98, MBM*98, LMS96b, SW92], its application in real situations requires both efficiency and scalability across large numbers of sources, quantities and kinds of data, and demand for services. The scalability and optimization of this mediation processing for large numbers of sources across multiple collaborative domains and contexts is important. In a heterogeneous and distributed environment, the mediator transforms a query written in terms known in the user or application program context (i.e., according to the user's or program's assumptions and knowledge) into one or more queries in terms of component sources. Individual subqueries at this stage may involve one or multiple sources. Subsequent planning, optimization and execution phases [AKS96, Fynn97] take into account the limitations of the sources and the topology and costs of the network (especially when dealing with non-database sources, such as web pages or web services). The execution phase schedules execution of steps in the query execution plan and the realization of the integrative operations not handled by the sources individually (e.g. a join across sources) [Tar02].

10. Domain Knowledge Acquisition – Improving Human Performance. Domain and context knowledge acquisition are also essential. One critical property to be emphasized is the independence of domains and sources. The COIN approach is non intrusive and respects source and receiver independence (i.e. autonomy). To effectively use the expressive power of the constructs and mechanisms in COIN, it is important that subject matter domain experts be able to easily provide the needed domain and context knowledge. It is therefore essential to have an appropriate flexible methodology and tools supporting this methodology. Where a large number of independent information sources are accessed (as is now possible with the global Internet), flexibility, scalability, and non-intrusiveness will be of primary importance. Traditional tight-coupling approaches to semantic interoperability rely on the *a priori* creation of federated views on the heterogeneous information sources. These approaches do not scale-up efficiently or reliably given the complexity involved in constructing and maintaining a shared schema for a large number of possibly independently managed and evolving sources. Loose-coupling approaches rely on the user's intimate knowledge of the semantic conflicts between the sources and the conflict resolution procedures. This reliance becomes a drawback for scalability when this knowledge grows and changes as more sources join the system and when sources are changing. The COIN approach is a middle ground between these two approaches. It allows queries to the sources to be mediated, i.e. semantic conflicts to be identified and solved by a context mediator through comparison of contexts associated with the sources and receivers concerned by the queries. It only requires the minimum adoption of a common Domain Model, such as that developed for GSSD, that defines the domain of discourse of the application [Lee03].

11. Relationship with Evolving Semantic Web. Although the initial COIN and GSSD research and theories preceded the emerging activities now described as the Semantic Web, there are many areas of overlap, especially involving the development of the OWL ontology standards and the use of rules and reasoning. The LIGHT research contributes to the maturing of the Semantic Web and, at the same time, LIGHT exploits relevant ontologies, standards and tools that are emerging from the Semantic Web activities.

12. Operational System for Harmonized Information Processing. A key development is the new System for Harmonized Information Processing (SHIP), a distributed information infrastructure that will be used to support the types of challenges listed in Section 1, incorporating all the components identified above. This system has maximum flexibility and extensibility that permits new and existing applications to seamlessly extract data from an array of changing heterogeneous sources. The utility of many data bases in the national priority areas has been seriously constrained by the difficulties of reconciling known disparities and conflicts within and across sources. SHIP directly addresses this problem. (Data reconciliation itself has become an important focus of scholarly inquiry in various parts in political science, as recognized by the NSF).

13. Policy Implications Regarding Data Use and Re-use. There are widely differing views regarding the use and re-use of even publicly available information. In particular, the USA has taken a largely "laissez faire" approach whereas the European Union is pursuing a much more restrictive policy (as embodied in its "Data Base Directive"). We have been applying principles from the domain of economics to develop a more scientific approach to studying and evaluating the current and proposed policies and legislation in this area [ZMS02].

3. National and Homeland Security

National and homeland security (NHS) is an important research area. In this section, we describe some of the most fundamental barriers to the reliable use of information systems in this area. *Our goal is to reduce serious barriers, enhance understanding and meaning across substance, topics, and ontologies, and provide new tools for national security analysis in international relations (IR) research.* For example, data on incidences of conflict and war are available on the web sites of a wide range of institutions with different capabilities and objectives, such as the US Department of State, SIPRI in Sweden, the UN HCR, the Correlates of War Project.[6] Despite all this information, we cannot compute the 'actual' number of deaths and casualties in a conflict – at one point in time, over time, and as the contenders change and reconfigure their own jurisdictions – largely due to differences in definitions of key variables. These are typical questions that have plagued researchers, as far back as 1942, with classics in the field such as Quincy Wright's *A Study of War,* [Wri65] and even earlier, with Lewis Fry Richardson's *Statistics of Deadly Quarrels* (1917) [Rich60].

3.1 Pressing Demands on Information Systems

The proliferation of new actors on the international landscape (i.e. new states, non-governmental organizations, cross-border political groups, non-state actors, international institutions, global firms, etc.) reflects diverse perspectives, creates new sources of data, legacy problems, and new difficulties for access, interpretation and management. A persistent challenge to national security is to reduce the **distinction between reality and representation**. Reality is the empirical domain and is the referent of representations. Representations (ontologies) are idealized frameworks that identify salient aspects of reality and allow us to organize and manipulate them as information. The properties of the database scheme or application ontology define the domain of analysis, types of inferences, and nature of conclusions drawn. While representations are the interface to reality, organizations take action in reality. To date, efforts to

[6] http://www.pcr.uu.se/ research/UCDP/ conflict_dataset_catalog/data_list.htm

address the problem of domain-specific representation in international relations remain costly and time consuming, yet acting without them may be even more costly – or simply impossible.

Indeed, an often cited recent review of empirical challenges in a noteworthy issue of *International Political Science Review* (2001), devoted to "Transformation of International Relations – Between Change and Continuity" arguing that "reconfiguration of the founding concepts of international relations … is linked to important paradigmatic changes" [Sind01, p. 224] and that state-centric modes of analysis and information configuration must be augmented by methods that help capture changes in both structure and process in the international arena. This is one of the major challenges in the new domain of inquiry, termed CyberPolitics, as noted in the *International Political Science Review* (2000) issue "CyberPolitics in International Relations" [Cho00, Cho99] which identifies new directions of research, research priorities, and critical next steps.

3.2 Defining the Research Problem: The Paradox of Plenty

While there exists no 'single authoritative view' of the international relations field as a whole, Katzenstien, Keohane, and Krasner, eds. [KKK99], illustrate dominant trends in the non-quantitative aspects of the field. By contrast, in quantitative international politics (QIP), theory development and analysis is more data-driven and thus invariably more vulnerable to limitations of information systems. Earlier quantitative works, such as Hoole and Zinnes [HZ76] and Russett [Russ72], as well as the more recent advances by Levy [Levy89], Choucri and North [ChoN93], Choucri, North and Yamakage [ChoNY92], and Pollins and Schweller [SP99], illustrate the general progression in the field and the persistent data representation problems. Concurrently, [Alk96] highlighted some analogous and fundamental challenges to humanistic approaches to international studies, illustrated by ranges of computer-assisted applications. Further, in the issue of *International Studies Quarterly* [CR96] devoted to evolutionary perspectives in international relations, leading scholars such as George Modelski, Robert Gilpin, Cioffi-Revilla, and others, articulated the importance of transformation and adaptation over time, as an important departure from the common focus on discrete events, or retrospective case-based interpretation, so dominant in the field. By far the most succinct statement about data reconciliation problems is made by a leading scholar who proceeds to demonstrate in considerable detail the "semantic carelessness … [that can] stand in the way of cumulative research" and then identifies a large set of specific examples that may be particular to international relations, but "most seem to be found all across the discipline [of political science]." [Singer, 2001:604]

The *Paradox of Plenty* is this. Despite the *abundance* of existing data and information, there is a *paucity* in the consistency, reliability, and connectivity of the information. For example, in the conflict theory and analysis domain, advances in the long tradition of tracking wars and casualties have been severely hampered by the difficulties of generating an integrated approach to diverse information resources, drawing upon large scale collaborative efforts in the profession and undertaken by a large number of research groups, nationally and internationally. The same point holds for the cooperation theory domain where, for example, efforts to measure "regime formation" and "compliance" in a wide range of specific issue-areas are hampered by the diversity of ontologies, data meanings, metrics and methods.

3.3 Context Mediation for National Security

Increasingly, the nation's intelligence agencies rely on information from all over the world to anticipate, identify, and develop strategic responses to security threats. As noted in

[NRC02, p.304]: "Although there are many private and public databases that contain information potentially relevant to counter terrorism programs, they lack the necessary context definitions (i.e., metadata) and access tools to enable interoperation with other databases and the extraction of meaningful and timely information." The tragic events of 9/11/2001 starkly indicate how changes in the scale, scope, type, and intensity of external threats to national security is surpassing existing practices in information access, interpretation, and utilization -- in both the scientific and policy-making communities.

The *Paradox of Plenty* is amply demonstrated by the large number of data sets compiled by international relations scholars on conflict, crises and war that are now found in central repositories such as the InterUniversity Consortium for Political Science Research (ICPSR), the Harvard MIT-Data Center, and others. Despite decades of painstaking research, cumulativeness remains hampered by barriers to information reconciliation. There are no mechanisms for extracting coherent and integrated information from these data sets, since the variables are defined differently, the formatting varies, content is represented in different forms and updated variously. It is nearly impossible to utilize these sets for purposes other than those intended by the initial compilers, and it is even more difficult to merge, streamline, or normalize. The NSF sponsored Data Documentation Initiative (DDI) offers the prospect of formal XML-based documentation of the coding and structure of social science data sets. The Context Mediation research draws on DDI results and enables information extraction and fusion in a collaborative environment hitherto unreachable[7].

For example, among the most notable data sets of the *Correlates of War Project*, a highly respected and well-structured data set, wars are reported in dyads, i.e. country X - country Y. Data are reported by war-months, for the warring dyads, devoid of context, which means that we cannot determine if it was an offensive or defensive war, or readily extract other salient features of the "situation." These problems could be reduced if systematic comparisons could be made with relevant information from other data sets (such as the CIA Factbook and the Uppsala Conflict Database). Achieving this integration of data sets on attributes and activities of states over time requires the ability to reconcile different coding schemes representing states as well as the ability to track and integrate the impacts of changes in territorial and jurisdictional boundaries (using, for example, the Uppsala Territorial Change data set). Working from the opposite direction, the CASCON research [BM97] developed a set of policy relevant factors relating to the potential for violence in conflict situations, but requires laborious hand coding of each new conflict that arises. With the context mediation technology, it should be possible to connect many of these factors to available data sources and thereby enable fact patterns to be readily filled in so that the method can be more readily applied to supporting the policy analytic process.

These are the challenges that are being addressed with the development of the next generation of context mediation technologies in LIGHT. New technologies cannot alter shifting realities, but they can provide functionalities to reduce barriers to information access, use, re-use, customization and interpretation.

3.4 Research Design in Practice – Approach, Test-Applications, Implementation

3.4.1 Approach

This research is based on the structural differentiation among *contextual* conditions, and

[7] see http://www.icpsr.umich.edu/DDI/index.html

on the *type of gap* between the variable of interest, the *referent* (such as actor, issue, institution, etc.) and the information-system and its properties, the *representation.* The goal is to reduce the gap between the two and increase the representation power of the information systems. Toward this end, we address the *context* of content develop specific classes of tools to represent *context-types,* and approach these computationally through test applications. For each of the applications, we focus on (i) properties of the *context*-situation; (ii) properties of the *data features*, and (iii) properties of the *data collection agencies*.

3.4.2 Example Applications

Our focus is on the 'tough cases', i.e. reducing barriers to information access and use when the *properties of the problem* themselves are changing as a function of *unfolding* conflicts and contentions, and when the *demands* for information change in the course of the contentions. This includes three sets of applications selected because of their known and powerful impediment to national security analysis. (Each of these context-problems has some similarities with the Balkans example earlier, but each highlights added complexities).

(1) Shifts in Spatial Configuration – e.g. the territorial boundaries problem. As any student of international relations knows, the dissolution of the Soviet Union is a major, but far from unique reconfiguration of territorial boundaries. Several data bases seek to capture these changes, and below we refer to one such example with cases spanning well over one century (1816-1996).

(2) Disconnects in Definitions of 'Conflict' – e.g. the wars and casualties problem. Of the leading 10 data sets on international conflict and violence over time, no two data sets are synchronized or reconciled (see below for two examples).,

(3) Distortions due to Data Temporality – e.g. economic and political 'currency' problem. The ongoing experiment in Europe on the formal shift from national currencies to the Euro must be addressed if we are to ask: How extensive are the individual countries' investments in their military systems compared to each other, to the US, and to past commitments?

3.4.3 Implementation & Examples

To deploy the technical work put forth in Section 2 toward solving specific problems in the NHS domain, we proceed in the following steps (with of a degree of overlap as needed): (1) identify the referent situations, such as shifts in the Balkan countries' boundaries, war casualties in region X, or US troop casualties over the past X years, (2) create the case-catalogue, i.e. in such cases, list of all spatial reconfigurations over the past 20 years, and verify the degree of congruence among alternative sources for representing the shifts, (3) identify the similarities and differences between the variable definitions of the problem in various information systems or relevant data bases and compare these to the topic and/or domain specific ontology in GSSD, (4) use the results to design context features for computational purposes of new context mediation tools, (5) construct the pilot study for the case in point, (6) test viability of specifications against at least three different information systems or data bases (see below), and on this basis, (7) make adjustments, changes, etc. and, (8) undertake the actual test-application

To illustrate parts of the design, we refer below to application Case 2, namely, international conflict and war, so fundamental to the nation's security. For example, the *Correlates of War Project* (COW) and the Project on *Assessing Soceital and Systemic Impact of Warfare (SSIW)*, both deal with deaths due to violence and hostility, but they define war (terms and categories) in different ways: COW defines war as "sustained armed combat between two or more state member of the international system which meets the violence threshold", and uses 1,000 battle-related fatalities as the threshold, with no fixed time within which these deaths must occur, and proceed to differentiate between intra-state war, interstate war, and extra-state war (each defined specifically). ASSW develops a 10-point scale for assessing magnitude, intensity, and severity of war, differentiating among interstate warfare, wars of independence, civil warfare, ethnic warfare, and genocide.) In the absence of a common frame of reference spanning these two information systems it is extremely difficult to get a sense of what in fact may have taken place (i.e. clarifying the 'dependent' variable as a necessary precursor to statistical, simulation, modeling or policy analysis of any type.) For this reason, we use the ontology for the 'conflict and war' domain developed for GSSD as our platform, to provide the base line for developing the new operational ontology. This latter task, of course, is guided by the dominant theories of conflict and war in international relations[8].

At the same time, however, we know from historical and situational analysis that the very act of war (variously defined) is often preceded by, or results in, territorial shifts in legal political jurisdiction. This means that (a) reconciliation of definitions is only the first step; (b) accounting for spatial reconfiguration is a necessary next step. Both steps must be completed before we can address the question of 'how many casualties? Interestingly, the *Territorial Change Coding Manual*, showing the different dimensions across which spatial changes are coded, notes that these include "at least one nation-state" of the COW information system, and then identifies six specific procedures by which special changes take place (conquest, annexation, cession, secession, unification, mandated territory) – and as any international lawyer knows, these are contentious conditions.

The current *information base* for the GSSD platform consists of web based resources from over 250 institutions worldwide, representing a diverse set of data sets by type, scale and scope that is then cross-referenced and cross-indexed for ease of retrieval and analysis, according to an integrated and coherent conceptual framework covering the knowledge domain [Cho01]. The domain consists of a hierarchical and nested representation spanning 14 key socio-economic 'sectors' of human activities, attendant known problems to date related to each, responses to these problems, in terms of scientific and technological activities, social and regulatory instruments, as well as modes of international collaboration. GSSD was chosen as our platform because it: (1) provides a *domain-specific ontology* based on rigorous applications of social science theories, and related domains in science and technology, (2) offers practical reasoning rules for forming additional ontologies, (3) presents scenarios for broad applications of the new technologies to be developed in this project, (4) regularly updates its representation of, and links to, large and important set of information sources, and (5) spans local and global data information sources.

3.5 Generalizing the Research Tasks

To illustrate specific aspects of the research design, we note two key issues:

3.5.1 Comprehensive information-base survey.

First, is to more fully understand attributes of the data types in the GSSD knowledge base that are relevant to the specific domain selected for a test-application. The outcomes of this phase include: (a) an assessment of the context of data types within the domain, including the following aspects: data source, format, organization, equational and temporality attributes, provision rules, and utility for user-driven query; and (b) typologies of barriers to access, noted above.

[8] *See* "Using GSSD- GSSD Knowledge Strategy" at http://gssd.mit.edu/GSSD/gssden.nsf

3.5.2 Multi-disciplinary and distributed user survey for the test-applications

Second, is to develop and apply methods to survey current and future information demands from diverse NHS actors, differentiated in terms of (i) data users, (ii) data providers, and (iii) data intermediaries (or brokers). Test cases to capture the impacts and represent the views of different user types on information and data needs will emerge from this assessment. Specific activities include:

(a) **Multi-dimensional assessments of information demand** from different user types within the diverse conflict domains noted earlier (e.g. sections 1.2.1 and 1.3), based on surveys, workshops, and in-depth interviews.

(b) **Development of new or refined ontologies and a knowledge repository** to represent specific NHS domains and provide a test bed for the emergent information technologies.

(c) **Refined substantive applications of the new technologies for enhancing information capabilities in theory and methods development, and results of tests for effectiveness of the design.** This would demonstrate the performance of the technologies' domain specific and practical applications test cases, and to generate some guidelines of relevance for similarly complex domains.

(d) **Collaborative assessments** and evaluations of the technologies' effectiveness to address NHS information issues and LIGHT's capacity for scalability and cross-domain applicability.

4. Laboratory for Information Globalization and Harmonization Technologies

The **Laboratory for Information Globalization and Harmonization Technologies (LIGHT)** has been established to address the strategy, application, development and deployment of this next generation of intelligent information technologies that are designed to support the national priority areas. Its purpose is to examine 'frontier' issues, such as transformations in patterns of conflict and cooperation, changes in modes of international business, emergent dimensions of globalization and system change, negotiation systems for new global accords, among others. In addition, LIGHT hosts the technical infrastructure of our System for Harmonized Information Processing (SHIP).

In practice, this multidisciplinary Laboratory brings together faculty and students with interdisciplinary interests and activities from a number of departments of MIT, including Information Technologies, Political Science, Management Science, and the Technology, Management and Policy program.

More specifically, the Laboratory is the central entity for developments in four areas: (1) Software Platforms, (2) Knowledge Repositories, (3) Application Demonstrations, and (4) Education and Research. The software platforms include, but not limited to, SHIP with Collaborative Domains Spaces (CDS) including one or more Ontology Library Systems, Context and Conversion Management Systems, Context Mediation Engine, Execution and Planning Module, and Application and Source Support Tools. The Knowledge Context Repositories include the NHS domain specific knowledge represented in ontologies, context and conversion libraries, source schemas and capabilities. The Application Demonstrations are being developed at MIT, with the participation of collaborators, nationally and internationally.

5. Conclusions

The LIGHT project, building on the COIN and GDSS systems, will lead to major advances in information technology applicable to the national priority areas. The outcomes of this innovative project address many of the challenges facing our nation:

1. Theory and Technology. The LIGHT System for Harmonization of Information Processing (SHIP) provides an effective mechanism for effective and meaningful information interchange among very large scale (in terms of size and geographical locations) and diversified (in terms of media, schemas, and domains) systems. The reliability of systems is significantly improved by dynamically incorporating semantically equivalent sources into the interconnected system. It allows new applications to be built quickly to facilitate information sharing among diverse groups of people, devices, and software systems. Since it facilitates semantic level information interchange, any information receiver (people, devices, or software) can obtain customized information accurately and in a form and meaning that the receiver prefers.

2. National Priorities. This effort significantly augments the effective use of information in our society and expands the frontiers of political science and information technology. This has important applicability for increasing national security and prevention and attribution of terrorism. These findings help us to meet the goal of improved information utilization that also can be applied and extended to other important areas. Through international collaborators we will be able to obtain a more robust handle on matters of context, culture, multiple interpretations, multilingualism, imperatives of localization, etc. This also will lead to more effective use of information in society.

3. Knowledge acquisition and interpretation. Two of the fundamental goals of this effort are (1) the acquisition of information context knowledge (both for sources and users) and (2) the ability to use our SHIP's reasoning ability about this knowledge to correctly and effectively organize and interpret the information. A third goal will be shaped as a result of work on the fundamental ones, namely, articulating and formalizing logics required for reasoning about emergent knowledge acquisition and interpretation needs in the evolution of a 'context' (i.e., situation, conflict, etc.) over time.

The technical infrastructure and intellectual advances developed by the new Laboratory for Information Globalization and Harmonization Technologies (LIGHT) will be shared to encourage collaboration with the broader community**.** The materials will be made publicly available on the Internet including: literature reviews, survey results, theoretical models, reports, the System for Harmonized Information Processing technology, and other analyses conducted during the life cycle of the project, and an evaluative discussion forum. We expect this effort will generate important impacts for the research and practitioner communities, as well as society, in general.

CITED REFERENCES

[AKS96] Arens, Y., Knoblock, C. and Shen, W.M. (1996). "Query Reformulation for Dynamic Information Integration." *Journal of Intelligent Information Systems*, 6: 99-130.

[Alk96] Alker, H. R. (1996). Rediscoveries and Reformulations: Humanistic Methodologies for International Studies. Cambridge, UK, Cambridge University Press.

[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.

[BM97] Bloomfield, L.P., Moulton, A. (1997) Managing International Conflict: From Theory to Policy. New York: St. Martin's Press. (supplemented by CASCON web site http://mit.edu/cascon).

[CFM*01] Chen, X., Funk, J., Madnick, S., Wang, R. (2001) "Corporate Household Data: Research Directions", *Proceedings of the Americans Conference on Information Systems*, AMCIS'01, Boston, August.

[Chen99] Chen, P (1999). "ER Model, XML and the Web". *18th International Conference on Conceptual Modeling*.

[CZ98] Cherniak, M. and Zdonik, S (1998). "Inferring Function Semantics to Optimize Queries". *VLDB*, 239-250.

[Cho00] Choucri, N. (2000). "CyberPolitics in International Relations." International Political Science Review. **21**(3): 243-264.

[Cho01] Choucri, N. (2001) "Knowledge Networking for Global Sustainability: New Modes of Cyberpartnering", in Richard, D.J., Allenby, B.R. and Compton, D.W. Information Systems and the Environment. Washingtopn: National Academy Press, 195-210

[Cho99] Choucri, N. (1999). "*nnovations in Uses of CYberspace*," in Becker, E. and Jahn, T. eds. Sustainability and the Social Sciences. New York: Zed Books-St. Martin's Press, 274-283.

[ChoN93] Choucri, N., North, R.C. (1993). "Growth, Development, and Environmental Sustainablity: Profiles and Paradox." Global Accord: Environmental Challenges and International Responses. Cambridge, MA, MIT Press. 67-131.

[ChoNY92] Choucri, N., North, R.C. and Yamakage, S. (1992). The Challenge of Japan Before World War II and After. London: Routledge

[Chu00] Chu, W. (2000): "Introduction: Conceptual Models for Intelligent Information Systems". *Applied Intelligence 13*, (2).

[CR96] Cioffi-Revilla, C. (1996). Origins and Evolution of War and Politics. International Studies Quarterly. **40:** 1-22.

[DFen01] Ding, Y., Fensel, D. (2001) "Ontology Library Systems: The Key to Successful Ontology Reuse", International Semantic Web Working Symposium, Stanford University, CA.

[DSW*99] Duineveld, A.J., Stoter, R., Weiden, M.R., Kenepa, B., Bejamins, V.R. (1999) "WonderTools? A Comparative Study of Ontological Engineering Tools", Proceedings of the 12th International Workshop on Knowledge Axuisition, Modeling, and Management (KAW'99), Banff, Canada.

[DT95] Dobbie, G. and Topor, R. (1995). "On the declarative and procedural semantics of deductive object-oriented systems," *Journal of Intelligent Information Systems*, 4: 193-219.

[Fensel01] Fensel D. (2001) Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, Springer.

[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

[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

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

[Gams68] Gamson, W. A. (1968). Power and discontent. Homewood, Illinois, The Dorsey Press.

[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.

[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.

[GS97] Geller, D.S., Singer, J.D. (1997). Nations at War: A Scientific Study of International Conflict. Cambridge, England, New York, NY, Cambridge University Press.

[HelfH00] Helfin, J. and Hendler J. (2000) "Dynamic Ontologies on the Web", Proceedings of the 11th National Conference on Artificial Intelligence (AAAI-2000), Menlo Park, CA: 443-449.

[HZ76] Hoole, F.W., Zinnes, D. A. eds. (1976). Quantitative International Politics: an Appraisal. New York, NY, Praeger.

[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.

[KKK99] Katzenstein, P., Keohane R., and Krasner, S. eds. (1999). Exploration and Contestation in the Study of World Politics. Cambridge, Massachusetts, The MIT Press.

[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.

[Las58] Lasswell, H.D. (1958). Politics: Who Gets What, When and How. New York, NY, McGraw-Hill.

[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.

[Levy89] Levy, J. S. (1989). The Causes of War: A Review of Theories and Evidence. Behavior, Society, and Nuclear War. Tetlock, Jervis, Stern, and Tilly, eds. New York, Oxford University Press.

[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.

[Mad01] Madnick, S. (2001) "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, September: 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.

[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.

[McN00] McNeill, J.R. (2000). Something New Under the Sun. New York, NY, W.W. Norton & Company.

[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.

[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.

[NRC02] National Research Council (2002), "Making the Nation Safer: The Role of Science and Technology in Countering Terrorism." Washington, D.C., National Academies Press. [http://www.nap.edu/html/stct/index.html]

[Nuna01] Nunamaker, J. F. (2001): "Collaboration Systems and Technology Track – Introduction". *Hawaii International Conference on System Sciences*.

[RAC01] Richard, D.J., Allenby, B.R., and Compton, W.D., eds. (2001). *Information Systems and the Environment*. National Academy of Engineering. Washington, D.C., National Academy Press.

[Rich60] Richardson, L.F. (1960). Statistics of Deadly Quarrels. Pittsburgh, PA, Boxwood Press.

[Rose90] Rosenau, J.N. (1990). Turbulence in World Politics: A Theory of Change and Continuity. Princeton, NJ. Princeton University Press.

[Russ72] Russett, B. M. ed. (1972). Peace, War, and Numbers. Beverly Hills, Sage Publications.

[Sind01] Sindjoun, L. (2001). "Transformation of International Relations – Between Change and Continuity." International Political Science Review **22**(3): 219-228

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

[Sin02] Singer, J.D. (2002) "Accounting for Interstate War: Progress and Cumulation", in Brecher, B. and Harvey, F.P. eds. Millennial Reflections on International Studies. Michigan: The University of Michigan Press, 598-615.

[SP99] Schweller, R. and Pollins B. (1999). "Linking the Levels: The Long Wave and Shifts in U.S. Foreign Policy, 1790-1993." American Journal of Political Science **43**(2): 431-464.

[SW92] Sheng, O. R. L., Wei (1992), C-P, "Object-Oriented Modeling and Design of Coupled Knowledge-base/Database Systems". *International Conference on Data Engineering*: 98-105.

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

[TBD87] Templeton, M., Brill, D., Dao, S.K., Lund, E., Ward, P., Chen, A. L.P., and MacGregor, R. (1987). "Mermaid --- a front end to distributed heterogeneous databases," *Proc of the IEEE*, **75**(5): 695-708.

[TM98] Tu, S.Y., Madnick S. (1998) "Incorporating Generalized Quantifiers into Description Logic for Representing Data Source Contents", *Data Mining and Reverse Engineering: Searching for Semantics,* Chapman & Hall.

[WAHW03] Wang, R., Allen, T., Harris, W., and Madnick, S., "An Information Product Approach for Total Information Awareness" (with R. Wang, T. Allen, W. Harris), Proceedings of the 2003 IEEE Aerospace Conference, Big Sky, Montana, March 8-15, 2003

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

[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.

[Wri65] Wright, Q. (1965). A Study of War. Chicago, IL, University of Chicago Press.

[WS99] Walter, B. and Snyder, J (1999). Civil Wars, Insecurity, and Intervention. New York, Columbia University Press

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