

Robust Mediation of Construction Supply Chain Information

William O'Brien¹ and Joachim Hammer²

Abstract

We describe the architecture of a robust and flexible mediation and wrapping infrastructure to support access to and sharing of semantically heterogeneous process information in the construction supply chain. This infrastructure allows us to collect and share process related information from numerous legacy systems, each with its own interface and schema (that may or may not subscribe to data standards). Our research extends current capabilities to allow automated mediation across a large number of sources and with a high degree of semantic heterogeneity. In particular, we extend the query processing capabilities of the mediator to make use of domain knowledge provided by experts and domain knowledge in the information sources accessed. Accomplishing this requires formalization of knowledge about the representation and use of information by construction subcontractors and suppliers, and development of theory about knowledge transformation in construction. Our mediation approach extends existing integration theories and is a novel complement to existing efforts in data standardization, embracing the heterogeneity of information representation in the supply chain.

Introduction

This paper describes a set of enabling technologies to support computerized analysis of supply chain operations in the AEC industry. Today, there exist a growing number of analytic tools capable of analyzing complex supply chains to provide both operational and strategic decision support. However, all these tools need data to provide useful results; it is our view that obtaining this data from firms in the decentralized AEC industry is one of the most difficult tasks on projects. Automating assembly of this information is a requisite challenge to computerized supply chain analysis. More broadly, automating assembly of distributed information is a necessary component of the FIAPP vision.

Recent years have seen the development of data standards such as the IFC (IAI 1996) and AECXML (aecXML 1999) to facilitate automated sharing of data by applications designed to operate on those standards. While useful, we view these standards as limited and incapable of sharing all the data needed for supply chain analysis. Our worldview and assumptions are similar to those of Zamanian and Pittman (1999) of Autodesk who bring a balanced perspective of theory and practice. They suggest that the rigid approach to specification of data standards provided by the IFC and similar approaches will “prove most useful for very general and very specific AEC information but not the information that falls between these two extremes.” (p. 224) Given the temporal and multi-perspective use of information on AEC projects, they suggest that a flexible schema (e.g., similar to the Semantic Modeling Extension by Clayton et al. 1996)) is more appropriate for

¹ Dept. of Civil & Coastal Engineering, 345 Weil Hall/PO Box 116580, University of Florida, Gainesville, FL 32611-6580, 352-392-7213, wjob@ce.ufl.edu

² Dept. of Computer and Information Sciences and Engineering, 301 CSE Building/PO Box 116120, University of Florida, Gainesville, FL 32611-6120, jhammer@cise.ufl.edu

information management in the AEC industry. Zamanian and Pittman further suggest that what will evolve over time is not a universally accepted standard but a multiple protocols suitable for use by certain disciplines at a specific phase of the project. The development of trade group standards such as CIMsteel Integration Standards (<http://www.cis2.org/>) supports their view. We note that CIMsteel and the IFC, while each developed from STEP constructs, are not interoperable. Considerable future development work is required to make them so (Crowley 1999).

Practically, we believe that it is implausible that all the firms in a project supply chain will uniformly subscribe to a single data standard. Birrell (1980) notes that on a given project there may be 6-60 subcontractors. As most subcontractors will have many suppliers (and, in turn, suppliers have sub-suppliers), the chance that a significant number of firms in a broad supply chain will subscribe to a uniform data standard is low. It is more likely that groups of firms will develop standards such as CIMsteel for their use for specific applications. Moreover, as much process information in the supply chain relates to cost, time and production capabilities (i.e., much of firms' core operational and competitive information), it is likely that many firms will prefer to use legacy applications to manage this information, avoiding the expensive transition to new applications. This makes connection to heterogeneous information systems and translation of semantically heterogeneous data in those systems dual challenges that must be overcome to support automated supply chain analysis.

Further, there are challenges in knowledge composition. Beyond an ability to connect to and provide basic translation of data in legacy systems in the supply chain, raw data must often be transformed to a form suitable for decision making. Consider that much of the data used for operations in firms is detailed in nature, often mimicking accounting details or a detailed work breakdown structure (Barrie and Paulson 1992). This data is too detailed for most supply chain decision support models (see for example, the models in Tayur et al. 1999)). More broadly, we must compose the data needed as input for analysis tools from data used by applications with other purposes in mind. This knowledge composition may involve considerable pre- and post-processing of data in firms' legacy systems before it is usable by supply chain analysis tools.

Our proposed infrastructure addresses these collective challenges. We describe a mediation and wrapping infrastructure that can connect to databases at the various firms involved on a project, automatically extract supply chain related information (in particular, cost and resource data), and provide this in a well defined form for further use. Our contribution to practice is to remove details of the location and representation of the underlying data from end users and higher level application, allowing them to focus on processing data rather than collecting it. Our contributions to knowledge are the theory and methods behind a uniquely robust mediation infrastructure.

Overview of Information Infrastructure

Our proposed mediation infrastructure and its relation to other components in the AEC information environment is shown in Figure 1. End users require various answers to managerial questions. These are typically provided by higher level applications (in our case, supply chain analysis tools), which in turn need information distributed in various firms. Our mediation layer exists between firms' data and applications. Here we note that

some of the output of the mediator may be shown directly to end users who are principally concerned with the assembly of information and less with subsequent processing by applications. The main task of the layer is presenting mediated information to clients (applications or end users). Clients can then focus on asking questions and exploring the results and ignore where the information comes from, how it is locally represented, and how it is assembled. We note that the proposed mediation layer is not envisaged as a multi-purpose system; rather it will allow connection of specific classes of applications such as supply chain tools with data in firms. There may be multiple mediation systems for different types of applications; this is shown in figure 2 where multiple wrappers connect to a single firm.

The wrappers in figure 1 connect the mediation layer with sources in each firm in the supply chain. Wrappers perform the important tasks of translating raw data in the language of firm to the internal language of the mediation layer. Semi-automatic instantiation of wrappers is an important aspect of this research as without rapid generation of wrappers it will be difficult to establish rapid connection of supply chain analysis tools to firm data. Current commercial applications such as Bentley’s PlantSpace Integration Server (<http://www.bentley.com>) allow only manual generation of wrappers. Some research tools exist to semi-automate the generation of wrappers (e.g., Hammer et al. 1997a) although these tools must be made significantly more robust to accommodate the degree of heterogeneity in AEC applications.

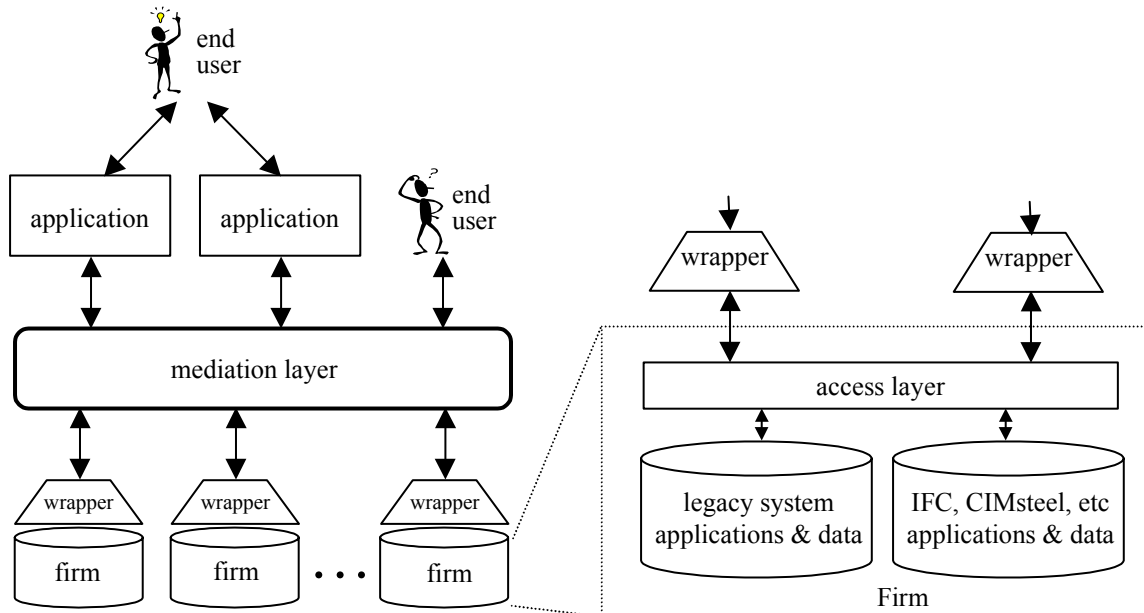


Figure 1: Overview of mediation infrastructure - mediation layer and wrappers in relation to firms' system and applications that generate queries

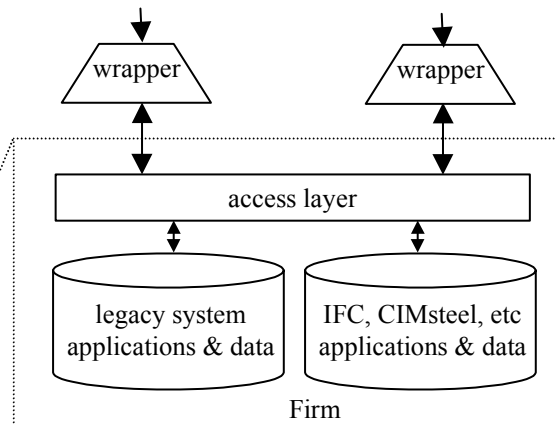


Figure 2: Details of firm information system and connection to multiple wrappers/mediation layers via an access layer with security set by the firm

Figure 2 shows in more detail the structure of systems within the firm. A firm will likely have multiple sources of data and applications. Some of these may be built from

standards such as the IFC, CIMsteel, etc. Many others will be firm specific legacy systems from older applications and applications developed in-house. We specifically show the access layer between the firm's applications and the wrappers. Through the access layer, the firm will specify security and access privileges that will restrict what the wrapper can access. We expect that privileges will vary greatly across firms in the supply chain. This presents a significant challenge to mediation as varying the level of detail of data available makes it difficult to maintain consistent responses to queries.

Gathering information and transforming it for use by higher level applications is the role of our proposed mediation infrastructure. Mediation is a value-added service (Wiederhold 1992). The services of mediators broadly include: Assembly or fusion of information from distributed data sources, transformation of that data into another form (often increasing the information density by combining or summarizing raw data), and cleansing the data to improve its reliability or accuracy (e.g., resolving conflicting information, identifying and interpolating missing data, providing a confidence level of the mediated information). Beyond Wiederhold's description of mediation tasks, we add a value-added wrapper to the mediation infrastructure. The wrapper resolves heterogeneities at the hardware and access layer level including establishing a connection to the legacy source, converting queries from the mediator into requests supported by the API of the access layer, and translating the query result into a format that is consistent with the integrated schema used by the mediator.

Example of Mediation Needs in Construction Supply Chain Analysis

Supply chain models generally address the coordination of the production activities of multiple firms to improve system performance (Cohen and Huchzermeir 1999; Davis 1993). Supply chain models cover a broad spectrum of approaches and objective functions. In general, these models use underlying cost, resource, and productivity information from firms to compute overall supply chain performance characteristics. It serves our purpose to consider a simple scheduling example as, under the definition above, classic network scheduling models and optimization techniques (Antill and Woodhead 1990; Fondahl 1991) are a form of supply chain model.

Consider as a (simple) example, two subcontractors follow each other around a roof edge of a building. (This example is taken from O'Brien (1995).) The first installs a sheet metal gutter around the roof edge. The gutter must be installed before the roofing subcontractor can do its work; the gutter acts as a stopper for an ice and water shield and as a kick-up plate for the base course of roofing tiles. Thus there is a finish-to-start precedence constraint between the work of the sheet metal subcontractor and the work of the roofing subcontractor. We show in table 1 the relevant data from the sheet metal and roofing subcontractors, with its representation and corresponding location.

Firm	Type	Value	Location
Roofing	Production rate	0.9 squares/man-day	Subcontractor database (e.g., Timberline)
	Crew size	11 workers	Subcontractor database
	Crew cost	\$250/man day	Subcontractor database
	Crew composition	4 journeymen 7 apprentices	Payroll database
Sheet Metal	Production rate	20 feet/hr	Subcontractor database
	Crew size & composition	1 foreman	Subcontractor database
	Crew cost	\$50/hr	Subcontractor database

Table 1: Sample data for roofing and sheet metal subcontractors.

The data in table 1 is needed by scheduling applications (more broadly, many supply chain applications will use this data.) It is a straightforward task for humans to process the information in table 1 and answer questions about productivity, crew composition, etc. It is difficult for computers to automatically collect and process this information for use by higher level applications such as scheduling software. Several challenges in mediation are quickly apparent from table 1:

1. For these two firms, there is a large amount of data to mediate, located in three different locations. In order to access the desired data, the mediator needs to establish a connection to each source using a wrapper, query for the relevant data using three separate queries, and merge the result before returning it to the client. Accomplishing these seemingly simple tasks requires detailed information about the underlying sources, their structure and contents.
2. Units are not consistent between firms. The roofing subcontractor produces squares, an area measure, and works in man-days, whereas the sheet metal subcontractor produces in linear feet per hour. Similarly, cost is considered in days and hours. Overcoming these so-called semantic heterogeneities at the data level requires context-specific transformations (from one format into another) which must be encoded in reusable, computer-executable programs. Once the mediator has identified the particular type of conflict, it can apply the proper transformation to convert the values into the same format in order to present units in a consistent manner to higher level applications.
3. Information is aggregated at different levels. The roofing subcontractor records a crew as eleven workers, whereas the sheet metal subcontractor records a crew as one foreman. Foreman is a more meaningful label than worker, and the mediator should be able to parse this and pass on the correct information to higher level applications where the distinction between foreman and other classes of worker is meaningful. Besides semantic knowledge about the structure and context of the data, this example also illustrates the need for a mediation data model that is flexible and expressive enough to accommodate and represent all of the data encountered in the sources.

4. Related to representing data at different levels of aggregation, there is a data fusion and cleansing problem. To obtain values for crew composition for the roofing subcontractor, the mediator must go to a different database (payroll), and look up values for the workers. Here, the classes (journeyman and apprentice) may not be clearly recorded, but could be inferred from difference in pay. The mediator would require some basic information about crew composition and rules for how to classify workers based on pay. This example clearly outlines one of the most challenging problems in mediation: (a) identifying related objects which have different names or descriptions, (b) joining data values across sources without global identifiers, (c) extracting partial data from complex fields. In the absence of sufficient knowledge and unique identifiers, inferencing must be used to solve problems (a) and (b).

These challenges get progressively more difficult, although all fall within Wiederhold's (1992) mediation tasks of fusion, transformation, and cleansing. Challenge 4 (above) blurs the line between traditional mediation services and knowledge composition. Clearly, the mediation layer requires considerable reasoning ability to parse and assemble data about crew composition across various sources.

A further challenge in knowledge composition for scheduling/supply chain applications comes from a need for pre-and post-processing of information in the firm. Consider an extension to traditional scheduling project activities with direct cost curves in a time-cost trade-off framework (Antill and Woodhead 1990). O'Brien and Fischer (2000) note that a subcontractor's resource commitments to projects, known as its capacity utilization, will affect its direct cost curves. This is shown in figure 3. Because these curves are dependent on resource use, they will likely need to be computed for each query. Answering that query may require an application to be invoked at the firm (pre-processing). The results from the application at the firm may require further development to put in a format suitable for the querying application (post-processing). Baring a computational ability to generate cost-time curves in the firm (or a limited ability to access data in the firm due to security restrictions), the mediation layer may need to generate the curves from raw data in the firm.

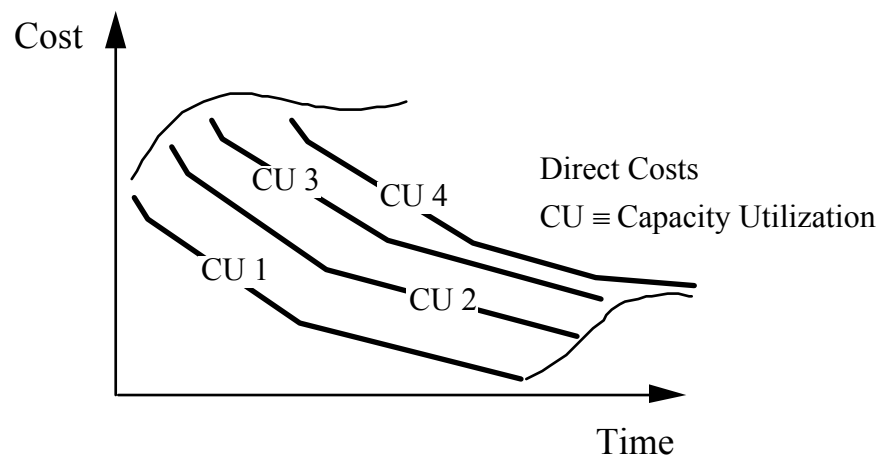


Figure 3: The direct cost curves for a project activity will vary with a subcontractor's capacity utilization (CU 1-4).

Implementation Discussion and Example

To-date, we have built a proof-of-concept wrapper system that connects to and extracts process information from an MS-Project database. MS-Project is representative of the type of information sources in AEC firms. Further, the data in MS-Project supports implementation of the example of the previous section. We have also implemented simple mediation tasks such as the ability to rewrite the incoming query into multiple source queries that fetch the necessary data as well as the ability to conduct resource planning on the extracted data on behalf of the client. The former was necessary since the knowledge about which exact data items were needed was not known until the resource planning task had started.

We first discuss our choice of data model that is used to represent the heterogeneous process data inside the mediator and the wrappers after it has been extracted from the sources. The major challenge is to use a data model that is flexible and extensible enough to accommodate the heterogeneous, semantically rich data from the different firms. To date, there exist several data models suitable for representing heterogeneous information in an integrated system (e.g., Stanford's Object Exchange Model (Papakonstantinou et al. 1995), ODMG's object model (Cattell June 1994), rule- and logic-based systems such as KIF (Genesereth and E 1992) and KQML (McKay et al. 1990), and lately XML (eXtensible Markup Language) (Connolly 1997)). We believe XML together with the Document Object Model (DOM) API (W3 Consortium 1998) gives us the most flexibility for storing, manipulating, and exchanging semantically rich legacy data (Hammer et al. 1997b).

Our mediation architecture is depicted in figure 4. The mediator consists of a Query Processing Module (represented by the dashed box on the left-hand side of the mediator) responsible for decomposing a request for data into the subqueries which are submitted to the relevant data sources, and a Data Merge Engine (represented by the dashed box on the right-hand side of the mediator) responsible for integrating, cleansing, and reconciling the result data that is returned from the firms via wrappers.

Within the Query Processing module, the *Query Translation component* parses and translates the incoming queries (which may be formulated in a high-level language such as SQL) into the internal format based on XML. The *Query Decomposition component* performs the query rewriting of the client query into one or more firm-specific mediated subqueries. This re-writing is necessary since the data model and representation used by the firms is different from the integrated view of the data that is supported by the mediator. As a result, the mediated subqueries are expressed using the schema and terminology used by the underlying firms. Query re-writing is a challenging task and we continue to use as much as possible of existing techniques in this area (e.g., Genesereth and Duschka 1997).

When the result is returned by the wrappers, the Data Merge Engine produces the integrated result that is returned to the client application. Specifically, the *Result Fusion component* joins related data based on a set of merging rules. Data restructuring is done by the *Cleansing & Reconciliation component* and includes the removal of duplicate information, resolution of conflicts, as well as the aggregation and grouping of information into high-level units.

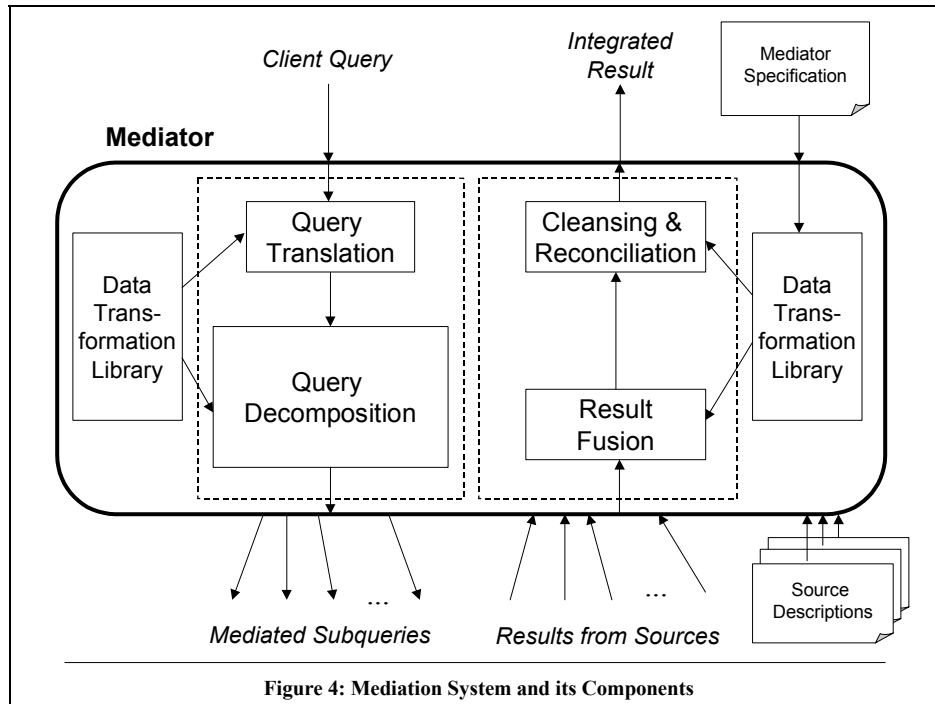


Figure 4: Mediation System and its Components

The mediation knowledge for query processing and data merging is provided to the mediator in three ways: (1) In the form of a high-level mediation specification which contains the query decomposition and rewrite plans, (2) in the form of data transformation library which contains application-specific transformations and conversion routines and, (3) in the form of source descriptions which specify the structure and contents of the legacy systems in the firms and the supported queries.

Both the mediation specification and the data transformation library are compiled into the mediation system at build time (when the mediator is instantiated). The source description is provided by each firm (using a pre-determined format) and is parsed by the mediator during initialization. Note that the mediation specification must also contain a list of firms and their locations so that the mediator knows from where to obtain the source descriptions.

An important prerequisite for mediating process data as described above is the ability to access firm-specific information through value-added wrappers. The main functions of the wrappers in existing information management architectures are to provide a low-level access mechanism to a data source, and to overcome syntactic and semantic heterogeneities to enable the exchange of data between the source context and the wrapper context. These wrappers must be (1) deployable and configurable with minimal effort and technical expertise, (2) able to interface with a wide range of legacy sources including business process systems and hence must be able to support exchange of knowledge such as rules and processes in addition to just data, and (3) provide value-added services to the firm, such as providing functionality that may be missing at the firm but necessary to participate in the project (a necessary service to smaller firms that may lack an ability to invest in expensive information infrastructure).

A sample query between a high level supply application (e.g., master schedule) and one of the firms (e.g., the roofing subcontractor) using the prototype wrapper is shown in figure 5. We assume the contractor is interested in all feasible start/end date combinations for an existing activity (AID = 4) in the sample construction project (PID = 1). The activity start date is left blank which assumes the current date as the first possible start date (see input window in upper left-hand corner); furthermore, in the absence of a cost parameter, the effort is to remain the same. The client query is represented in XML and transmitted to the wrapper (shown in the upper right-hand window). The answer, shown in its XML representation (lower right window) and formatted by the GUI (lower left window) consists of a set of feasible start/finish dates. The results shown in figure 5 support the construction of curves shown in figure 3.

This seemingly simple example requires advanced analysis capabilities in the wrapper which must be tightly coupled with the underlying query mechanism. For example, to compute the resource allocation above, the execution plan in the wrapper generates multiple SQL queries against the firm's MS-Project database to retrieve for the given activity and project all available resources. Subsequently, for each resource the wrapper queries for their current allocations. Using this data, the analysis module computes new allocation plans and the resulting start and end dates for the activity. Although the wrapper provides some of the functionality of MS-Project, there is no way to invoke this functionality through MS-Project's proprietary access layer. Therefore, we must provide this capability through the wrapper-mediator combination to allow appropriate post-processing of data from the firm to enable consistent response to queries.

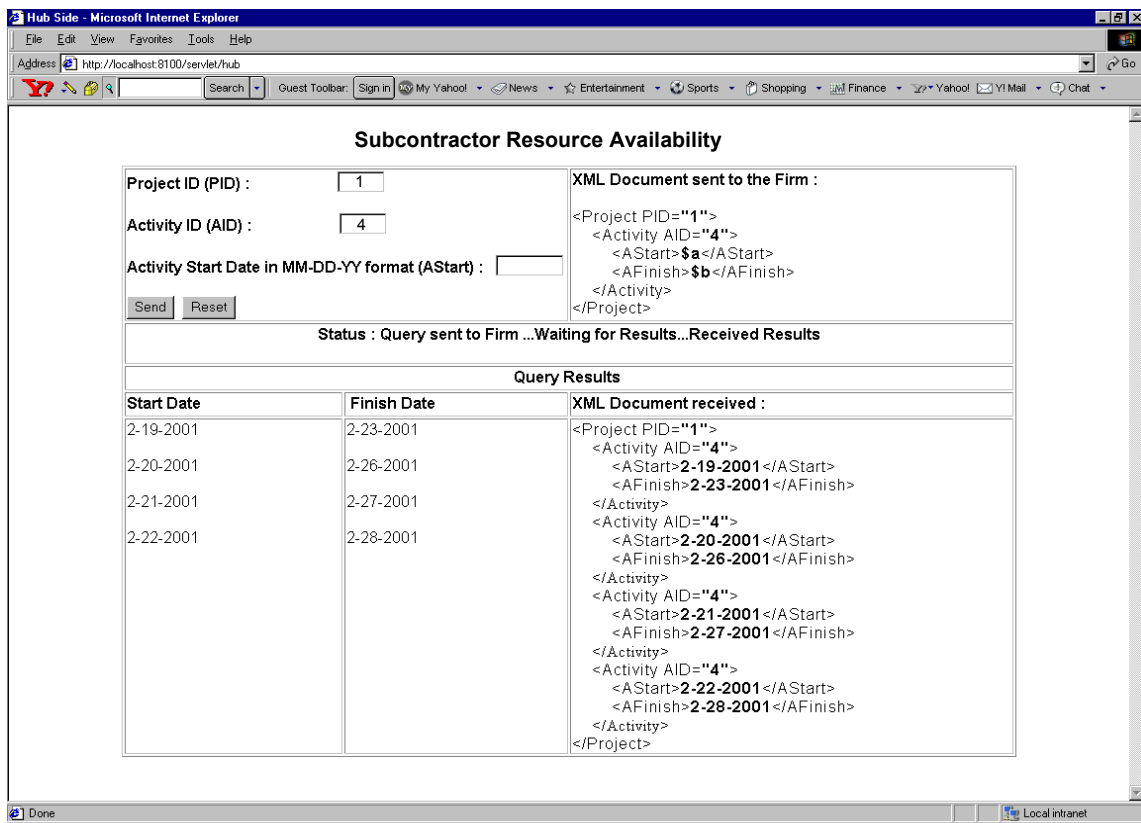


Figure 5: Screen-shot of contractor-subcontractor interaction using prototype wrapper/mediator.

Conclusions and Future Research

We described above a mediation and wrapper infrastructure in the context the broader AEC information environment. While we have specifically focussed on supply chain applications, our proposed infrastructure can serve as a model for different types of knowledge extraction and composition. Hence, our model and development of associated knowledge and technologies broadly supports the FIAPP vision by providing a basis for automating collection and composition of distributed AEC data.

Building from a paper example showing the mediation and wrapping challenges related to AEC knowledge composition, we have implemented a value-added wrapper prototype capable of extracting process knowledge from an MS-Project database. This serves as a prototype from which to build a broader mediation and wrapper infrastructure. In order to provide maximum scalability towards the development of a more robust system, we have carefully separated the wrapper components into three groups: (1) client-specific components, (2) wrapper internal components and, (3) firm-specific components. By modularizing the wrapper architecture in this way, we can make changes to one part of the code without impacting rest of the wrapper. The basic wrapper which connects and queries the MS-Project database (stored in MS-Access) can be instantiated quickly using our wrapper generation toolkit (see Hammer 1999; Hammer et al. 1997a). However, additional manual effort is needed to configure the integration of the operational knowledge of MS-Project with the analysis module in the wrapper.

In this prototype, the knowledge about MS-Project's analysis capabilities was obtained from the online specification of its API. The knowledge about how the sample data is represented and stored was extracted manually from the MS-Project database catalog. Finally, the knowledge about what data is needed in the analysis and how to translate queries and data between the wrapper and MS-Project was obtained by consulting one of the domain experts who helped develop the conversion mappings. This knowledge extraction and assembly alone took a CS graduate student several days. However, based on our experience, we believe that it is possible to automate the extraction of the necessary schema and operational knowledge from MS-Project and other applications to drastically reduce the manual effort that is needed to configure the wrapper.

Our work to-date suggests that we need future research in three areas: First, further development of knowledge about the underlying representation of supply chain information in subcontractors and suppliers as well as knowledge about necessary transformations. This provides the domain knowledge necessary to automate mediation tasks and drive semi-automated setup of wrappers. Second, development of basic methods of wrapping heterogeneous information systems. Our research work to-date uses formalized domain knowledge to direct automatic wrapping, an advance over existing computer science techniques. Third, development of more advanced query processing capabilities in the mediation layer to effectively direct the pre- and post-processing of data in and from the source. Collectively, our research promises advances in both construction and computer science knowledge, providing tools that address the on-going structural changes in industry towards computational support for networked or extended-enterprise collaboration

Acknowledgements

We would like to thank the National Science Foundation who partially supported this research under grant CMS-0075407.

References

- aecXML. (1999). "A framework for electronic communications for the AEC industries." White Paper, IAI aecXML Domain Committee.
- Antill, J. M., and Woodhead, R. W. (1990). *Critical Path Methods in Construction Practice*, Wiley, New York.
- Barrie, D. S., and Paulson, B. C. (1992). *Professional Construction Management*, McGraw-Hill, New York.
- Birrell, G. S. (1980). "Construction planning — beyond the critical path." *ASCE Journal of the Construction Division*, 106(CO3), 389-407.
- Cattell, R. G. G. (June 1994). "ODMG-93: A Standard for Object-Oriented DBMSs." *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 23(2), 480-80.
- Clayton, M. J., Kunz, J. C., and Fischer, M. A. (1996). "Rapid conceptual design evaluation using a virtual product model." *Engineering Applications of Artificial Intelligence*, 9(4), 439-451.
- Cohen, M. A., and Huchzermeir, A. (1999). "Global supply chain management: a survey of research and applications." *Quantitative Models for Supply Chain Management*, S. Tayur, R. Ganeshan, and M. Magazine, eds., Kluwer Academic Publishers, Boston/Dordrecht/London, 669-702.
- Connolly, D. (1997). "Extensible Markup Language (XML)." W3C.
- Crowley, A. (1999). "The development of data exchange standards: the legacy of CIMsteel.", The CIMsteel Collaborators, 6 pages.
- Davis, T. (1993). "Effective supply chain management." *Sloan Management Review*, 34(4), Summer, 35-46.
- Fondahl, J. W. (1991). "The development of the construction engineer: Past progress and future problems." *ASCE Journal of Construction Engineering and Management*, 117(3), 380-392.
- Genesereth, M., and Duschka, O. (1997). "Answering Recursive Queries Using Views." *ACM Symposium on Principles of Database Systems*, Tucson, AZ, 109-116.
- Genesereth, M. R., and E., F. R. (1992). "Knowledge Interchange Format Technical Report Logic-92-1.", Stanford University.
- Hammer, J. (1999). "The Information Integration Wizard (IWiz) Project." Technical Report TR99-019, University of Florida, Gainesville, FL.

- Hammer, J., Breunig, M., Garcia-Molina, H., Nestorov, S., Vassalos, V., and Yerneni, R. (1997a). "Template-Based Wrappers in the TSIMMIS System." *Twenty-Third ACM SIGMOD International Conference on Management of Data*, Tucson, Arizona, 532.
- Hammer, J., McHugh, J., and Garcia-Molina, H. (1997b). "Semistructured Data: The TSIMMIS Experience." *First East-European Symposium on Advances in Databases and Information Systems (ADBIS '97)*, St. Petersburg, Russia.
- IAI. (1996). "End user guide to Industry Foundation Classes, enabling interoperability in the AEC/FM industry." , International Alliance for Interoperability (IAI).
- McKay, D., Finin, T., and Ohare, A. (1990). "The Intelligent Database Interface: Integrating AI and Database Systems." *National Conference of the American Association for Artificial Intelligence*, Boston, MA.
- O'Brien, W. J. (1995). "Coordination and Change in Construction: the Haas School of Business." Unpublished teaching case , Stanford University, Stanford, CA.
- O'Brien, W. J., and Fischer, M. A. (2000). "Importance of capacity constraints to construction cost and schedule." *ASCE Journal of Construction Engineering and Management*, 125(6), 366-373.
- Papakonstantinou, Y., Garcia-Molina, H., and Widom, J. (1995). "Object Exchange Across Heterogeneous Information Sources." *Eleventh International Conference on Data Engineering*, Taipei, Taiwan, 251-260.
- Tayur, S., Ganeshan, R., and Magazine, M. (ed). (1999). *Quantitative Models for Supply Chain Management*, Kluwer Academic Publishers, Boston/Dordrecht/London.
- W3 Consortium. (1998). "Document Object Model (DOM) Level 1 Specification." W3C Recommendation , W3C.
- Zamanian, M. K., and Pittman, J. H. (1999). "A software industry perspective on AEC information models for distributed collaboration." *Automation in Construction*, 8, 237-248.