

Inducting Subcontractor Process Ontologies: Challenges, Methods, and Illustrative Results

William O'Brien*, Raja R. A. Issa**, Rodrigo Castro-Raventós***,
Jaehyun Choi ****, and Joachim Hammer*****

* Assistant Professor, M.E. Rinker, Sr. School of Building Construction and Adjunct Assistant Professor, Dept. of Civil & Coastal Engineering, University of Florida, wjob@ufl.edu

** Professor, M.E. Rinker, Sr. School of Building Construction, University of Florida

*** Graduate Assistant, M.E. Rinker School, Sr. of Building Construction, University of Florida

**** Graduate Assistant, Dept. of Civil & Coastal Engineering, University of Florida

***** Assistant Professor, Dept. of Computer & Information Science & Engineering, University of Florida

Abstract

This paper reports on experience inducting subcontractor process ontologies as part of research on the SEEK: Scalable Extraction of Enterprise Knowledge project at the University of Florida. We briefly review approaches to ontology generation and the application of those approaches to product/process ontologies in the AEC arena as well as process-oriented ontologies generated from a general business perspective. With limited exceptions, existing approaches have not taken an inductive approach. As such, this paper contributes a description of a two-phase inductive methodology to develop firm specific process ontologies. The methodology combines documentation of firm processes and information systems with a synthetically derived base ontology to specify and test a firm specific ontology. Brief findings from application of this methodology are presented. These findings frame description of the challenges eliciting ontological concepts and related data from firms as well as the challenges in representing process knowledge within existing frameworks and tools. Overall, induction is seen as a viable approach to AEC ontology specification, with a useful role in generating detailed ontologies that represent firms' existing processes, procedures, and terminology.

Introduction

Subcontractors are increasingly making use of information systems to manage their operations. This presents an opportunity as information stored electronically can be shared and leveraged to improve coordination among the many firms that compose the construction supply chain. However, such sharing is currently difficult as subcontractors use a wide range of computer hardware and software applications (a.k.a., legacy systems). These legacy systems generally do not represent data in same way, and thus extracting the data for subsequent use generally requires considerable translation. Hence, beyond very basic data, computer-enabled sharing of a subcontractor's information requires formal representation of the firm's processes and procedures. Further, to be reusable, a formal representation of firms' data and processes requires an explicit, formally represented specification or ontology (Gruber 1993).

To-date, very little research has been conducted concerning subcontractor information systems and associated ontologies. Considerable research and development effort has been spent generating a variety of proposed standard data models or standard ontologies to support

interoperability among AEC information systems. Unfortunately, these standard models have had few tests in practice. Further, the standard data models are generally more developed in their representation of products as opposed to processes. Hence there is a need to document process models and information needs for firms. Our aim is to contribute to knowledge about processes through development of detailed process ontologies and methods to generate such ontologies from existing practices. Borrowing from Schlenoff et al. (2000), our basic definition of a *process ontology* is a set of terms capable of describing a process or a flow of processes, including supporting parameters and settings. We focus on process ontologies for construction subcontractors (a thrust of the SEEK project).

Process Ontologies: Review of Efforts and Approaches to Generation

Holsapple and Joshi (2002) define five approaches to the design or specification of an ontology: (1) *Inspiration*, where ontologies are designed from an individual's viewpoint about the domain. (2) *Induction*, which uses case studies as references for ontology design. (3) *Deduction*, where general principles and knowledge about a domain are used to generate ontological formalisms. (4) *Synthesis*, which merges several independent ontologies within a domain to create a unified ontology. (5) *Collaboration*, where several individuals work together to define a shared ontology for a domain. These approaches are not mutually exclusive; indeed, almost every ontology definition effort will include aspects from each approach.

The dominant approaches to ontology specification in the AEC industry are collaboration and synthesis. These efforts are directed at generating shared, common standards. For example, the Industry Foundation Classes (IFC) (IAI 1996) and related aecXML (aecXML 1999) have been generated by committees working in collaboration. Extensions to the IFC specifications have been undertaken by small groups and reintegrated in subsequent, committee approved specifications, following a synthetic approach to ontology definition. In contrast, later efforts such as the OmniClass Construction Classification System (OCCS; OCCS 2001) began with a synthetic approach that combined ontologies from a number of existing efforts. The IFC and related ontologies have elements of processes as part of their specification, but are more developed and used to represent product models. Zamanian and Pittman (1999) suggest that the traditional divide between process and product modeling communities will continue to limit the development and use of IFC and related approaches for process modeling. Other than limited tests (e.g., Froese et al. (1999); Staub-French and Fischer (2000)), the IFC, aecXML, and OCCS have not been widely used for process modeling.

Several ontologies and specifications for business processes have been generated outside the AEC arena. These efforts include ARIS (Scheer 2000), the Process Exchange Format of the Workflow Management Coalition (Workflow Management Coalition 1999), the Enterprise Ontology (Uschold et al. 1998), and the Toronto Virtual Enterprise (TOVE) project (Fox et al. 1996). As with AEC efforts, these ontologies have been generated using collaborative and synthetic approaches although the scale of these efforts is generally somewhat smaller than the construction efforts. The general business ontologies have not yet been widely used or tested outside of academia. The Process Specification Language (PSL; Schlenoff et al. (2000) synthetically combines the business process ontologies referenced above and uses a strong deductive approach to cleanse and evaluate its process ontology. Perhaps most interesting about

PSL is that while very powerful, it is not envisioned as a standard process language, but rather as a lingua franca that can be used as translators between other process ontologies in practice.

Froese (1996) reviewed several general or core approaches to modeling construction processes, commenting that meaningful standards may evolve slowly. The purpose of PSL suggests a subtle modification of this statement, suggesting that a broad standard may emerge for translation purposes but not for uniform acceptance in practice. Following this argument and given the limited adoption of existing process models, it is apparent that our knowledge of processes in practice is limited. There is a role for inducted ontologies to enrich our knowledge of process modeling needs and requirements. Some starting work at inducting process models includes documentation of the types of information used by contractors (Shahid and Froese 1998) and process mapping in the residential construction industry (Wakefield et al. 2001). While useful starting points, this research has not achieved the level of formal specification needed for a robust ontology. In the general business domain, Gandon (2001) outlines the O'CoMMA ontology building project using an inductive methodology. They describe a four-step approach, starting with scenario and data collection, translation from semi-informal data to semi-formal abstractions and definitions, formalization and implementation in an ontology (in their case, using RDFS as a description language), and finally navigation and use of the ontology.

The SEEK Project: Inducting Subcontractor Process Ontologies

The Scalable Extraction of Enterprise Knowledge (SEEK) project (O'Brien et al. 2002) aims to provide a toolkit that enables semi-automatic instantiation of connections between heterogeneous legacy sources and higher level decision support tools such as scheduling and supply chain management applications. SEEK thus enables scalable deployment of decision support tools across a project. As such, part of SEEK research is directed at building a detailed domain ontology that represents the processes of firms participating in the project supply chain. This process ontology is used to guide data reverse engineering (DRE) and schema matching (SM) algorithms; results from these algorithms are used to semi-automatically configure a value-added wrapper. (See Hammer et al. (2002); O'Brien et al. (2002) for technical details.) To-date, we have focused on generation of process ontologies for subcontractors.

Scope of Ontologies Generated for SEEK

The SEEK toolkit employs extensive use of domain specific knowledge in both its DRE and SM processes. The scope of this domain knowledge is determined by the nature of the queries that are generated from the decision support application. Implicitly, the broader the range of knowledge needed for decision support, the more difficult it is to accurately match legacy data with data needed by the tool. Hence, SEEK is not envisioned as a broad data extraction toolkit or as a universal translator for legacy sources. Rather, it is designed to extract a narrow range of information from a variety of sources to support specific types of decision support applications. SEEK thus requires rich definitions of both the knowledge being extracted as well as good definitions of the range in the way that knowledge is represented by firms. Thus there is a need for a correspondingly narrow but specific ontology that relates data in legacy sources with data needed as input (e.g., queries) to the decision support tool(s). The specific class of decision support chosen for exploration in the SEEK project concerns resource availability and

scheduling/rescheduling in the construction supply chain; corresponding focus is placed on developing a resource and scheduling ontology for participating subcontractors.

Ontology Induction Method

Building from the experience of Holsapple and Joshi (2002) and Gandon (2001), we developed a two-phase method for inducing process ontologies. In the first phase, a base ontology was generated using a combined inductive/deductive approach. Concepts regarding resources and scheduling were identified and defined from the general construction literature, allowing deduction of the base process ontology. This base ontology was augmented by a limited case study of a small subcontractor, allowing limited validation and extension. In the second phase, a more detailed ontology was inducted from case studies using the base ontology as a starting point. Where base ontological concepts accurately described the more detailed case study, the base ontology was reused. Otherwise, extensions or revisions were generated.

Early in the project it was decided not to use existing ontologies such as the IFC and PSL. While building from such existing ontologies would allow more specific tests and perhaps richer induced ontologies, it was felt that existing ontologies could bias the data collection and ontology specification efforts. Neither PSL nor the IFC are specific to subcontractors, which might bias understanding of terms and concepts. Further, the rich detail of existing ontologies could cause the researchers to look for certain types of information at the expense of data in use. While safeguards could be built into the data collection and analysis methodology to avoid these difficulties, a naïve approach provided the same benefits and more time to work with practitioners (rather than conducting an extensive review of PSL, the IFC, etc.). Thus the researchers were aware of approaches to process modeling at the level detailed by Froese (1996), and subcontractor resource management issues at the level detailed by Hegazy and Erashin (2001) and O'Brien and Fischer (2000).

Case study descriptions of subcontractors to support ontology induction followed the two-phase approach: An initial case study was conducted to gain experience and knowledge about subcontractors' representation of information. The initial case guided subsequent development of more detailed cases. This approach allowed a separation of efforts on the part of the research team. Graduate student Rodrigo Castro-Raventós took the lead on documenting the subcontractor processes, documents, and information systems (Castro-Raventós 2002) while Jaehyun Choi focused on ontology specification. Both students attended all the meetings with the subcontractors studied.

With regard to generation of ontologies within each phase, three steps were taken. (See figure 1.) First, building from the case study source data, the scope and overall design of the ontology was determined. Ontology class and attribute definitions were also determined from the information on the

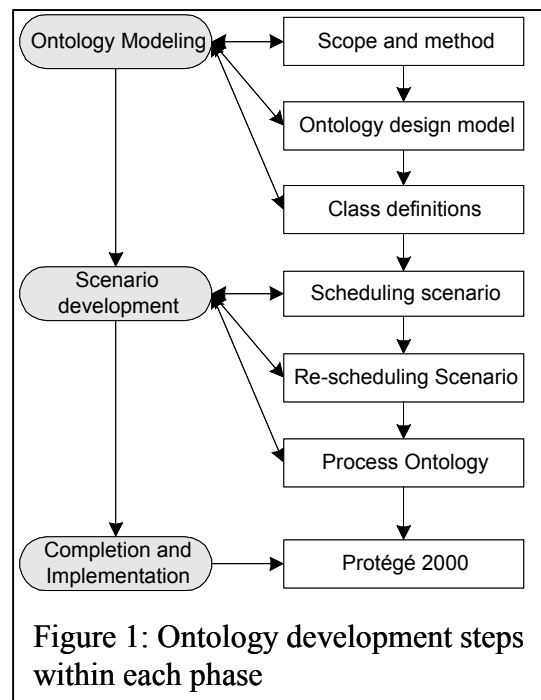


Figure 1: Ontology development steps within each phase

source documents and information records. This provided the basic set of firm specific terms definitions. Concepts and definitions were developed in part by using a matrix to matching document type and concept, closely following the approach used by Shahid and Froese (1998).

The second step within each phase was to review specific process scenarios related to scheduling and rescheduling resources. Employing the basic class definitions developed in the first step, process scenarios were used to define relations between classes, as well as additional classes, sub-classes, and attributes needed for a complete description of the processes. As such, the scenarios provided a test of the internal validity of the definitions generated from the source data. In this sense, the ontology developed cannot be said to be formally cleansed of redundancies or inconsistencies (Guarino and Welty 2002). However, we can be reasonably sure the ontology is capable of representing firm specific processes and related concepts.

The third step is to implement the ontology in a formal language. We employed Protégé 2000 (<http://protege.stanford.edu>; Noy and McGuinness (2001)), a popular tool for representing frame-based ontologies. Protégé has the advantages of simplicity and the ability to export ontologies in XML, increasing capabilities for reuse. However, as a general-purpose tool, Protégé is not specifically designed to support process modeling or provide an executable for testing. Further testing of the process ontologies will be conducted as we place the ontology into the domain models in the SEEK toolkit and explore the ability of the SEEK tools to extract data.

Sample Results: Miller Electric, Inc.

We briefly describe aspects of case research and ontology development for one of the firms studied, Miller Electric, Inc. (Miller). Further documentation can be found in (Castro-Raventós 2002) and in the forthcoming thesis of Choi. Based in Jacksonville, FL, Miller is the 16th largest electrical subcontractor in the United States with annual revenues of \$100 million. Miller is highly data intensive in its management procedures, using a suite of software including J.D. Edwards for accounting data, MS Project and Primavera Project Planner for scheduling, AutoCAD, electrical contractor specific estimating software ConEst, MS Word for correspondence, and MS Excel for cost and quantity calculations and project controls. One of their superintendents even makes all his notes in a PDA. The company information systems are networked and much of the data is accessible to and uploaded from field offices. While much current cost control and planning work is performed using forms that are stored as templates in MS Excel, a corporate effort is underway to better integrate the data currently stored in Excel with the J.D. Edwards accounting software. In addition to software used, the research team was able to identify and map the information on twenty-eight distinct document types and reports used to estimate project costs and manage resources on projects underway.

The rich information that Miller collects, combined with interviews with project managers and the site supervision, have enabled generation of a detailed process ontology specific to Miller. Examination of Miller's source data, processes, and inducted ontology reveal several interesting items: First, Miller's cost codes are firm specific, following a proprietary numbering system where a base number established the main category (e.g., wiring devices, fixtures, etc.) and a standard extension modifies to code for materials, labor, burden, or miscellaneous expense. Second, these cost codes are used to develop schedule activities. For example, a building scheduled by Miller divided work per floor, following the schedule of the construction manager. However, each activity on the floor was named as a cost code (directly linking project activities to costs). Third, Miller does not resource load its schedules. Rather, a

projected manpower curve is generated for each project in MS Excel. Each week, the superintendent compares progress and resource utilization against that curve, with a goal of staying on target. Manpower costs are similarly estimated; superintendents can track their average labor costs and are encouraged to adjust their crew mix to stay below that average.

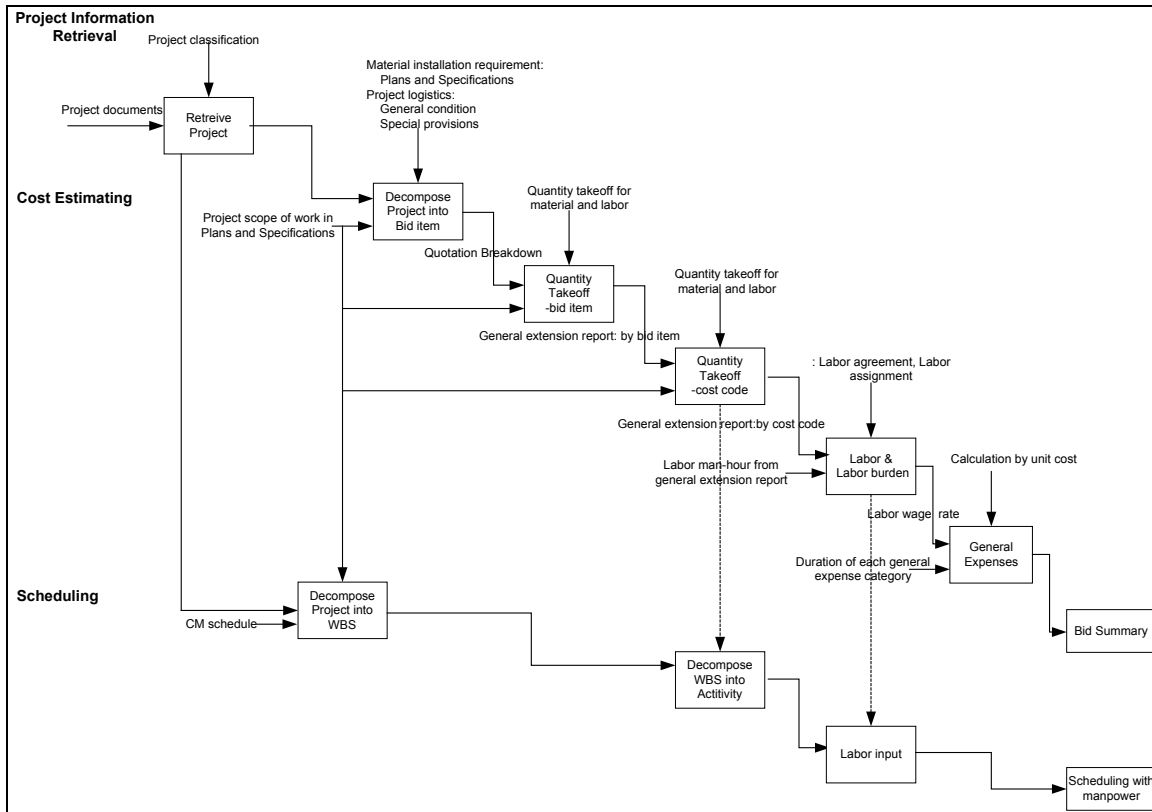


Figure 2: Miller Electric estimating and scheduling process diagram with supporting information inputs (pre-cursor to ontology specification)

Concluding Discussion: Challenges, Learning and Future Research

Details such as those described above are important to the SEEK ontology and extraction algorithms; it is necessary to know how costs and resource are represented by each firm. Abstracting from the details, Miller as an organization is managed from an estimating centric viewpoint. (Part of the integration between estimating and scheduling at Miller can be seen in figure 2, a process diagram generated to as a precursor to formalizing the ontology.) We suspect other firms might be similarly managed; further research is needed to determine what aspects of the Miller ontology can be generalized.

More broadly, our experience suggests that induction is a viable method to generate process ontologies in the AEC domain. It is possible to generate with two students in (large) Master's Thesis projects highly specific descriptions of (aspects of) firms' information and management systems and formalize these descriptions in an ontology. Further, we have observed a large learning curve in the ontology induction process suggesting that professional efforts could have considerably shorter development times. That said, we have also observed the benefit

of having a reflective practitioner as a partner in ontology development. Miller Electric as a firm is extensive in its documentation and has many centralized processes; other firms in this study were found to have much looser documentation and a broader range of decision processes.

Existing tools and definitions may further speed induction. Data mining tools may help identify concepts for inclusion in the ontology; Kosovac et al.'s (2000) development an AEC/FM thesaurus is an example of such work. Similarly, building from existing ontologies such as the OCCS could help generate ontologies better suited to translation and help individuals identify important concepts that may be overlooked in a purely inductive approach. As noted, our research took a naïve approach with only limited reference to existing ontologies. We have not yet compared our findings to the standardized models; it remains an open research issue to determine how well such standards map to current practice and what limitations (if any) may be imposed on induction by starting from existing efforts.

A specific learning from our efforts is that process models do not neatly fall into general ontology definitions and models. In our documentation of process related concepts and translation of those concepts to the SEEK toolkit, we have found it useful to categorize process ontologies as (1) basic definitions of static terms, such as <activity>, (2) business rules that contain well defined methods such as calculating cost from a productivity rate and estimate, and (3) process definitions that contain process steps and relationships between classes. These three steps roughly correspond to increasing levels of abstraction and/or complexity. From our experience, general tools such as Protégé are best equipped to handle definition level ontologies. There is a need for more descriptive representation languages that can address the temporal and state aspects of processes; as we have taken a naïve approach, it is unclear that tools such as PSL provide a better framework. In future research, we will review the ability of PSL to describe our detailed ontologies.

Acknowledgements

The authors thank the National Science Foundation for its support of the SEEK project under grants CMS-0075407 and CMS-0122193. We also thank the firms Centex Rooney, Miller Electric, W.W. Gay Mechanical, and Perry Roofing for their time and support of this research.

References

- aecXML. (1999). "A framework for electronic communications for the AEC industries." IAI aecXML Domain Committee, <http://www.aecxml.org/docs/aecwhite.doc>, 9 pages.
- Castro-Raventós, R. (2002). "Comparative Case Studies of Subcontractor Information Control Systems," M.S. Thesis, University of Florida.
- Fox, M. S., Barbuceanu, M., and Gruninger, M. (1996). "Organization ontology for enterprise modeling: Preliminary concepts for linking structure and behaviour." *Computers in Industry*, 29(1-2), 123-134.
- Froese, T. (1996). "Models of construction process information." *ASCE Journal of Computing in Civil Engineering*, 10(3), 183-193.
- Froese, T., Fischer, M., Grobler, F., Ritzenthaler, J., Yu, K., Sutherland, S., Staub, S., Akinci, B., Akbas, R., Koo, B., Barron, A., and Kunz, J. (1999). "Industry foundation classes for project management - a trial implementation." *Electronic Journal of Information Technology in Construction*, 4, 17-36.

- Gandon, F. (2001). "Engineering an ontology for a multi-agents corporate memory system." *Proceedings of the Eight International Symposium on the Management of Industrial and Corporate Knowledge (ISMICK '01)*, Université de Technologie de Compiègne, France, 22-24 October 2001, <http://www.hds.utc.fr/~barthes/ISMICK01/papers/IS01-gandon.pdf>, 24 pages.
- Gruber, T. R. (1993). "A translation approach to portable ontologies." *Knowledge Acquisition*, 5(2), 199-220.
- Guarino, N., and Welty, C. (2002). "Evaluating ontological decisions with OntoClean." *Communications of the ACM*, 45(2), 61-65.
- Hammer, J., Schmalz, M., O'Brien, W., Shekar, S., and Haldavnekar, N. (2002). "Enterprise knowledge extraction in the SEEK project part I: data reverse engineering." *TR02-008*, Department of Computer and Information Science and Engineering, University of Florida, <ftp://ftp.cise.ufl.edu/cis/tech-reports/tr02/tr02-008.pdf>, 30 pages.
- Hegazy, T., and Erashin, T. (2001). "Simplified spreadsheet solutions. I: Subcontractor information system." *ASCE Journal of Construction Engineering and Management*, 127(6), 461-468.
- Holsapple, C. W., and Joshi, K. D. (2002). "A collaborative approach to ontology design." *Communications of the ACM*, 45(2), 42-47.
- IAI. (1996). "End user guide to Industry Foundation Classes, enabling interoperability in the AEC/FM industry." International Alliance for Interoperability (IAI).
- Kosovac, B., Froese, T., and Vanier, D. (2000). "Use Of Keyphrase Extraction Software For Creation Of An AEC/FM Thesaurus." *ITcon*, 5, 25-36.
- Noy, N. F., and McGuinness, D. L. (2001). "Ontology development 101: A guide to creating your first ontology." *SMI-2001-0880*, Stanford University, http://protege.stanford.edu/publications/ontology_development/ontology101.pdf, 25 pages.
- 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.
- O'Brien, W. J., Issa, R. R., Hammer, J., Schmalz, M., Guenes, J., and Bai, S. (2002). "SEEK: Accomplishing enterprise information integration across heterogeneous sources." *ITcon - Electronic Journal of Information Technology in Construction - Special Edition on Knowledge Management*, 7, 101-124.
- OCCS. (2001). "Overall Construction Classification System: A Strategy for Classifying the Build Environment." OCCS Development Committee, 334 pages.
- Scheer, A.-W. (2000). *Aris : Business Process Modeling.*, Springer Verlag.
- Schlenoff, C., Gruninger, M., Tissot, F., Valois, J., Lubell, J., and Lee, J. (2000). "The Process Specification Language (PSL): Overview and Version 1.0 Specification." *NISTIR 6459*, NIST, <http://www.mel.nist.gov/psl/pubs/PSL1.0/paper.doc>, 83 pages.
- Shahid, S., and Froese, T. (1998). "Project management information control systems." *Canadian Journal of Civil Engineering*, 25(4), 735-754.
- Staub-French, S., and Fischer, M. A. (2000). "Practical and research issues in using Industry Foundation Classes for construction cost estimating." 56, Stanford University, Stanford, CA, 45.
- Uschold, M., King, M., Moralee, S., and Zorgios, Y. (1998). "The Enterprise Ontology." *The Knowledge Engineering Review. Special Issue on Putting Ontologies to Use (eds. Mike Uschold and Austin Tate)*, 13.
- Wakefield, R. R., O'Brien, M. J., and Beliveau, Y. (2001). "Industrializing the Residential Construction Site - Phase II Information Mapping." Department of Housing and Urban Development, Office of Policy Development and Research, Washington, D.C., 80 pages.
- Workflow Management Coalition. (1999). "Process Definition Interchange Model." Workflow Management Coalition.
- 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.