# Data Integration for Capital Projects via Community-Specific Conceptual Representations

Yimin Zhu[1], Mei-Ling Shyu[2], Shu-Ching Chen[3]
[1,3]*Florida International University*
[2]*University of Miami*
*E-mail: [1]zhuy@fiu.edu, [2]shyu@miami.edu,[3]chens@cs.fiu.edu*

## Abstract

*Although data integration has been a research subject for decades in the AEC (Architecture Engineering Construction) industry whose data is usually highly fragmented, nowadays we still see problem statements regarding data integration, which are often very similar to those discussed in articles dated long time ago. Representing a departure from existing methods, this paper suggests exploring a novel method, the community-specific conceptual representation, based on the Markov Model framework and the mediator concept. The community-specific conceptual representation provides a de facto standard for collaborating systems in order to share data; while the Markov Model framework and the mediator concept provide a theoretic foundation for the generation of the conceptual representation with expert assistances. This paper offers a theoretic discussion on the relevant theories and concepts. Further studies are needed to validate the feasibility and the practicality of this approach in the context of the AEC industry.*

## 1. Introduction

Large-scale capital programs, often involving many concurrent projects with a huge amount of investment, multiple years' planning, design and construction, and thus great social and economic impacts, bear many features of ad-hoc collaboration. This is mainly due to the fact that they have multiple, volatile yet collaborative organizations, as well as federated heterogeneous supporting information systems.

Such a situation puts great constraints on the collaborating systems. In reality, the users' ability and their need to maximize the value of their information asset is hampered by poor interoperability between systems, competing standards for managing the data, lack of a common methodology for managing a project's information assets, and the complexity of the human-oriented business process. For example, a recent National Institute of Standards and Technology (NIST) report [1] has further confirmed the need for better coordination of electronic data and the development of standards. The report found that better data interoperability could result in potential cost savings of $15.8 billion a year for the construction industry alone.

The capital facilities industry significantly lags in its ability to share and process information. One of the largest problems facing any capital project or capital facility operation is timely access to accurate and complete information. The data integration problem is further complicated by disparate business processes, lack of an accepted industry standard for data exchange, and oftentimes a reluctance to share business-sensitive information.

The need to improve this situation imposes major challenges on the supporting information systems to be 1) *ready* to integrate with heterogeneous data sources; 2) *adaptive* to dynamic interactions among project participants as well as business process changes; and 3) *sensitive* to dynamic needs of information ownership, as well as knowledge protection and sharing.

Existing mainstream integration approaches in AEC are based on shared semantics, bearing significant limitations as those approaches assume a priori knowledge of heterogeneity during system design time, which many studies in computer science have pointed out to be not practical [2, 3, 4, 5]. For example, past experience in information management of capital projects shows that it is difficult for relative static standards to keep up with the complexity and the dynamics of business applications [6, 7], especially when significant human involvement is present such as the case in the construction management process. On the other hand, several studies have also pointed out that mapping between heterogeneous data sources, which seems practical, is not a simple task, e.g.,

due to the existence of semantic heterogeneity and the need for bi-directional mappings [7]. Most of those approaches are semi-automatic, which requires the input from domain experts from time to time [5]. Unfortunately, there is still lack of mechanisms to assist domain experts in managing and controlling the semi-automatic process.

Clearly understanding the standard-based strategy and the mapping-based strategy, this proposed study seeks a new strategy for data integration. Instead of requiring collaborating systems to share an industry standard, the proposed approach turns to the idea of sharing a *community-specific conceptual representation* [8], which is a de facto "standard" only for the collaborating systems of a project, generated based on the *sharable ontological definitions* among heterogeneous data sources.

The organization of this paper is as follows. Section 2 introduces the theory of the MMM mechanism and our proposed Community-Specific Mediator (CSM) conceptual representation. Case studies and several potential applications are discussed in Section 3 to demonstrate the advantages of our proposed method. The conclusion is given in Section 4.

## 2. Theory of MMM

The community-specific conceptual representation, or Community-Specific Mediator (CSM), is based on the MMM (Markov Model Mediator) mechanism, which adopts the *Markov Model* framework and the *mediator* concept. A Markov model is a well-researched mathematical construct, which consists of a number of states connected by transitions; while a previous study [9] defines a mediator to be a program that collects information from one or more sources, processes and combines it, and exports the resulting information. In other words, mediators can be said to be a program or a device which expresses how to integrate different databases/data sources. The MMM mechanism has been experimented in some application scenarios including document management on the World Wide Web [10], multimedia database management [11], and content-based image retrieval [12, 13], but its feasibility and practicality in the capital project management environment remain untested.

The CSM model is suitable to serve as the conceptual representation for the data sources in the management environment of large-scale capital projects with respect to the following aspects:

1. Schema Independence Support: No matter what the underlying schemas of the data sources are, the integrated CSM can still provide querying and browsing of the information as well as control of data capture and presentation without the schema translation processes. The only information required from each data source is its objects and their attributes/features. Since things in the world have defining attributes, each data source can be represented by a set of objects and attributes/features. A local CSM is constructed for each data source and the semantic relationships among the objects within a data source are captured in that local CSM. An integrated CSM is then constructed from the local CSMs. As long as the relationships among the objects in each local CSM are captured, the same object relationships are also retained in an integrated CSM. Hence, as long as each data model of a data source provides the defining objects and their corresponding sets of features (attributes), the semantic relationships can be maintained in an integrated CSM without the mapping between schemas. In other words, schema independence can be supported by using the CSM model.
2. Data Source Autonomy Preservation: One of the important characteristics of the conceptual representation is the ability to preserve the autonomy of each data source; that is, the data in a data source can be created and manipulated independently of other data sources. Data source autonomy can be easily maintained by using the proposed CSM model. No data source needs to expose all of its information to its integrated schema; instead, only the higher level information such as the objects and their corresponding set of attributes/features need to be provided. Data in a data source can be manipulated via its own schema without any data conversion.

There are two types of CSMs. Each data source is modeled as a local CSM, and all the data sources in the program management environment can be modeled as an integrated CSM. The compact notion $\delta=(S, F, A, B,\pi)$ is adopted for the CSM model, where

1. S is a set of objects called states: A CSM consists of a sequence of states which represent the objects (in S) in the data source(s). The states are connected by directed arcs (transitions) which contain probabilistic and other data used to determine which state should be selected next. All transitions $S_i \rightarrow S_j$ such that $\Pr(S_j \mid S_i) > 0$ are said to be allowed, the rest are prohibited.
2. F is a set of attributes/features: A class of attributes or features, associated with an object, is used to characterize the object and to represent

the information pertaining to the data source available to the application queries. Each object has its own set of attributes/features.

3. A is the state transition probability distribution: The state transition probability denotes the probability that a traversal choice to node *j* given the current node is in *i*. Here, the node represents a state in the CSM.
4. B is the observation symbol probability distribution: The observation symbol probability denotes the probability of observing an output symbol from a state. Here, the observed output symbols represent the attributes and the states represent the objects. Since an object has one or more attributes and an attribute can appear in multiple objects, the observation symbol probabilities show the probabilities that an attribute is observed from a set of objects.
5. $\pi$ is the initial state probability distribution: For any object in a data source, the initial state probability is defined as the fraction of the number of occurrences of this object with respect to the total number of occurrences for all the objects in that data source.

## 3. Case Studies and Potential Applications

In a previous study [14, 15], an affinity-based association rule mining (ARM) approach was proposed to discover the quasi-equivalent entities in heterogeneous data sources. The discovered knowledge can then be used to establish the links among the local MMMs to build the integrated MMM. In the following, three examples will be used to illustrate the potential of using the proposed CSM in resolving the data integration problem.

Figure 1 shows a typical problem with three independently defined data models, **Cost, Schedule and Product**. The integration of cost, schedule and product is a classic example of managing capital project data and information. In the following, the authors present three types of cases to discuss how the problems addressed in the case studies can be resolved by using the CSM.

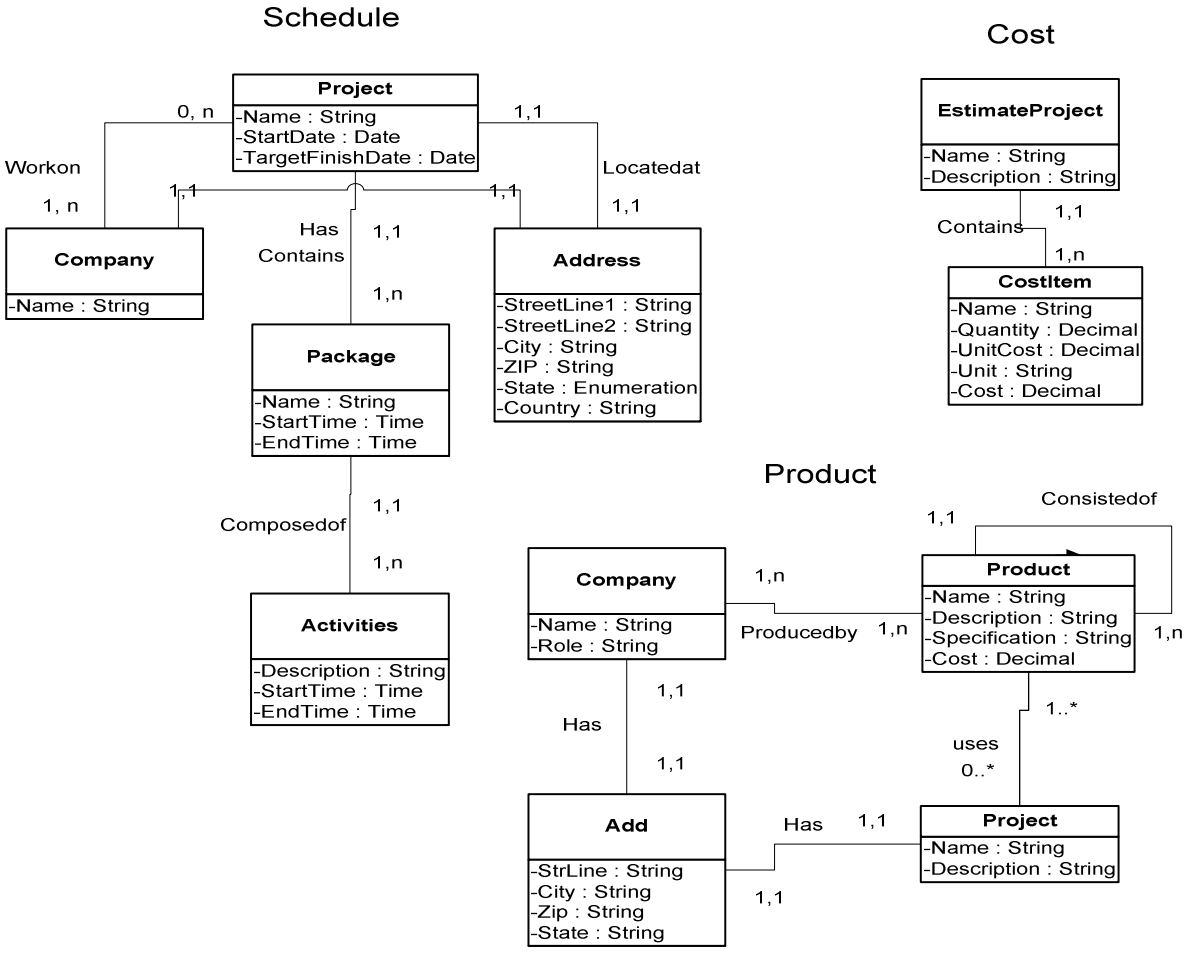


Figure 1. Sample Data Models for Cost, Schedule, and Product

### 3.1 Case 1: Static Structural Conflicts

Assume that the CSM needs to support the user application in such a way that a user application should understand that both addresses ("Address of" **Schedule** and "Add" of **Product** in Figure 1) are supposed to be correct. If the address information needs to be used in generating a report, using either one of the addresses is fine. However, the CSM needs to be able to identify the differences in their attribute sets.

To resolve such a case, the constructed integrated CSM will allow the resolution of structural conflicts. For example, in the integrated CSM, there will be a link between the *Address* entity of the **Schedule** database and the *Add* entity of the **Product** database. Though the attributes in *Address* and in *Add* are not exactly the same, the users can access the complete address information through the links between the two entities. In other words, the retrieval of the data is completely transparent to the users in the sense that these two entities complement each other to provide the complete address information to the users, and the users do not need to know which part of the address comes from which database.

### 3.2 Case 2: Application-Related Heterogeneity

Assume monthly reports are to be generated. The reports need to reflect the schedule status and the budget for a building component. In this case, the schedule information is defined by **Schedule**, budget information is defined by **Cost**, and building component information is defined by **Product**. The CSM needs to be able to represent that these three entities are related so that if information stored in those databases share the same key structure, e.g., using the same Work Breakdown Structure (WBS) for activities, cost items and products, the user application can generate such a report by using the CSM.

To solve this problem, in the integrated CSM, the links of those entities that have high affinity relationships are established based on the usage patterns obtained from the previous monthly reports. To generate the current monthly report, the information from multiple entities in multiple databases can be retrieved for the report. For example, since there are links among the *Project* entity of the **Schedule** database, the *EstimateProject* entity of the **Cost** database, and the *Product* entity of the **Product** database, for a particular building component, its schedule information in the **Schedule** database, its budget information in the **Cost** database, and its product information in the **Product** database can be pulled out from the system to generate the report.

### 3.3 Case 3: Dynamic Semantics

Assume that the *Product* entity of **Product** has another attribute called "budget", which is a lump-sum cost of a product. Such a budget format serves the application of "Product" very well. However, an ad-hoc executive report requires information regarding budget amount at a certain date, such as project encumbrance date. The CSM should not only be able to identify the link between *Activity* entity of **Schedule** and *Product* entity of **Product**, but also manage additional dynamic semantics that handle the additional relationship between them, i.e., the budget of that activity up to the date of project encumbrance.

In construction management, a lot of information such as situation changes, relationship changes, process changes, etc. is very dynamic and it is not feasible to store them in the static structures. However, such information is very important in project operations. In this study, we explore the incorporation of some additional constructs into our proposed framework such as representing the dynamic semantics in Resource Definition Framework (RDF) in a policy database and adding an additional tuple in CSMs.

This policy database will store the dynamic semantics (including situations, relationships, and process changes, etc.) as policies for the system, where all the policies are represented in ontology such as RDF. For example, though the *Budget* entity of the **Product** database is available, it provides only the lump-sum cost of the building component. The detailed break-ups of the budget for the particular product based on the dates of project encumbrance will be kept in this policy database as policies and represented in RDF. We propose to represent CSM as a 6-tuple $\lambda = (S, F, A, B, \Pi, D)$, where D is a semantic-dynamic vector with the size 1 x |S| and |S| is the cardinality of the set S. Initially, all the entries in D are initialized as NULL values. Every time a new policy with respect to a particular entity is added to (or removed from) this policy database, the corresponding entry in D is updated by adding a pointer (or resetting the pointer to NULL) to the policy database. During the retrieval process, each CSM will first check whether the corresponding entry for the entity in D is NULL or not. If it is not NULL, then the pointer will point to the policy database to access the additional information to model the dynamic semantics.

## 4. Conclusion

Having an integrated data source for a capital or facility project that can support seamless accesses to the

data, information and knowledge for optimal decision-making throughout the lifecycle of the project has been a vision of many research studies for decades. However, methodologies for realizing this vision differ. While the mainstream school of thinking in integrating heterogeneous data sources of capital projects still prefers a shared semantic model for integration, this paper envisions a different methodology that possess much potential to the solving the integration problems of capital projects.

The discussions on the three cases show the potential of using the MMM mechanism to generate community-specific conceptual representations in different scenarios.

The proposed CSM community-specific conceptual representation, based on the MMM mechanism, can be more scaleable and adaptive to the dynamics in capital project management. However, further studies are necessary to validate its feasibility and practicality.

## 5. References


[1] M.P. Gallaher, A.C. O'Connor, J.L. Dettbarn, Jr. and L.T. Gilday, "Cost Analysis of Inadequate Interoperability in the U.S. Capital Facilities Industry", *NIST GCR 04-867*, 2004.

[2] C.H. Goh, S. Bressan, S. Madnick, and M. Siegel, "Context Interchange: New Features and Formalism for the Intelligent Integration of Information", *Journal of ACM Transactions on Information System*, Vol. 17, No. 3, 1999, pp. 270-291,

[3] E. Rahm and P.A. Bernstein, "A Survey of Approaches to Automatic Schema Matching", *the VLDB Journal*, Vol. 10, 2001, pp. 334-350.

[4] B.S. Mitchell, S. Mancoridis and M. Traverso, "Reverse Engineering: Search-Based Reserve Engineering", *Proceedings of the 14th International Conference on Software Engineering and Knowledge Engineering*, 2002, pp. 431- 438.

[5] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, "Learning to Map between Ontologies on the Semantic Web", t*he VLDB Journal*, Vol. 12, 2003, pp. 303-319.

[6] M.K. Zamanian, and J.H. Pittman, "A Software Industry Perspective on AEC Information Models for Distributed Collaboration", *Automation in Construction*, Vol. 8, 1999, pp. 237-248.

[7] R., Amor and I. Faraj, "Misconceptions about Integrated Project Database", *ITCON*, Vol. 6, 2001, pp. 57-66.

[8] Y. Zhu and S.-C. Chen, "A Conceptual Framework of Ontology-based Scope Alignment", *the Second LACCEI International Latin American and Caribbean Conference for Engineering and Technology*, Miami, Florida, USA, 2-4 June, 2004.

[9] G. Wiederhold, "Mediators in the Architecture of Future Information Systems", *IEEE Computer*, March 1992, pp. 38-49.

[10] M.-L. Shyu, S.-C Chen and C. Haruechaiyasak, "Mining User Access Behavior on the WWW", *Proc. of IEEE International Conference on Systems, Man, and Cybernetics*, Tucson, Arizona, USA, October 7-10, 2001, pp. 1717-1722.

[11] M.-L Shyu, S.-C. Chen, and R.L. Kashyap, "A Probabilistic-Based Mechanism for Video Database Management Systems", *Proc. of IEEE Intl. Conf. on Multimedia and Expo (ICME'00)*, New York City, USA, 2000, pp. 467-470.

[12] M.-L Shyu, S.-C. Chen, M. Chen, C. Zhang and K. Sarinnapakorn, "Image Database Retrieval Utilizing Affinity Relationships," *Proceedings of the First ACM International Workshop on Multimedia Databases (ACM MMDB'03)*, New Orleans, Louisiana, USA, November 7, 2003, pp. 78-85.

[13] M.-L. Shyu, S.-C. Chen, M. Chen, C. Zhang and C. Shu, "MMM: A Stochastic Mechanism for Image Database Queries," *Proceedings of the IEEE Fifth International Symposium on Multimedia Software Engineering (MSE2003),* Taichung, Taiwan, ROC, December 10-12, 2003, pp. 188-195.

[14] M.-L Shyu, S.-C. Chen and R.L. Kashyap, "Discovering Quasi-Equivalence Relationships From Database Systems", P*roceedings of the ACM Eighth International Conference on Information and Knowledge Management (CIKM'99)*, Kansas City, MO, USA, November 2-6, 1999, pp. 102-108.

[15] M.-L Shyu, S.-C. Chen and R.L. Kashyap, "Generalized Affinity-Based Association Rule Mining for Multimedia Database Queries", *Knowledge and Information Systems (KAIS): An International Journal*, vol. 3, no. 3, August 2001, pp. 319-337.