MIT Sloan School of Management

MIT Sloan Working Paper 4488-04
CISL Working Paper No. 2004-05
February 2004

Improving National and Homeland Security through a proposed Laboratory for Information Globalization and Harmonization Technologies (LIGHT)

Nazli Choucri, Stuart Madnick, Michael Siegel, Richard Wang

Improving National and Homeland Security through a proposed Laboratory for Information Globalization and Harmonization Technologies (LIGHT)

February 2004

Working Paper CISL# 2004-05

Nazli Choucri {nchoucri@mit.edu}
Stuart Madnick {smadnick@mit.edu}
Michael Siegel {msiegel@mit.edu}
Richard Wang {rwang@mit.edu}

Composite Information Systems Laboratory (CISL)
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, Massachusetts 02142

Improving National and Homeland Security through a proposed Laboratory for Information Globalization and Harmonization Technologies (LIGHT)

Abstract

A recent National Research Council study found that: "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*" [NRC02, p.304, emphasis added] That sentence succinctly describes the objectives of this project. Improved access and use of information are essential to better identify and anticipate threats, protect against and respond to threats, and enhance national and homeland security (**NHS**), as well as other national priority areas, such as Economic Prosperity and a Vibrant Civil Society (**ECS**) and Advances in Science and Engineering (**ASE**). This project focuses on the creation and contributions of a **Laboratory for Information Globalization and Harmonization Technologies** (LIGHT) with two interrelated goals:

(1) Theory and Technologies: To research, design, develop, test, and implement theory and technologies for improving the reliability, quality, and responsiveness of automated mechanisms for reasoning and resolving semantic differences that hinder the rapid and effective integration (**int**) of systems and data (**dmc**) across multiple autonomous sources, and the use of that information by public and private agencies involved in national and homeland security and the other national priority areas involving complex and interdependent social systems (**soc**). This work builds on our research on the **COntext INterchange** (COIN) project, which focused on the integration of diverse distributed heterogeneous information sources using ontologies, databases, context mediation algorithms, and wrapper technologies to overcome information representational conflicts. The COIN approach makes it substantially easier and more transparent for individual receivers (e.g., applications, users) to access and exploit distributed sources. Receivers specify their desired context to reduce ambiguities in the interpretation of information coming from heterogeneous sources. This approach significantly reduces the overhead involved in the integration of multiple sources, improves data quality, increases the speed of integration, and simplifies maintenance in an environment of changing source and receiver context – which will lead to an effective and novel distributed information grid infrastructure. This research also builds on our **Global System for Sustainable Development** (GSSD), an Internet platform for information generation, provision, and integration of multiple domains, regions, languages, and epistemologies relevant to international relations and national security.

(2) National Priority Studies: To experiment with and test the developed theory and technologies on practical problems of data integration in national priority areas. Particular focus will be on national and homeland security, including data sources about conflict and war, modes of instability and threat, international and regional demographic, economic, and military statistics, money flows, and contextualizing terrorism defense and response.

Although LIGHT will leverage the results of our successful prior research projects, this will be the first research effort to simultaneously and effectively address ontological and temporal information conflicts as well as dramatically enhance information quality. Addressing problems of national priorities in such rapidly changing complex environments requires extraction of observations from disparate sources, using different interpretations, at different points in times, for different purposes, with different biases, and for a wide range of different uses and users. This research will focus on integrating information both over individual domains and across multiple domains. Another innovation is the concept and implementation of Collaborative Domain Spaces (CDS), within which applications in a common domain can share, analyze, modify, and develop information. Applications also can span multiple domains via Linked CDSs. The PIs have considerable experience with these research areas and the organization and management of such large scale international and diverse research projects.

The PIs come from three different Schools at MIT: Management, Engineering, and Humanities, Arts & Social Sciences. The faculty and graduate students come from about a dozen nationalities and diverse ethnic, racial, and religious backgrounds. The currently identified external collaborators come from over 20 different organizations and many different countries, industrial as well as developing. Specific efforts are proposed to engage even more women, underrepresented minorities, and persons with disabilities.

The anticipated results apply to any complex domain that relies on heterogeneous distributed data to address and resolve compelling problems. This initiative is supported by international collaborators from (a) scientific and research institutions, (b) business and industry, and (c) national and international agencies. Research products include: a System for Harmonized Information Processing (SHIP), a software platform, and diverse applications in research and education which are anticipated to significantly impact the way complex organizations, and society in general, understand and manage critical challenges in NHS, ECS, and ASE.

Section 1. Project Overview and Significance

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. *This proposal seeks to identify new ways of addressing these challenges by significantly improving access to diverse, distributed, and disconnected sources of information.* Although this effort will focus on the realm of National and Homeland Security (NHS), the results have relevancy to economic prosperity and a vibrant civil society (ECS), as well as to the advancement of most scientific and engineering (ASE) endeavors that have such information needs.

1.2 Relevance to National Priority Areas

1.2.1 National and Homeland Security (NHS)

This project will focus on information needs in the realm of national and homeland security, involving emergent risks, threats of varying intensity, and uncertainties of potentially global scale and scope. Specifically, we propose 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, involve routinized searches, but 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.

Due to space limitations, this proposal document will focus primarily on the NHS national priority. There are similar and/or analogous needs and opportunities in the other national priority areas.

1.2.2 Economic Prosperity and Vibrant Civil Society (ECS)

The need for intelligent harmonization of heterogeneous information is important to all information-intensive endeavors – which encompasses many aspects of our economy and society, including business, government, research, and education. The fundamental technology research proposed has broad relevancy for all complex inter-organizational applications, such as Manufacturing (e.g., Integrated Supply Chain Management), Transportation/Logistics (e.g., In-Transit Visibility), Government (e.g., Electronic Voting), Military (e.g., Total Asset Visibility), and Financial Services (e.g., Global Risk Management). Our LIGHT team is involved in research in all of these areas. People from different organizations and different parts of our societies have different perspectives (i.e. "contexts"). Rather than requiring them all to change to some imposed "standard", it is much more viable to have the information systems able to adapt to the people's needs (i.e., "context mediate"). Laws or policies that may unnecessarily limit or impair the effective use and re-use of information will also be examined.

1.2.3 Advances in Science and Engineering (ASE)

Similarly, the advancement of science and engineering involves the accumulation and use of information and knowledge, often gathered by multiple organizations, in different formats, and for differing purposes. We are working with colleagues at MIT and other institutions in several areas, such as biology, healthcare, engineering product design, and manufacturing, to draw on their experience with these types of barriers.

The field of biology, for example, has become increasingly information-intensive. Information generated in life sciences research is so large that no single person or group owns or controls all the needed data sources. A pharmaceutical company, for example, combines information from 40 sources on average to conduct research in drug development. Although much of this information is publicly available, heterogeneity in data structure and semantics limits the ability of life science researchers to easily integrate and exploit research data. Biologists often think in terms of pathways, may it be sequence analysis, functional genomics, proteomics or literature search. Pathways, discovered by different groups do not have a uniform representation. Pathway integration will be critical to systemic understanding how the cell works and will significantly speed up advances in the field. LIGHT will enable semantic interoperability between life science information sources, which have diverse data representations and semantics. In contrast to more constrained approaches, LIGHT will simultaneously support multiple views. For example, rather than adopting a single gene centric view as the standard way of viewing data, the system will adjust data automatically if the researcher wants to view the data in terms of function, disease, phenotype, or organ. Similarly, data semantics will be adjusted automatically reflecting the assumptions of a particular researcher: be it a biologist, geneticist or a medical researcher.

1.3 Addressing Information Needs

1.3.1 Operational Example

For illustrative 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 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.

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 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, which will be explained in more detail in Section 2.)

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.

1.3.3 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:

<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 2).

<u>For time T1:</u>

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>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><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">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><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>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><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.

1.3.4 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 this research effort is to create the Laboratory for Information Globalization and Harmonization Technologies (LIGHT) with capabilities that can 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 will be 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 propose to utilize the foundation of COIN to develop theories and methodologies for our proposed 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). Developing such a flexible, scalable software platform will require significant additional research in a number of key research areas as described in Section 2.4.

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 testing ground for the technologies we propose to develop, including automated methods for information aggregation from various sources, context mediation capabilities, customized information retrieval capabilities, and ontology representations.

1.5. Research Team

Due to the multi-disciplinary nature of LIGHT, we have composed a research team that is uniquely qualified to conduct this work. The PIs of this project come from MIT's School of Humanities, Arts, and Social Sciences (Choucri), School of Engineering (Madnick and Wang), and School of Management (Siegel and Madnick), and the students who will contribute significantly to the research come from all these diverse Schools. Furthermore, the PIs have extensive research experience in critical areas characterized by rapid change, system instabilities, and demands for rapid response to information need. These are all necessary to accomplish the goals of this project.

1.6. Proposal Organization

The remainder of this proposal elaborates on the intended research tasks. Section 2 describes research needs

in IT theory and technologies. How these capabilities can address the national priorities is discussed in Section 3. Section 4 provides a brief description of the new laboratory, its intellectual and research strategy, and how it will ensure coherence among the components of the project and also handle outreach and dissemination activities. Finally Section 5 presents the anticipated contributions of the project, with a focus on educational impacts.

Section 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. It is proposed that the application focus for this research effort be in the domains of the national priority areas with specific emphasis on national and homeland security, which by definition, takes into account internal as well as external dimensions of relations among actors in both the public and the private domains.

This research effort will:

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 in NHS: 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 focus of this part of the effort is **the acquisition, organization, and effective intelligent usage of distributed context knowledge to support information harmonization and collaborative domains**. {*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.*}

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).

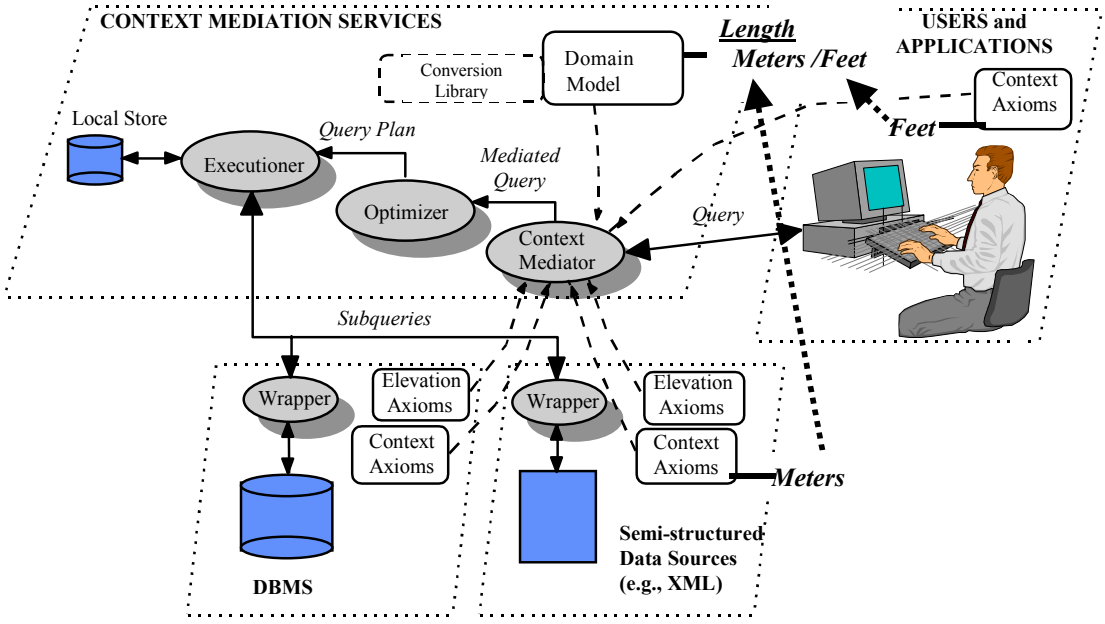

Figure 1. The Architecture of the Context Interchange System

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.

2.4. Research Tasks and Expected Contributions in Integrating Systems (int) and Data (dmc) Involving Complex and Interdependent Social Systems (soc)

Sometimes research is viewed as either "impossible" or "trivial." We believe that the thirteen research goals below ideally match the "high risk, high impact" goals of the NSF ITR. First, they build on our proven COIN and GSSD efforts and, in many cases, we have working papers describing approaches toward solutions (due to space limitations, it is difficult to present many details) – so we strongly believe that our goals are definitely "possible." On the other hand, each of these research goals separately is challenging and we also believe that no one has attempted to accomplish them all in unison, so it is definitely "not trivial." Even if we succeed in accomplishing only a subset of these goals, it would be a major contribution – but our goal is to accomplish and integrate them all.

1. Extended Domain of Knowledge – Equational Context. In addition to the representational context knowledge currently handled by the COIN framework, we need to perform research 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. We propose to extend the original COIN design 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 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. *{See [FMG02] for more details on proposed solution approach.}*

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 jurisidiction, and often changes in modes of information provision and formatting. This is a common problem facing a new government after a revolution. We propose to augment the COIN context knowledge representation to include a specification of the history of all contextual attributes in the ontology. 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]. *{See [ZMS04] for more details on proposed solution approach.}*

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 in many other 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 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 – corporate structure related and task related. Corporate 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 will be extended to comprehend the general semantics of the organization hierarchies that must be navigated. *{See [MWX03] for more details on proposed solution approach.}*

4. Linked Collaborative Domain Spaces. The existing COIN framework provides 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. The primary focus of this overall research effort 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]. *{See [Kal03] for more details on proposed solution approach.}*

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 intensional 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 will build upon the formal system underlying our current 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. By selecting applications, where fundamental shifts in relationships, systems, and pressures, we are opting for the 'tough test' where 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. 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 (ie

who is being counted). *{See [TM98] for more details on proposed solution approach.}*

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. Although there has been some basic research on modeling the semantics of data quality [WKM93], significant additional research must be done to advance and formalize these notions and then incorporate them into the SHIP system. *{See [Mad03] for more recent details on proposed solution approach.}*

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. *{See [Lee02] for more recent details on proposed solution approach.}*

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 will be 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 be handled by the sources individually (e.g. a join across sources). *{See [Tar02] for more details on proposed solution approach.}*

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. Our 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 domain experts be able to easily provide the needed domain and context knowledge. It is therefore essential to develop 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. Our 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. *{See [Lee03] for more details on proposed solution approach.}*

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 will contribute to the maturing of the Semantic Web and, at the same time, LIGHT will exploit relevant ontologies, standards and tools that emerge from the Semantic Web activities.

12. Operational System for Harmonized Information Processing. A critical goal of this project is to develop a fully operational System for Harmonized Information Processing (SHIP), a distributed information grid infrastructure, that will be used to support the types of challenges listed in Section 1, incorporating all the components identified above. It is essential that this system be developed with maximum flexibility and extensibility that will permit 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 is seriously constrained by the difficulties of reconciling known disparities and conflicts within and across sources. (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 started to apply 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. *{See [ZMS02] for more details on proposed solution approach.}*

Section 3. National Priority Area Research – Focus on National and Homeland Security (NHS)

National and homeland security (NHS) is our primary research priority area. In this section, we describe some of the most fundamental barriers to the reliable use of information systems in this area. They are also directly relevant to EVS and ASE. *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 [http://www.pcr.uu.se/ research/UCDP/ conflict_dataset_catalog/data_list.htm]. 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 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] which identifies new directions of research, research priorities, and critical next steps. {*For social science logic application see* [Cho99]}.

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 Research 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, the 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 proposed here will draw on the DDI results and enable information extraction and fusion in a collaborative environment hitherto unreachable {for details see, http://www.icpsr.umich.edu/DDI/index.html.}

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 technology developed in this project, 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 we seek to address with 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

The proposed research design is based on the structural differentiation among *contextual* conditions, and 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 test applications in the research design (see below) we will focus on (i) properties of the *context*-situation; (ii) properties of the *data features*, (iii) properties of the *data collection agencies*.

3.4.2 Test-Applications

The research tasks identified in Section 2 above are framed below in terms of the 'tough case', 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. The research design includes three sets of test 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 Spacial 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 propose to 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 spacial 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 research design, we refer below to test-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 propose to use the ontology for the 'conflict and war' domain developed for GSSD. as our research 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. {*See* "Using GSSD- GSSD Knowledge Strategy" at *http://gssd.mit.edu/GSSD/gssden.nsf }*

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 research platform currently 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 is chosen as a research 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 and Expected Contributions

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

3.5.1 Undertake a 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 anticipated contributions 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.

3.5.2 Conduct an extensive 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 deliverables 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.

Section 4. Laboratory for Information Globalization and Harmonization Technologies

The **Laboratory for Information Globalization and Harmonization Technologies (LIGHT** will be 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 to the research activities, the lab will host the technical infrastructure of the project, in particular our System for Harmonized Information Processing (SHIP), and the publication and dissemination of research tools and findings, and will serve as a focal point for new educational and research initiatives, both here and overseas.

In practice, the research activities in this multidisciplinary Laboratory will bring 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, as well as key research centers relevant to this work, notably the Center for eBusiness (CeB), Center for Technology, Innovation, and Policy Development (CTIPD), the Center for International Studies (CIS), and the Laboratory for Energy and Environment (LFEE).

More specifically, the proposed Laboratory will be the central entity for producing products in four areas: (1) Software Platforms, (2) Knowledge Repositories, (3) Application Demonstrations, and (4) Education and Research. The software platforms will include but not be 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 will include the NHS domain specific knowledge represented in ontologies, context and conversion libraries, source schemas and capabilities. The Application Demonstrations will be developed at MIT, with the participation of the Project collaborators, nationally and internationally. Significant efforts will be placed on technology transfer and open source Web presence.

In Education and Research, the Laboratory will have three sets of outreach activities to the scholarly and the policy communities: (a) an ongoing Workshop on Innovations in Harmonizing Information, designed largely for experimental work across disciplines and domains, (b) a periodic Symposium on Advances in Information Technology in National Priority Areas, targeted as an interface to the national and international policy-making communities, and (c) a web site that will include access to our SHIP, host the Studies, house ongoing research activities, and useful links that are relevant to our research, as well as electronic discussion forums. The Laboratory will also issue its own working papers and, as appropriate, organize its Book series, potentially with the MIT Press, and it will coordinate the Project's educational activities, research materials, and outreach initiatives.

We have designed an initial plan for engaging women, underrepresented minorities, and persons with disabilities in this effort, in particular, we support travel and registration expenses to LIGHT workshops and conferences and, as part of MIT's Affirmative Action policy and institutional support, will attempt to recruit for the post-doc position.

Section 5. Anticipated Contributions and Broader Impacts

This project will lead to major advances in information technology and applicable to the national priority areas. The outcomes of this innovative project will address many of the challenges facing our nation:

1. Theory and Technology. This project will produce a robust platform, the LIGHT System for Harmonization of Information Processing (SHIP), 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. Reliability of systems will be significantly improved by dynamically incorporating semantically equivalent sources into the interconnected system. The general-purpose platform will allow new applications to be built quickly to facilitate information sharing among diverse groups of people, devices, and software systems. Since the platform will facilitate 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. Address National Priorities. This project will significantly augment the effective use of information in our society and expand the frontiers of political science and information technology. This has important applicability for increasing national security and prevention and attribution of terrorism. These findings will help to meet the goal of improved information utilization that also can be applied and extended to other important areas, such as economic effective of our society and advances in science and engineering. 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 contribution also will lead to more effective use of information in society enabling more informed citizen participation.

3. Knowledge acquisition and interpretation. Two of the fundamental goals of this project are (1) the acquisition of information context knowledge (both for sources and users) and (2) the ability to use our proposed SHIP's reasoning ability about this knowledge to correctly and effectively organize and interpret the information.

4. Education. Our project will contribute to education in specific ways: it will help to transform the traditional IT educational setting by incorporating various disciplines into the development of new IT theories and tools. In addition, by facilitating the **integrated study** of complex issues, this research will help to develop and foster new multidisciplinary learning environments. Our project will also contribute to the education of new researchers, including post-doctoral associates, graduate students, and undergraduate students, who will take an active role in the research of this project. We anticipate that the impact to education will be profound and continuous as our international collaborators begin to adapt the project's curricula to their own contexts, educational programs, and institutional conditions. We propose to interface with the MIT OpenCourseWare administration to draw on the most recent educational technology outreach system.

In conclusion, the research team plans to utilize the technical infrastructure and intellectual advances developed by the new Laboratory for Information Globalization and Harmonization Technologies (LIGHT) to share its findings and encourage collaboration with the broader research community**.** The materials that will be publicly available on the Internet include: literature reviews, survey results, theoretical models, reports, the System for Harmonized Information Processing technology, other analyses conducted during the life cycle of the project, and an evaluative discussion forum. We expect the results will generate profound impacts for the research, education, and various practitioner communities, as well as society, in general.

COORDINATION PLAN AND INTERNATIONAL COLLABORATIONS

Recognizing that advances in information technology are essential for achieving the NSF-defined national priorities, we propose to integrate and manage all components of the proposed research under a newly created laboratory, named the **Laboratory for Information Globalization and Harmonization Technologies (LIGHT).** The lab will oversee and coordinate all research activities, host the technical infrastructure, coordinate outreach activities of the project, and disseminate the products of LIGHTS research (such as publications, platforms, tools, and educational materials) and host the proposed Symposia and Workshops.

The laboratory will be jointly run by the co-PIs (Choucri, Madnick, Siegel, Wang) who have effectively worked together (in groups of two or three) on other projects. One of the PIs (Siegel) will take the key role in the day-to-day management and coordination of the Laboratory. This management team is dedicated to providing results that will directly address the information technology problems and applications central to national priorities in IT.

A **steering committee** of approximately eight individuals will be formed from the national and international collaborators, drawing approximately one individual from each of the categories listed below. This steering committee will meet at least twice annually and provide both feedback and priorities to this research effort.

The proposed project consists of three components that will focus on different, but related, areas of interest: (1) identifying barriers to access of information for education, research, decision making, and performance in the national priority areas, (2) development of new information technologies to address these needs, where there are multiple actors and domains of salience, and rapidly changing conditions, and (3) advancing innovation in the use of the technologies to facilitate interdisciplinary research and contribute to new education materials, approaches, tools, and methods.

The NHS research component will be directed by one PI (Choucri) and will include the efforts of one full-time doctoral student and several research assistants. The IT development will be directed by one PI (Madnick), with specific technical areas assigned to the other co-PIs (Siegel and Wang), and will include the efforts of one full-time doctoral student and several research assistants. The education component of the project will be supported by all four PIs, and will include the efforts of all full-time doctoral students and graduate research assistants. All of the PIs have considerable prior experience with the organization and management of large scale, international, diverse and distributed research projects.

At the foundation of this proposal is a network-in-place of national and international collaboration. These include a wide range of collaborators, each with their own distinctive operational context and expected participation. The list below names some of the initial collaborators that have verbally committed to this effort (*letters of confirmation from thirteen of the collaborators, marked with *, have been included in the Supplemental Documents*). The Table highlights four types of contributions: (1) **reviewers** (who contribute valuable input on the research), (2) **data sources** (who provide data for application testing), (3) **users** (potential users of the technology who help with the problem definition and who provide challenging test cases), and (4) **active researchers** (who will directly participate in and contribute to our research). None of these collaborators will be receiving any of the NSF funds, but they will significantly leverage the funds that are provided.

Names and Institutions of Collaborators	Institution Type	Anticipated Roles	Benefits to the Research
C. von Furstenberg, **UNESCO** B. Plescovic, **World Bank**	International governmental organizations	**Data sources and users,** contribute to understanding changing policy contexts & impact on information needs.	Direct inputs on policy deliberations affecting context and framework for of international information systems.
J. Cares, **Alidade Consulting** * M. Laguerre, **U. Berkeley Institute of Global Studies** * P. Brecke, **Georgia Tech, Nunn School of International Affairs** B. Pollins, **Ohio State University** M. Feldman , **Stanford University** A. White and R. Massie, **Global Reporting Initiative**	Scientific research and policy institutions	**Reviewers, users, and active researchers,** who will also participate in workshops and help to develop new applications.	Provide comparative bases for assessing generalizability and collaborate on new applications.

Names and Institutions of Collaborators	Institution Type	Anticipated Roles	Benefits to the Research
* B. Allenby, **AT&T** * W. R. Baker,**Baker & McKenzie** * Dan Schutzer, **Citibank** U. Wennberg, **Global Responsibility, International** K. Cavanaugh, **IBM** * J. D. Funk, **S.C. Johnson Company** * L.G. Scheidt, **Sony International Advanced Technology Center**	Global firms – Information Technology, Legal Services, Financial Services, Consumer Products, and Electronics	**Reviewers and users,** contributing to improved applications, including relevance of changing contexts. Insights into integration issues in large multinational environments with heterogeneous global data sources.	Diversity of professional and domain expertise, covering variations in legal contexts, environmental research, and responses to the cultural diversification of the global workplace. These organizations are currently working with various of the co-PIs.
* B. Davidson, **Cedars Sinai Health System** * C. Marshall, **New York State Office for the Aging**	Non-profit org – health care and elderly	**Reviewers and users,** important applications and issues in complex governmental and non-profit environments with heterogeneous data sources.	Currently working with co-PI Wang on improving the use of information in their organizations, especially improving information quality.
G. Kochendoerfer-Lucius, **German Foundation for International Development** C. Brodhag, **Ecole des Mines a St. Etienne**, **France** S. Chengyoung, **Ministry of Science & Technology**, **China**	Governmental scientific agencies	**Data source and active researchers,** contributing to contextual evaluation, cross-cultural interpretation and meanings, local knowledge provision, and comparison across contexts.	Currently working with PI Choucri on global knowledge networking. Direct input into contextual biases, or errors in assignment of meaning to recorded observations.
* T. Mezher, **American University of Beirut, Lebanon** A.Koshla, **Development Alternatives India** M. Tolba, L. Hassenien, **ArabDev, Egypt**	Researchers from institutions in developing countries	**Data source and active researchers,** with a focus on the provision of local and national knowledge.	Currently collaborating with PI Choucri on global knowledge networking. Important to comparative and diverse contextual applications, validation of internationalization.
* A. Segev, **U. Berkeley Center for information Technology** * Nor Adnan Yahaya, **Malaysia University of Science and Technology** * Tan Kian Lee and Stephane Bresson, **National University of Singapore**	Research Universities	**Data sources, reviewers users, active researchers (IT)**, providing complementary labs for development of theory and software platform.	Working with Co-PIs Madnick and Siegel. Active database researchers having significant experience with web-based information integration.

Given the highly multi-disciplinary nature of this effort, the research will be supported by this outstanding and diverse research team of international collaborators, with multiple demographics, experiences, and qualifications. We strongly believe that this project will lead to important developments in domains of IT research and national priority areas. In particular, their intersection will have a significant impact on the way organizations (e.g., governments, companies, world bodies) understand, react to, and manage the significant global challenges (e.g., war, terrorism, environment) of the 21st century.

FACILITIES, EQUIPMENT, AND OTHER RESOURCES

COMPUTING EQUIPMENT AND DATA SOURCES

The **Laboratory for Information Globalization and Harmonization Technologies and Studies**, to be formed, will primarily use existing computing equipment from the Context Interchange Systems (COIN) laboratory (within the Information Technology group of MIT's Sloan School of Management) and the Global System for Sustainable Development (GSSD) project (within MIT's Center for International Studies and Political Science department.) Both facilities are located in the same building, and most on the same floor, so coordination will be easy.

Equipment currently available within the COIN lab includes two Sun Unix servers, two Windows NT servers, a Linux server, and 16 current generation Intel workstations running versions of Windows or Linux as appropriate for research needs. Available software includes Microsoft development, systems, and server platforms as well as open source resources for software development, knowledge management, and database management. The latest version of the COIN context mediation prototype, for knowledge representation and reasoning, was developed within this lab and this software infrastructure will constitute a starting point for the proposed effort.

In addition, we will draw on the two Pentium workstations and 3 Windows NT servers, and data sources of the GSSD. GSSD is the knowledge networking and management system for the Alliance for Global Sustainability (which includes MIT, University of Tokyo, Chalmers University-Sweden, and ETH - the Swiss Technical University System). GSSD mirror sites are maintained in France (École Nationale Supérieure des Mines de Saint Etienne), China (Ministry of Science and Technology) and Japan (University of Tokyo).

OTHER RESOURCES

As part of its dual and integrated focus on education and research, there are more than 3,000 ongoing projects on campus at MIT. These projects utilize shared centralized facilities, such as contemporary computational aids and library facilities, as well as specialized facilities of individual departments, research centers, and labs. Each project is affiliated with a nodal department, but can access resources in other parts of MIT. This project will draw particularly on MIT's extensive communications and network infrastructure.

The co-PIs are affiliated with various organizational units and research centers at MIT and will have access to their resources, especially the departments of Information Technologies, Political Science, Management Science, and the Technology, Management and Policy program, as well as key research centers, notably the Center for eBusiness (CeB), Center for Technology, Innovation, and Policy Development (CTIPD), the Center for International Studies (CIS), the Technology and Development Program (TDP), the Total Data Quality Management (TDQM) program, the Productivity from Information Technology (PROFIT) program, and the Laboratory for Energy and Environment (LFEE).

A primary mission of MIT is education and many MIT resources will be used to facilitate the development, testing, and deliver of new educational materials. In particular, we plan to work with MIT's OpenCourseWare initiative, which reflects MIT's institutional commitment to disseminate knowledge across the globe. One of the co-PIs (Madnick) serves on the OCW Advisory Board. We will also make use of other media development, presentation, and transmission facilities, such as MIT's new Learning International Network Consortium (LINC), which supports collaboration and cooperation across international borders through technology-enabled media for higher quality education to 'learners' worldwide.

In this research effort we plan to work with collaborators as reviewers, data sources (who provide data for application testing), users (potential users of the technology who help with the problem definition and who provide challenging test cases), and active researchers (See Management Plan for more details). As a result of the active participation of these collaborators (i.e., international and governmental organizations, scientific research and policy institutions, researchers from institutions in developed and developing countries, global commercial firms, non-profit organizations and universities) we expect to have access to and involve a number of resources from these organizations, including databases, applications, algorithm and theory development, software, and facilities for meetings and demonstrations.

BUDGET JUSTIFICATION

Key Personnel
A1. This project represents a major multidisciplinary effort with significant distinct but interrelated components: (1) theory and technology development, (2) applications and studies in the national priority areas, (3) knowledge collection, (4) educational material development, and (5) outreach for education and global impact. We expect the Principals to lead these efforts, to coordinate across the components, and to facilitate their success.

B1. A post-doctoral student will work alongside Dr. Siegel to facilitate coordination across these diverse efforts, between the schools at MIT, and with our national and international collaborators. We recognize the importance of ensuring timely activities and outputs, appropriate sequencing of tasks, and effective streamlining of interactions among all participants, as well as managing report preparations, working papers, and internal and external research communication. We feel that these efforts will require the times allocated by the PIs and the post-doc.

B2. We have allocated 1/10 of a financial analyst's time to assist in management of budget, internal MIT requirements, and financial reports required of a large research project .

B3. This project seeks to, and will, depend heavily on graduate and undergraduate students, as an important contribution to their education (in terms of basic research as well as the 'pre-testing' of educational materials we will prepare). For graduate students we will be using one doctoral and one masters student to focus on the development of the technology platform, SHIP. For the collection of NHS related data, the generation of new data, and integration of information on key data-generating institutions, we will be using an additional doctoral and an additional masters student. All students, including the undergraduates, will be assisting in the development of coursework, meetings, seminars and other outreach programs. In addition, for efforts such as conferences, courses, and larger meetings we can call upon our undergraduates and graduate students to assist. Such involvement has a multiplier benefit, namely that of providing these students with a closer working relationship with the project, tighter connections to their educational programs, and new experiences working with its collaborators -- while limiting our support staff requirements. Graduate students are noted in Section B of the budget. Tuition support for graduate students is detailed in section G6.

B4. We intend to involve undergraduate students as well, to assist the project in select tasks. More specifically, these students will be UROP (Undergraduate Research Opportunity Program) students. UROP is a unique program at MIT that has a long track record of providing undergraduates the ability to work with research projects while providing faculty with a low cost, low overhead, high motivation workforce. These students have proven to be particularly helpful in developing software, collecting data, and supporting faculty and advanced graduate students. We are requesting support for five UROP students for the duration of this project.

B5. This is a substantial project with global scope that will require careful support, in terms of clerical requirements as well as financial tracking and analysis. For this reason, we plan to fund 1/3 of the time of an administrative assistant. We believe this to be very conservative given our overall needs.

MIT budgeting guidelines:
(1) Non-faculty salaries are inflated at 3% per year on January 1 of each year
(2) Faculty salaries are inflated at 3% per year on July 1 of each year

C1. Fringe Benefits
Employee Benefits are calculated at 25% of base salary for all staff positions.

D1. Equipment
This category includes all equipment with costs over $3000. At these costs MIT does not charge overhead on such equipment. Also, it is possible to make upgrades to this equipment without occurring additional overhead. A majority of the machine costs are included in this category because of the favorable overhead situation. Complete systems including peripherals can be priced in a bundle to fit into this budget line. In the first year we plan to use a larger sum ($15,400) to set-up two servers and 3 complete desktops. The servers will be used to support a development platform and a release platform to collaborators and other organizations, nationally and internationally.

It will support the website for outreach and education programs and the center point for dissemination of publications. Access to all applications, domain knowledge and contributions by collaborators will be coordinated through this site. The three desktops will be used to support the post-doc and two of the graduate students. After the first year we will use portions of the smaller equipment budget ($5400 in year 2 and 3) to upgrade existing machines and one new machine per year to replace older machines that were in place prior to the project but that have become obsolete. In the last year we use a smaller budget ($3000) to upgrade existing machines. A smaller budget ($3200 first year and $1800 in later years) is used for smaller equipment (fax machines, printers) and networking equipment and one desktop per year for students or principals. Overhead is applied to such equipment purchases.

Travel

E1. Travel is central to our outreach activities. In all years we plan one trip for two principals for meetings with NSF and other government agency interested in the research and the results of the research. We target three domestic trips to meet with collaborators. These would be meetings where we can get feedback on our application development, access to new sources of information, additional technology and domain expertise, and transfer results. In addition we plan to submit and present publications at domestic and international conferences and have budgeted a total of three conference attendances to allocate over our faculty and students at a cost of $10,400 in year one, with similar expenses, adjusted for inflation, in subsequent years. Our travel is important for outreach. However to accommodate the number of faculty and students we plan to use MIT as the primary meeting place for collaboration. Meeting costs will be discussed in the section G6.

E2. We plan to conduct one international trip to meet with collaborators and one trip to attend an international conference in each year of the project.

F1. Participant Support

We intend to subsidize travel costs for women, underrepresented minorities, and persons with disabilities attendees of each symposium in order to broaden the diversity of the participants. In addition, we will offer the same subsidy for attendance at each of the annual workshops. The total subsidy requested is $5000.

G1. Materials and Supplies

These are based on a first year rate of $6100. These include all costs for postage, telecommunications, network communication charges, back-up charges, and office supplies. These will be used by the individual investigators and for the new laboratory.

G2. Publications/Documentation/Dissemination

We plan to run a number of meetings/workshops and develop courseware for MIT's Open Courseware Initiative. Publication and dissemination costs are escalated in Years 2, 3, and 4 (i.e., by $2000) to accommodate a significant level of publication and outreach. Courses will be developed based on this research for the Sloan School of Management, the Political Science Department and for use in MIT-Singapore and MIT-Malaysia Alliances.

G4. Computer Services

This includes a reasonable number for software purchases such as licenses for the Laboratory and individual licenses as needed.

G6. Other

This category includes tuition expense for our graduate students and expenses for meetings with our collaborators and workshop and conference.

Tuition: Tuition costs are similar in all four years ($65000 first year no overhead charged). This represents a 50% subsidy of tuition costs for the four graduate students working on this project. Tuition support is required for all RAs under MIT's guidelines.

Meetings: In all years we plan to conduct quarterly meetings of the collaborators, biannual steering committee meetings and annual workshops. These will be working meetings and launch of application to selected collaborators. In addition, we are planning a symposium in year two with collaborators and outside organizations to present results and gain further buy-in to the project from the broader community across research, business, and policy domains and at regional as well as global levels. We expect to capture both the multiple perspectives and the diversity of organizational and institutional impacts upon information generation, compilation, and dissemination worldwide. Finally in year four we will host a broad symposium and application rollout to the community and transfer technology to appropriate collaborators to support a global platform. Expenses for these meetings include costs for

space, audiovisual equipment, networking, refreshments, and appropriate office supplies. These meetings and workshops will be web cast to reach the largest possible audience and minimize travel expenses.

Summary of Planned Meetings

Collaborators' Meetings	Quarterly	Up to 12 participants
Steering Committee Meetings	Twice per year	Up to 10 participants
Workshops	Once per year	Up to 30 participants
Symposia	One in year 2 and one in year 4	Up to 100 participants

Leveraging of these budgeted resources

The requested budget resources will be significantly leveraged in several ways: (1) much of the critical initial basic research, especially for the COIN and GSSD platforms, have been previously funded from NSF, DARPA, and industry sources, (2) the LIGHT collaborators are committing significant internal resources that will greatly assist this effort as well to facilitate the dissemination of results and impact of this research, (3) relevant and unique resources at MIT, such as MIT's OpenCourseWare and LINC technologies and distribution resources, will be utilized; as well as (4) recent experience and ongoing participation in national and international policy-related exercises – such as the National Academy of Sciences (NAS) Committee initiative on Terrorism, DARPA-National Academy of Sciences Committee on *Understanding Terrorism in order to Deter Terrorism,* and the United Nations – Information and Communication Taskforce (UN-ICT) background work in preparations for World Summit on Information Technology (WSIS).

Note: An overhead rate of 60% is applied to all direct costs except major equipment and tuition.

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.

OTHER RELEVANT REFERENCES

Theory and Technology

[ACH93] Arens, Y., Chee, C., Hsu, C., and Knoblock, C. (1993). "Retrieving and integrating data from multiple information sources," *International Journal on Intelligent and Cooperative Information Systems*, http://www.isi.edu/sims/papers/93-sims-ijicis.ps.

[ASD91] Ahmed, R., Smedt, P.D., Du, W., Kent, W., Ketabchi, M.A., Litwin, W.A., Raffi, A., and Shan, M.C. (1991). "The Pegasus heterogeneous multidatabase system," *IEEE Computer*, 24(12): 19-27.

[BFG*97] Bressan, S., Fynn, K., Goh, C., Madnick, S., Pena, T. and Siegel, M. (1997) "Overview of a Prolog Implementation of the COntext INterchange Mediator", *Proceedings of the Fifth International Conference and Exhibition on the Practical Applications of Prolog*, London, England, April 1997.

[BFG96b] Bressan, S., Fynn, K., Goh, C.H., Madnick, S., Pena, T., and Siegel, M. (1996). "Overview of a prolog implementation of the COntext INterchange mediator," In *Practical Applications of Prolog 97*.

[BGF*97a] Bressan, S., Goh, C., Fynn, K., Jakobisiak, M., Hussein, K., Kon, H., Lee, T., Madnick, S., Pena, T., Qu, J., Shum, A. and Siegel, M. (1997) "The Context Mediator Prototype" *Proceedings of the ACM Conference on the Management of Data (SIGMOD '97*), February.

[BGT*97] Bressan, S., Goh, C.H. , Lee, T., Madnick, S., Siegel, S. (1997) "A Procedure for Mediation of Queries to Sources in Disparate Context", *Proceedings of the International Logic Programming Symposium,* October.

[CHS91a] Collett, C., Huhns, M., and Shen, W. (1991). "Resource Integration Using an Existing Large Knowledge Base," *MCC Technical Report Number ACT-OODS-127-91*.

[CHS91b] Collet, C., Huhns, M.N., and Shen, W.M. (1991). "Resource integration using a large knowledge base in Carnot," *IEEE Computer*, 24(12): 55-63.

[Date87] Date, C. (1987). *An Introduction to Database Systems, volume 1, fourth edition*. Addison-Wesley Publishing, Menlo Park, CA.

[DG96] Duschka, O. and Genesereth, M. (1996). "Query Planning in Infomaster," http://logic.stanford.edu/people/duschka/papers/Infomaster.ps.

[DGH*96] Darawala, A., Goh, C., Hofmeister, S., Madnick, S. and Siegel, M. (1996) "Context Interchange Network Prototype" in *Database Applications Semantics*, Chapman & Hall (London, UK), October.

[DGH95] Daruwala, A., Goh, C.H., Hofmeister, S., Hussein, K., Madnick, S., and Siegel, M. (1995). "The context interchange network prototype," In *Proc of the IFIP WG2.6 Sixth Working Conference on Database Semantics (DS-6)*, Atlanta, GA.

[FDF95] Faquhar, A., Dappert, A., Fikes, R., and Pratt, W. (1995). "Integrating information sources using context logic," In *AAAI-95 Spring Symposium on Information Gathering from Distributed Heterogeneous Environments*.

[FLM*01a] Fan, W., Lu, H., Madnick, S., Cheung, D. (2001) "Discovering and Reconciling Value Conflicts for Numerical Data Integration," *Information Systems Journal*, **26**(8), November: 635-656.

[FLM*01b] Fan, W., Lu, H., Madnick, S., Cheung, D.W. (2001) "DIRECT: A System for Mining Data Value Conversion Rules from Disparate Sources, *Decision Support Systems* (in press).

[GGKS95] Geddis, D., Genesereth, M., Keller, A., and Singh, N. (1995). "Infomaster: A Virtual Information System," http://logic.stanford.edu/papers/iiaw95_infomaster.ps.

[GMP95] Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., Vassalos, V., and Widom, J. (1995). "The TSIMMIS Approach to Mediation: Data Models and Languages," *Next Generation Information Technologies and Systems*, Naharia, Israel, November, Anonymous ftp to db.stanford.edu/pub/papers/mis-models-languages.ps.

[Guha91] Guha, R.V. (1991). "Contexts: a formalization and some applications," *Technical Report STAN-CS-91-1399-Thesis*, Department of Computer Science, Stanford University.

[Jako96] Jakobisiak, M. (1996). "Programming the web -- design and implementation of a multidatabase browser," *Technical Report CISL WP #96-04*, Sloan School of Management, Massachusetts Institute of Technology.

[JS95] Jonker, W. and Schuetz, H. (1995). "The ECRC multidatabase system," In *Proc ACM SIGMOD*, page 490.

[KS91] Kim, W. and Seo, J. (1991). "Classifying Schematic and Data Heterogeneity in Multidatabase Systems," *IEEE COMPUTER*, December: 12-18.

[KuL88] Kuhn, E. and Ludwig, T. (1988). "VIP-MDBMS: A logic multidatabase system," In *Proc Int'l Symp.*

on Databases in Parallel and Distributed Systems.

[LaR82] Landers, T. and Rosenberg, R. (1982). "An overview of Multibase," In *Proc 2nd International Symposium for Distributed Databases*: 153-183.

[LeG89] Lenat, D.B. and Guha, R. (1989). *Building large knowledge-based systems: representation and inference in the Cyc project*. Addison-Wesley Publishing Co., Inc.

[LFG*97] Lu, H., Fan, W., Goh, C.H., Madnick, S., Cheung, D.W. (1997) "Discovering and Reconciling Semantic Conflicts: A Data Mining Perspective", *Proceedings of the IFIP 2.6 Conference on Database Semantics*, October.

[LFG*98] Lu, H., Fan, W., Goh, C.H., Madnick, S. and Cheung, D.W. (1998) "Discovering and Reconciling Semantic Conflicts: A Data Mining Perspective", *Data Mining and Reverse Engineering: Searching for Semantics,* Chapman & Hall.

[Lit92] Litwin, W. (1992). "O*SQL: A language for object oriented multidatabase interoperability," In Hsiao, D.K., Neuhold, E.J., and Sacks-Davis, R., editors, *Proc of the IFIP WG2.6 Database Semantics Conference on Interoperable Database Systems (DS-5),* Lorne, Victoria, Australis. North-Holland: 119-138.

[LitA87] Litwin, W. and Abdellatif, A. (1987). "An overview of the multi-database manipulation language MDSL," *Proc of the IEEE*, **75**(5): 621--632.

[LM96] Lee, J. and Madnick, S. (1996) "Resolving Context Heterogeneity: Applying Insights from Semantics and Ontology", *Proceedings of the Workshop on Information Technology & Systems*, December.

[LMS96a] Lee, J., Madnick, S. and Siegel, M. (1996) "The Context Interchange Approach to Large Scale Data Integration: Benefits and Challenges", *Proceedings of the International Conference on Data Engineering*, August.

[Mac88] MacGregor, R. (1988). "A Deductive Pattern Matcher," In Proceedings of AAAI-88, The National Conference on Artificial Intelligence. St. Paul, MN, August.

[Mad00] Madnick, S. (2000) "Chapter 6: The MIT Context Interchange Project" in *Data Quality*, Ed: R.Y. Wang, M. Ziad, and Y.W. Lee, Kluwer Academic Publishers: 79-92.

[Mad97a] Madnick, S. (1997) "Database in the Internet Age", *Database Programming & Design*, **10**(1), January: 28-33.

[Madn96] Madnick, S.E. (1996). "Are we moving toward an information superhighway or a Tower of Babel? The challenge of large-scale semantic heterogeneity," In *Proc of the IEEE International Conference on Data Engineering*, pages 2-8. Also reprinted in 21st (a Web 'Zine at http://www.vxm.com/), Vol 1, No. 6, April 1996, pp. Speed 1-10.

[MMS01a] Moulton, A., Madnick, S., and Siegel, M. (2001) "Cross-Organizational Data Quality and Semantic Integrity: Learning and Reasoning about Data Semantics with Context Interchange Mediation", *Proceedings of the Americans Conference on Information Systems* (AMCIS, Boston), August.

[MMS01b] Moulton, A., Madnick, S., Siegel, M. (2001) "Knowledge Representation Architecture for Context Interchange Mediation: Fixed Income Securities Investment Examples", *Proceedings of the First International Workshop on Electronic Business Hubs:* XML, Metadata, Ontologies, and Business Knowledge on the Web (WEBH; Munich, Germany), published by IEEE, September: 50-54.

[MMS98] Moulton, A., Madnick, S. and Siegel, M. (1998) "Context Interchange on Wall Street", *Proceedings of the International Conference on Cooperative Information Systems* (CoopIS'98), September: 271-279.

[MWF*01] Madnick, S., Wang, R., Dravis, F. and Chen, X. (2001) "Improving the Quality of Corporate Household Data: Current Practices and Research Directions", *Proceedings of the Sixth International Conference on Information Quality* (IQ2001, Cambridge), November: 92-104.

[PFPS92] Patil, R.S., Fikes, R.E., Patel-Schneider, P.F., Mckay, D., Finin, T., Gruber, T., and Neches, R. (1992). "The DARPA Knowledge Sharing Effort: progress report," In *Principles of Knowledge Representation and Reasoning: Proc of the Third International Conference*, Cambridge, MA.

[PGM95] Papakonstantinou, Y. and Garcia-Molina, H. (1995). "Object Fusion in Mediator Systems (Extended Version)," Anonymous ftp to db.stanford.edu/pub/papakonstantinou/1995/fusion-extended.ps.

[PGMU96] Papakonstantinou, Y., Garcia-Molina, H., and Ullman J. (1996). "MedMaker: A Mediation System Based on Declarative Specifications," *International Conference on Data Engineering*, New Orleans, LA, February p. 132-41. Anonymous ftp to db.stanford.edu/pub/papers/medmaker. ps.

[PGMW95] Papakonstantinou, Y., Garcia-Molina, H., and Widom, J. (1995). "Object exchange across heterogeneous information sources," In *Proc IEEE International Conference on Data Engineering*.

[QRS95] Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., and Widom, J. (1995). "Querying semistructured heterogeneous information," In *Proc International Conference on Deductive and Object-Oriented*

Databases.

[Shum96] Shum, A. (1996). *Open Database Connectivity of the Context Interchange System*, Master's thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.

[SL90] Sheth, A.P. and Larson, J.A. (1990). "Federated database systems for managing distributed, heterogeneous, and autonomous databases," *ACM Computing Surveys*, **22**(3): 183--236.

[SM91b] Siegel, M. and Madnick, S. (1991) "A metadata approach to solving semantic conflicts," In *Proc of the 17th International Conference on Very Large Data Bases*: 133-145.

[SSR94] Sciore, E., Siegel, M., and Rosenthal, A. (1994). "Using semantic values to facilitate interoperability among heterogeneous information systems," *ACM Transactions on Database Systems*, **19**(2): 254-290.

[SYE90] Scheuermann, P., Yu, C., Elmagarmid, A., Garcia-Molina, H., Manola, F., McLeod, D., Rosenthal, A., and Templeton, M. (1990). "Report on the workshop on heterogeneous database systems," *ACM SIGMOD RECORD*, 19(4): 23--31. Held at Northwestern University, Evanston, Illinois, Dec 11-13, 1989. Sponsored by NSF.

[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

[TM97] Tu, S.Y., Madnick S. (1997) "Incorporating Generalized Quantifiers into Description Logic for Representing Data Source Contents", *Proceedings of the IFIP 2.6 Conference on Database Semantics*, October.

[TRV95] Tomasic, A., Raschid, L., and Valduriez, P. (1995). "Scaling heterogeneous databases and the design of DISCO," In *Proc of the 16th International Conference on Distributed Computing Systems*, Hong Kong.

[ZMS01] H. Zhu, S. Madnick, and M. Siegel, "Information Aggregation - a Value-Added E-Service", Proceedings of the 5th International Conference on Technology, Policy, and Innovation – Theme: "Critical Infrastructures", Delft, The Netherlands, June 26-29, 2001.

National and Homeland Security (NHS)

[AK97] Arbetman, M., and Kugler, J. eds. (1997). Political Capacity and Economic Behavior. Boulder, CO, Westview Press.

[AM99] Amadae, S. M., De Mesquita, B. B. (1999). The Rochester School: The Origins of Positive Political Theory. Annual Review of Political Science. Nelson W. Polsby, ed. Palo Alto, CA, Annual Reviews. **2:** 269-295.

[Bate01] Bates, R. H. (2001). Prosperity and Violence - The Political Economy of Development. New York, London, W. W. Norton and Company.

[BHP92] Bright, M., Hurson, A., and Pakzad, S. (1992). "A taxonomy and current issues in multidatabase systems," *IEEE Computer*, 25(3):50--60.

[BJ99] Becker, E., and Jahn, T. eds. (1999). Sustainability and the Social Sciences. London, Zed Books, Ltd.

[Brech96] Brecher, M. (1996). "Introduction: Crisis, Conflict, War - State of the Discipline." International Political Science Review **17**(2): 127-139.

[Cho81] Choucri, N. (1981). International Energy Futures - Petroleum Prices, Power and Payments. Cambridge MA, The MIT Press.

[Cho92] Choucri, N. (1992). "Environment and Conflict: New Principles for Environmental Conduct." Disarmamemnt **XV**(1): 67-78.

[Cho93a] Choucri, N. (1993). "Political Economy of the Global Environment." International Political Science Review **14** (1): 103-116.

[Cho93b] Choucri, N. (1993). Environmentalism. Oxford Companion to Politics of the World. e. J. Krieger. New York, Oxford University Press.

[Cho93c] Choucri, N., ed. (1993). Chapter 1: Introduction: Theoretical, Empirical, and Policy Perspectives. Global Accord: Environmental Challenges and International Responses. Cambridge, Mass., MIT Press.

[Cho95] Choucri, N. (1995). "Globalization of Eco-Efficiency: GSSD on the WWW." UNEP Industry and Environment(October-December 1995): 45-49.

[Cho98] Choucri, N., McHugh and Millman (1998). Innovations in Cyberpartnering for Sustainability. Care Innovation '98: Proceedings, Second International Symposium, Brokerage Event and Environmental Exhibition, Vienna, Austria.

[Cho99a] Choucri, N. (1999). Innovations in Use of Cyberspace. Sustainability and the Social Sciences. Becker and Jahn, eds. London, Zed Books, Ltd.**:** 274-283.

[Cho99b] Choucri, N. (1999). Strategic Partnerships with Multilingual Functionality for Globalisation and Localisation. European Commission Directorate - General Information Society Workshop on "Sustainability and Environment", IST99, Helsinki, Finland.

[Cho99c] Choucri, N. (1999). The Political Logic of Sustainability. Sustainability and the Social Sciences. Becker and Jahn, eds. New York, Zed Books.

[ChoMM98] Choucri, N., McHugh and Millman (1998). Innovations in Cyberpartnering for Sustainability. Care Innovation '98: Proceedings, Second International Symposium, Brokerage Event and Environmental Exhibition, Vienna, Austria.

[ChoN75] Choucri, N., North, R. C. (1975). Nations in Conflict - National Growth and International Violence. San Francisco, CA, W. H. Freeman and Company.

[ChoN92] Choucri, N., North, R. C., and Yamakage S. (1992). The Challenge of Japan - Before World War II & After. New York, Routledge. [CHS91a] Collett, C., Huhns, M., and Shen, W. (1991). "Resource Integration Using an Existing Large Knowledge Base," *MCC Technical Report Number ACT-OODS-127-91*.

[Elki97] Elkins, D. J. (1997). "Globalization, Telecommunication, and Virtual Ethnic Communities." International Political Science Association **18**(2): 139-152.

[Fark96] Farkas, A. (1996). Evolutionary Models in Foreign Policy Analysis. International Studies Quarterly Hart, Rasler, Thompson, eds. Cambridge, MA, Blackwell Publishers, Inc. **40:** 343-362.

[Gilp96] Gilpin, R. (1996). Economic Evolution of National Systems. International Studies Quarterly. Hart, Rasler, Thompson, eds. Cambridge, MA, Blackwell Publishers, Inc. **40:** 411-432.

[HSG80] Holsti, O.R., Siverson, R.M., and George, A.L. eds. (1980). Change in the International System. Boulder, CO. Westview Press.

[HT99] Herst, P and Thompson, G. (1999). Globalization in Question. Malden, MA, Blackwell Publishers Inc.

[Jerv78] Jervis, R. (1978). "Cooperation under the Security Dilemma." World Politics: 167-214.

[KR96] Kang, H, and Reuveny R. (1996). International Conflict and Cooperation: Splicing COPDAB and WEIS Series. International Studies Quarterly. Hart, Rasler, Thompson, eds. Cambridge, MA, Blackwell Publishers, Inc. **40:** 281-306.

[Mesq00] Mesquita, B. B. (2000). Principles of International Politics - People's Power, Preferences, and Perceptions. Washington, D.C., CQ Press.

[Midl89] Midlarsky, M. I. ed. (1989). Handbook of War Studies, The University of Michigan Press.

[Mode96] Modelski, G. (1996). Evolutionary Paradigm for Global Politics. International Studies Quarterly. Hart, Rasler, Thompson, eds. Cambridge, MA, Blackwell Publishers, Inc. **40:** 321-342.

[NC93] North, R. C. and Choucri, N. (1993). Growth, Development, and Environmental Sustainability : Profiles and Paradox. Global Accord: Environmental Challenges and International Responses. N. Choucri., ed. Cambridge, MA, MIT Press.

[ND00] Nye, J. S. and Donahue, J. eds. (2000). Governance in a Globalizing World. Washington, D.C., Brookings Institution Press.

[OY96] Osherenko, G. and Young, O. (1996). "Testing Theories of Regime Formation - Findings from a Large Collaborative Research Project." The International Political Economy and International Institutions **1**: 573-602.

[PM96] Poznanski, K. and Modelski, G. (1996). Evolutionary Paradigms in the Social Science. International Studies Quarterly. Hart, Rasler, Thompson, eds. Cambridge, MA, Blackwell Publishers, Inc. **40:** 315-319.

[SK92] Sheth, A. and Kashypa, V. (1992). "So Far (Schematically) yet So Near (Semantically)," in Hsiao, D., Neuhold, E., and Sacks-Davis R., ed. *Proc of the IFIP WG2.6 Database Semantics Conference on Interoperable Database Systems (DS-5)*, p. 283-312.

[Stei00] Steinbruner, J. D. (2000). Principles of Global Security. Washington, D.C., Brookings Intitution Press.

[Vasq93] Vasquez, J. A. (1993). The War Puzzle. Cambridge, Cambridge University Press.

[Walt83] Waltz, K. N. (1983). Theory of International Politics. Reading, MA, Addsion-Wesley Publishing Company.

[Zieg84] Ziegler, D. W. (1984). War, Peace, and International Politics. Boston, Little, Brown and Company.