Information Integration for Counter Terrorism Activities: Requirements for Context Mediation[1,2]

Nazli Choucri
Stuart E. Madnick
Allen Moulton
Michael D. Siegel
Hongwei Zhu
MIT Sloan School of Management
30 Wadsworth Street, Cambridge, MA 02142, USA
{nchoucri, smadnick, amoulton, msiegel, mrzhu}@mit.edu
http://context2.mit.edu/coin

Abstract—The National Research Council has noted that "[A]lthough there are many private and public databases that contain information potentially relevant to counterterrorism programs, they lack the necessary context definitions (i.e., metadata) and access tools to enable interoperation with other databases and the extraction of meaningful and timely information." In this paper we present examples of these problems and a technology developed at MIT, called context mediation, which provides a novel approach for addressing these problems.

TABLE OF CONTENTS

1. INTRODUCTION

In the aftermath of the 9/11 tragedy it has became clear that the lack of effective information exchange among government agencies hindered the capability of identifying potential threats and preventing terrorism actions. It has been noted by the National Research Council that "Although there are many private and public databases that contain information potentially relevant to counterterrorism programs, they lack the necessary context definitions (i.e., metadata) and access tools to enable interoperation with other databases and the extraction of meaningful and timely information[3]". This report clearly recognized the important problem that the semantic data integration research community has been studying.

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

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

In this paper we present the COntext INterchange (COIN) technology. We begin in Section 2 with motivation for the requirements for integrating complex sources with different contexts. In Section 3 we present a detailed example of the context problem. In Section 4 we describe the COIN technology. We present a summary of the current status of this technology and on-going research challenges in Section 5. We present some conclusions in Section 6.

[1] 0-7803-8155-6/04/$17.00 © 2004 IEEE
[2] IEEEAC paper #1264, Version Final , Updated December 9, 2003
[3] Emphasis added

0-7803-8155-6/04/$17.00 ©2004 IEEE

Authorized licensed use limited to: MIT Libraries. Downloaded on April 03,2022 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

2. THE CHALLENGES OF CONTEXT IN THE TERRORISM DOMAIN

The important trends of unrelenting globalization, growing worldwide electronic connectivity, and increasing knowledge intensity in economic and social activities create challenging demands for information access, interpretation, provision and overall use. Unless IT advances remain 'one step ahead' of such realities and complexities, strategies for better understanding and responding to emergent global challenges will be severely impeded. For example, the new Department of Homeland Security relies on intelligence information from all over the world to develop strategic responses to a wide range of security threats. However, relevant information is stored throughout the world and by diverse agencies and in different media, formats, quality, and contexts. Intelligent integration of that information and improved modes of access and use are critical to developing policies designed to identify and anticipate sources of threat, to strengthen protection against threats on the United States, and to enhance the security of the nation.

2.1 Emergent Challenges to Global Information

There are critical new challenges to current modes of information access and understanding for counterterrorism. 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. We have learned about the costs of overlooking these challenges through tragic events. There are also significant opportunity costs, which potentially hinder both empirical analysis and theoretical inquiry so central to national policy.

2.2 Integration Requirements for Crisis, Conflicts and Prevention

The examples in Table 1 illustrate the types of information needs required for effective research, education, decision-making, and policy analysis on a range of conflict issues. The information needs in the conflict realm involve emergent risks, threats of varying intensity, and uncertainties of potentially global scale and scope. Three major categories of information requirements are: (a) crisis situations; (b) conflicts and war; and (c) anticipation, monitoring, and early warning. Information needs for research 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 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

Illustrative Cases	Example of Information Needs	Intended Use of Information
1. Strategic Requirements for Managing Cross-Border Pressures in a *Crisis* The UNHCR needs to respond to the dislocation and large numbers of Afghans into neighboring countries, triggered by war in Afghanistan.	Logistical and infrastructure information for setting up refugee camps, such as potential sites, sanitation, and potable water supplies.	Facilitated coordination of relief agencies with up-to-date information during a crisis for more rapid response (as close to real time as possible).
2. Capabilities for Management during an Ongoing *Conflict & War* The goal of the newly established UNEP-Balkans group is to assess whether the ongoing Balkan conflict has had significant environmental and economic impacts on the region. The data, extensive as it may be, is dispersed and presented in different contexts.	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 environmental and economic viability.	Improved decision making during conflicts and war - taking into account contending views and changing strategic conditions - in order to better prepare for, and manage, future developments and modes of resolution.
3. Strategic Response to Security Threats for *Anticipation, Prevention, and Early Warning* The newly-created Department of Homeland Security needs to coordinate U.S. government efforts with foreign governments using information from different regions of the world.	Intelligence data from foreign governments, non-governmental agencies, US agencies, and leading opinion leaders worldwide.	Streamline potentially conflicting information content and sources in order to facilitate coherent anticipation, preventive monitoring, and early warning.

Table 1. Illustrating Information Needs in Three Areas

Authorized licensed use limited to: MIT Libraries. Downloaded on April 03,2022 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

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 routine searches, but require extraction of information from sources that may evolve and change over time. Furthermore, in today's global context, 'preventative action' may even take on new urgency, and create new demands for information services.

All of these issues remain central to matters of security in this increasingly globalized world.

3. Challenges for Integrating Information with Multiple Contexts: A Detailed Example

For illustrative purposes only, this section elaborates on the challenges described above by presenting a detailed example. This example is particularly relevant to the types of problems illustrated by row 2 in Table 1, but it illustrates basic challenges to all areas.

The specific question that we want to address is: **to what extent have economic performance and environmental conditions in Yugoslavia been affected by the conflicts in the region?** The answer to this question could shape policy priorities for different national and international institutions, as well as 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, by any means, but one that illustrates concurrent challenges for information compilation, analysis, and interpretation – under changing conditions.

For example, if we are interested in determining the change of carbon dioxide (CO_2) emissions in the region, normalized against the change in GDP and population - before and after the outbreak of the hostilities – we need to take into account territorial and jurisdictional boundaries, changes in accounting and recording norms, and varying degrees of 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?

An even more subtle issue is: what does the user mean by "Yugoslavia"? Is it the country defined by its post-conflict borders, or the entire geographic area formerly known as Yugoslavia? One of the effects of the war is that the region, which used to be one country consisting of six republics and two provinces, has subsequently been reconstituted into five legal 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 [4].

In this example, 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 original Yugoslavia (see the alternative User 2 scenario in Table 2 for the post0-conflict Yugoslavia). 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?**

3.1 Diverse Sources and Contexts

By necessity, to answer this question, one needs to draw data from diverse types of sources (we call these differing *domains* of information) - such as, economic data (e.g., the World Bank, UN Statistics Division), environmental data (e.g., Oak Ridge National Laboratory, World Resources Institute), and country history data (e.g., the CIA Factbook), as illustrated in Table 2 below. Merely combining the numbers from the various sources is likely to produce serious errors due to different sets of assumptions driving the representation of the information in the sources These assumptions are often not explicit but are an important representation of 'reality' (we call these the meaning or *context* of the information.)

The purpose of Table 2 is to illustrate some of the complexities in attempting to answer 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 the user's information needs. As specified in the table, time T0 refers to a date b*efore 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

[4] To make the problem even more complex, more recently the country "Yugoslavia" disappeared entirely – there is now the "Republic of Serbia and Montenagro."

Authorized licensed use limited to: MIT Libraries. Downloaded on April 03,2022 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

Domain and Sources Consulted

Economic Performance
- World Bank's World Development Indicators database
- UN Statistics Division's database
- Statistics Bureaus of individual countries

Environmental Impacts
- Oak Ridge National Laboratory's CDIAC database
- WRI database
- GSSD
- EPA of individual countries

Country History:
- CIA
- GSSD

Mappings Defined:
- Country code
- Currency code
- Historical exchange rates*

* *Note: Hyperinflation in YUG resulted in establishment of a new currency unit in June 1993. Therefore, T1.YUN is completely different from T0.YUN.*

Sample Data Available

A. Annual GDP and Population Data:

Country	*T0.GDP*	*T0.Pop*	*T1.GDP*	*T1.Pop*
YUG	698.3	23.7	1627.8	10.6
BIH			13.6	3.9
HRV			266.9	4.5
MKD			608.7	2.0
SVN			7162	2.0

- GDP in billions local currency per year
- Population in millions

B. Emissions Data:

Country	*T0*	*T1*
YUG	35604	15480
BIH		1279
HRV		5405
MKD		3378
SVN		3981

- Emissions in 1000s tons per year

T0.{YUG} = T1.{YUG, BIH, HRV, MKD, SVN}
(i.e., geographically, YUG at T0 is equivalent to YUG+BIH+HRV+MKD+SVN at T1)

Country	*Code*	*Currency*	*Currency Code*
Yugoslavia	YUG	New Yugoslavian Dinar	YUN
Bosnia and Herzegovia	BIH	Marka	BAM
Croatia	HRV	Kuna	HRK
Macedonia	MKD	Denar	MKD
Slovenia	SVN	Tolar	SIT

C_From	*C_To*	*T0*	*T1*
USD	YUN	10.5	67.267
USD	BAM		2.086
USD	HRK		8.089
USD	MKD		64.757
USD	SIT		225.93

Basic Question, Information User Type & Usage

Question:
How did economic output and environmental conditions change in YUG over time?

User 1: YUG as a geographic region bounded at T0:

Parameter	T0	T1
CO_2	35604	29523
CO_2/capita	1.50	1.28
GDP	66.5	104.8
GDP/capita	2.8	4.56
CO_2/GDP	535	282

User 2: YUG as a legal, autonomous state

Parameter	T0	T1
CO_2	35604	15480
CO_2/capita	1.50	1.46
GDP	66.5	24.2
GDP/capita	2.8	1.1
CO_2/GDP	535	640

Note:

T0: 1990 (prior to breakup)
T1: 2000 (after breakup)
CO_2: 1000's tons per year
CO_2/capita: tons per person
GDP: billions USD per year
GDP/capita: 1000's USD per person
CO_2/GDP: tons per million USD

represents the Yugoslavian Dinar. The third column lists the outputs and units 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.

3.2 The Manual Approach

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

For time T0:
1. Get CO_2 emissions data for "YUG" from source B;
2. Convert it to tons/year using scale factor 1000; call the result X;
3. Get GDP data from source A;
4. Convert to USD by looking up currency conversion table, an auxiliary source; call the result Y;
5. No need to convert the scale for GDP because the receiver uses the same scale, namely, 1,000,000;
6. Compute X/Y (equal to 535 tons/million USD in Table 2).

For time T1:
1. Consult source for country history and find all countries in the area of former YUG;
2. Get CO_2 emissions data for "YUG" from source B (or a new source);
3. Convert it to tons/year using scale factor 1000; call the

Authorized licensed use limited to: MIT Libraries. Downloaded on April 03,2022 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

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).

The complexity of this task would be easily magnified if, for example, the CO_2 emissions data from the various sources were all 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 example 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, time-consuming, and error-prone. This makes it difficult under "normal" circumstances and likely impossible under time-critical circumstances.

3. 3 The Challenges for Counter-Terrorism Information Integration: Information Extraction, Dissemination, and Interpretation

We will now be more detailed about the information challenges that must be addressed:

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 [20], we will continue to live in a very heterogeneous world for quite a while to come. So we must be able to easily and rapidly extract information from all types of sources.

Information Dissemination: The users want to use the resulting "answers" in many 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, or application program for further processing. So we must be able to disseminate the information in many ways.

Information Interpretation: Although the problems of information extraction and dissemination are difficult, the most difficult challenges involve information interpretation, as introduced above and elaborated below.

Let us reconsider our 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 emission per GDP changed in Yugoslavia after the war?" may have multiple interpretations. Not only does each source have a context, but also 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 user might have a different preferred context for their answer, such as: tons/million USD or kilograms/billion EURO, etc. There are many information harmonization challenges.

Although seemingly simple, this example addresses some of the most complex issues in international relations: 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 is interested in demographic distributions around oil fields (by ethnic group) and before and after independence. Alternatively, UNEP is interested in CO_2 emissions per capita given that these are oil-producing regions. On the other hand, foreign investors will be interested in insurance rates before and after independence.

Authorized licensed use limited to: MIT Libraries. Downloaded on April 03,2022 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

The information shown as footnotes 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 many times it must be found in other sources, and on occasion someone must be asked to track down and explain the meaning of the data. The good news is that such context knowledge almost always exists, but it is often widely *distributed* within and across organizations. Thus, a central focus of semantic data integration technology is to support *the acquisition, organization, and effective intelligent usage of distributed context knowledge to support information harmonization and collaborative domains.*

4. A BETTER WAY: THE CONTEXT INTERCHANGE APPROACH

A key goal of our research is to create a system that can automatically determine and reliably perform the steps shown in Section 3.2 to answer such a user's request – thereby reducing the time delay from hours to seconds. The COntext INterchange (COIN) System, shown below, is such a system. COIN is capable of storing the necessary context about the sources and receivers of information. It has a reasoning engine capable of determining the necessary sources, conversions, and calculations.

The COIN Project has developed a basic theory, architecture, and software prototype for supporting intelligent information integration employing context mediation technology [11,12,13,21], It also has support tools to allow for applications' (i.e. receivers') context definition and source definitions to be added and removed easily (i.e., schemas, contexts, capabilities). COIN is a mediation approach [23,24] for semantic integration of disparate (heterogeneous and distributed) information sources as described in [2,11]. 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.

The wrappers are physical and logical gateways providing uniform access to the disparate sources over the network [3,8,9]. 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, and results are composed and sent to the receiver.

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 programmer's assumptions and knowledge) into one or more queries in the terms of the component sources. The individual subqueries may still involve several sources. However, subsequent planning, optimization and execution phases are needed [1, 10]. The planning and execution phases must consider the limitations of the sources and the

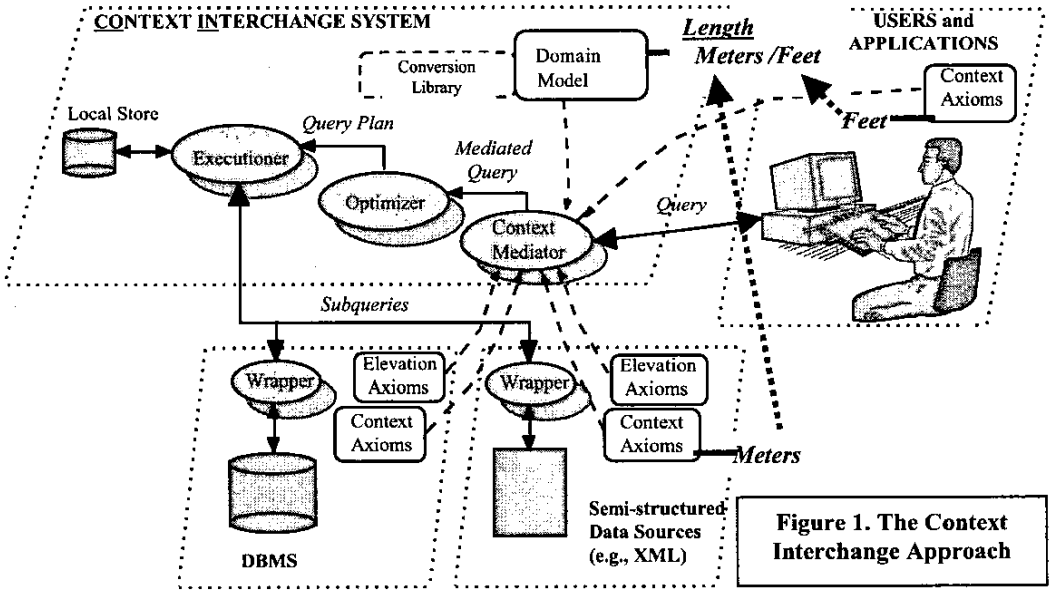

Figure 1. The Context Interchange Approach

Authorized licensed use limited to: MIT Libraries. Downloaded on April 03,2022 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

topology and costs of the network (especially when dealing with non-database sources, such as web pages or web services). The execution phase is in charge of the scheduling of the query execution plan and the realization of the complementary operations that could not be handled by the sources individually (e.g. a join across sources).

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 and loose-couple approaches are not suitable for such an environment.

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 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.

In contrast, the COIN approach is a middle ground between these two approaches. It allows queries to the sources to be mediated, i.e. semantic conflicts to be identified and solved by a context mediator through comparison of contexts associated with the receivers and sources associated with the queries. It only requires the minimum adoption of a common Domain Model that defines the domain of discourse of the application.

The knowledge needed for harmonization is formally modeled in a COIN framework [13], 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 [5,15]. 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 or from a receiver's point of view.

The comparison and conversion procedure itself is inspired by the Abductive Logic Programming framework [14] and can be qualified as an abduction procedure, to take advantage of its formal logical framework. One of the main advantages of the abductive logic programming framework is the simplicity in which it can be used to formally combine and to implement features of query processing, semantic query optimization and constraint programming.

We use a set of web-based authoring tools [16] to create and manage the ontology, the elevation axioms, and context definitions, which we call the knowledgebase for the application. This tool also imports RDF and exports RDF [18,19]. By this means we can utilize ontologies developed by other applications. The tool provides both a text-based and a graphical interface. Using this tool we gain the ability to develop context knowledge and to add easily new sources and to modify context.

The scalability of COIN architecture has been greatly extended by a new feature to allow for allow for application merging [6]. Applications are usually developed in particular domains of interest. It is important that the effort to develop these applications and associated domain models be reusable in other applications that may draw from one or more application domains. Our application merging technology reuses existing ontologies and enables easy creation of large applications by merging multiple smaller ones. Unlike other approaches we utilize existing domain models intact. We have developed a tool that creates merging axioms that reside with the new application and operate over existing ontologies and contexts.

5. Status of Project and On-going Research

We have demonstrated these context capabilities in a number of application domains, such as financial services [7], online shopping [26], disaster relief efforts [16], corporate house holding knowledge engineering [26], and larger applications built by combining existing ones (*e.g.*, combine an airfare aggregator and a car rental shopper into a travel planner, see demos at our website). Efforts are also underway to use COIN framework as a cost effective alternative to standardization in the financial industry. In addition, we have developed a .NET version of web wrapper and performed a preliminary study on accessing data and methods using Web Services. Progress in these areas will make COIN technology available to the Semantic Web community. Other planned extensions, such as temporal context, will further improve the applicability of COIN technology for various data integration needs.

This focus on context knowledge and data integration has allowed us to make significant progress, however, challenges exist in making such an approach fully scalable, maintainable and usable in an open environment. Our on-

Authorized licensed use limited to: MIT Libraries. Downloaded on April 03,2022 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

going research is addressing some of these extensions, as briefly described below..

5.1 Extended Domain of Knowledge – Equational and Temporal Context.

In addition to the types of domain and context knowledge currently handled by the COIN framework, we need to perform additional research to add capabilities for both the representation and reasoning to provide support for equational [FGM02] and temporal 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.

Temporal context refers to the fact that context not only varies across sources but also across time. Thus, the implied currency context for France's GDP prior to 2002 might be French Francs but after 2002 it is in Euros. If one were performing a longitudinal study over multiple years from multiple sources, it is important that this variation in context over time be understood and processed appropriately.

5.2 Advanced Mediation Reasoning and Services

The COIN abductive framework can also be extrapolated to problem areas such as integrity management, view updates and intentional updates for databases [4]. Because of the clear separation between the declarative definition of the logic of mediation into the COINL program from the generic abductive procedure for query mediation, 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.

The COIN approach holds the knowledge of the semantics of data in each context and across contexts in declarative logical statements separate from the mediation procedure. 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, 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, F-Logic and abductive reasoning, and extend the expressiveness and the reasoning capabilities leveraging ideas developed in different yet similar frameworks such as Description Logic and classification.

5.3 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 [17]. 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).

5.4 Gathering, Representing and Maintaining Context Knowledge for Unknown Tasks

Context Interchange capabilities have been used for specific applications. Though the semantic integration can ontologies developed in RDF to include modifiers and other context information. However, we expect a wide range of ontology languages and representations, context information must either be easily extracted from these ontologies or added through the use of context-authoring tools as developed on this project. Tools are needed to automatically assemble and maintain context knowledge.

6. CONCLUSIONS

The effective pursuit of counter-terrorism activities requests the rapid and semantically meaningful integration of information from diverse sources. Fortunately, context mediation technology offers the potential of addressing these needs. Further advances will allow this technology to be used effectively in conflict, crisis and prevention modes

Authorized licensed use limited to: MIT Libraries. Downloaded on April 03,2022 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

of counter-terrorism. We look forward to applying this technology to specific implementations in this area.

ACKNOWLEDGEMENTS

Work reported herein has been supported, in part, by Banco Santander Central Hispano, Citibank, Defense Advanced Research Projects Agency (DARPA), D & B, Fleet Bank, FirstLogic, Merrill Lynch, MITRE Corp., MIT Total Data Quality Management (TDQM) Program, PricewaterhouseCoopers, Singapore-MIT Alliance (SMA), Suruga Bank, and USAF/Rome Laboratory.

REFERENCES

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

[2] S. Bressan, C. Goh, N. Levina, S. Madnick, S. Shah, M. Siegel, "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, 2000.

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

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

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

[6] A. Firat, "Information Integration Using Contextual Knowledge and Ontology Merging" PhD Thesis, 2003.

[7] A. Firat, B. Grosof, S. Madnick, "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, 2002.

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

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

[10] K.D. Fynn, "A Planner/Optimizer/Executioner for Context Mediated Queries," MS Thesis, MIT, 1997.

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

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

[13] C.H. Goh, "Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Systems, PhD Thesis, MIT Sloan School of Management," 1997.

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

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

[16] P.W. Lee, "Metadata Representation and Management for Context Mediation," MIT EECS Master's Thesis, 2003.

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

[18] S. Liburd, "Conceptual Mapping and Structural Conversions between COIN and RDF," technical report, MIT Sloan School of Management, 2003.

[19] F. Manola, E. Miller, "RDF Primer", W3C working draft, 2003.

[20] S. Madnick, "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, 2001.

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

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

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

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

[25] X. Xian, Corporate Householding Knowledge Engineering and Processing using Extended COIN, MIT EECS Master's Thesis, 2003.

[26] H. Zhu, S. Madnick, M. Siegel, "Global Comparison Aggregation Services", *1st Workshop on E-Business*, December 14-15, Barcelona, Spain, 2002.

Authorized licensed use limited to: MIT Libraries. Downloaded on April 03,2022 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

BIOGRAPHIES

Dr. Nazli Choucri is Professor of Political Science at the Massachusetts Institute of Technology, and Director of the Global System for Sustainable Development (GSSD) a distributed multi-lingual knowledge networking system to facilitate uses of knowledge for the management of dynamic strategic challenges. To date, GSSD is mirrored (i.e. synchronized and replicated) in China, Europe, and the Middle East in Chinese, Arabic, French and English. As a member of the MIT faculty for over thirty years, Professor Choucri's area of expertise is on modalities of conflict and violence in international relations. She served as General Editor of the International Political Science Review and is the founding Editor of the MIT Press Series on Global Environmental Accord. The author of nine books and over 120 articles Professor Choucri's core research is on conflict and collaboration in international relations. Her present research focus is on 'connectivity for sustainability', including e-learning, e-commerce, and e-development strategies. Dr. Choucri is Associate Director of MIT's Technology and Development Program, and Head of the Middle East Program. She has been involved in research, consulting, or advisory work for national and international agencies, and in many countries, including: Abu Dhabi, Algeria, Canada, Colombia, Egypt, France, Germany, Greece, Honduras, Japan, Kuwait, Mexico, North Yemen, Pakistan, Qatar, Sudan, Switzerland, Syria, Tunisia, Turkey

Dr. Stuart Madnick is the John Norris Maguire Professor of Information Technology, Sloan School of Management and Professor of Engineering Systems, School of Engineering at the Massachusetts Institute of Technology. He has been a faculty member at MIT since 1972. He has served as the head of MIT's Information Technologies Group for more than twenty years. He has also been a member of MIT's Laboratory for Computer Science, International Financial Services Research Center, and Center for Information Systems Research. Dr. Madnick is the author or co-author of over 250 books, articles, or reports including the classic textbook, Operating Systems, and the book, The Dynamics of Software Development. His current research interests include connectivity among disparate distributed information systems, database technology, software project management, and the strategic use of information technology. He is presently co-Director of the PROductivity From Information Technology Initiative and co-Heads the Total Data Quality Management research program. He has been active in industry, as a key designer and developer of projects such as IBM's VM/370 operating system and Lockheed's DIALOG information retrieval system. He has served as a consultant to corporations, such as IBM, AT&T, and Citicorp. He has also been the founder or co-founder of high-tech firms, including Intercomp, Mitrol, and Cambridge Institute for Information Systems, iAggregate.com and currently operates a hotel in the 14th century Langley Castle in England. Dr. Madnick has degrees in Electrical Engineering (B.S. and M.S.), Management (M.S.), and Computer Science (Ph.D.) from MIT. He has been a Visiting Professor at Harvard University, Nanyang Technological University (Singapore), University of Newcastle (England), and Technion (Israel), and Victoria University (Australia).

Allen Moulton is a member of the research team of the Context Interchange Project at the MIT Sloan School of Management. His research focuses on applying automated reasoning technology to the problem of resolving semantic differences in the interchange of information among autonomous, heterogeneous data sources and receivers. In this research, he draws on over thirty years of industry experience in information technology consulting with an emphasis on systems for the fixed income securities industry. He has previously been involved in research on centralization and decentralization of information systems at the MIT Center for Information Systems Research and on applications of information technology to substantive foreign policy analysis at the MIT Center for International Studies. He is co-author of the book Managing International Conflict: from Theory to Policy, which includes the CASCON computer software for analyzing international conflict.

Dr. Michael Siegel is a Principal Research Scientist at the MIT Sloan School of Management. He is currently the Director of the Financial Services Special Interest Group at the MIT Center For eBusiness. Dr. Siegel's research interests include the use of information technology in financial risk management and global financial systems, eBusiness and financial services, global ebusiness opportunities, financial account aggregation, ROI analysis for online financial applications, heterogeneous database systems, managing data semantics, query optimization, intelligent database systems, and learning in database systems. He has taught a range of courses including Database Systems and Information Technology for Financial Services. He currently leads a research team

Authorized licensed use limited to: MIT Libraries. Downloaded on April 03,2022 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

looking at issues in strategy, technology and application for eBusiness in Financial Services.

Mr. Hongwei Zhu is a PhD candidate in the Technology, Management and Policy Program and a Research Assistant on the Context Interchange Project at Sloan School of Management. He develops technologies to enable meaningful information sharing. He also develops theories to address policy issues related to information sharing. Prior to coming to MIT, Mr. Zhu was a software engineer and IT consultant developing web based solutions for both private sector and government agencies.

Authorized licensed use limited to: MIT Libraries. Downloaded on April 03,2022 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.