Challenges Associated with Generating Accurate As-is Building Information Models for Existing Buildings by Leveraging Heterogeneous Data Sources

Bo Gu[1], Semiha Ergan[2] and Burcu Akinci[3]

[1] Ph.D. Student, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, email: bgu@andrew.cmu.edu
[2]Assistant R. Professor, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, email: semiha@cmu.edu
[3] Professor, Dept. of Civil & Environmental Engineering, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, email: bakinci@cmu.edu

ABSTRACT

Nowadays, facility management (FM) teams are facing challenges to generate accurate and semantically-rich as-is BIMs for existing buildings. Current model creation approaches, such as model generation based on point cloud data, mainly capture geometric information of a building and lack to provide additional semantic information about components and other project information. This paper provides the results of a detailed case study that aimed at leveraging existing data sources (e.g., archived documents and data in FM systems) to generate accurate and semantically-rich as-is BIMs. The initial findings from the case study highlighted two main challenges associated with model generation from existing data sources: information extraction and integration. Existing information for different components is typically stored in heterogeneous data sources with various formats and quality, and hence requires different approaches to extract information. The findings also showed that almost 40% of the component attributes investigated had conflicting values in existing sources. In order to address these challenges, formalized approaches are required to support conflict resolution, data extraction and integration.

INTRODUCTION

Building Information Modeling (BIM) is regarded as a remedy to manage facility life-cycle information (Eastman et al. 2011; Teicholz 2013). Although BIM is mainly used in design and construction phases in current practice, building owners and facilities management (FM) groups are leading an effort to the emerging concept of BIM for FM, and there is a significant need for an approach to support rapid BIM generation for existing buildings (Tang et al. 2010; McGraw-Hill Construction 2012). Current approaches for developing BIMs for existing buildings include automatic generation of models from paper-based or CAD-based architectural drawings, and leveraging laser scanners and cameras to capture as-is conditions and create models (e.g., Klein et al. 2012; Se et al. 1998). However, these approaches mainly capture geometrical information of a building. Other information that is important for FM activities, such as equipment performance information is not included in such models. These types of semantic information are typically stored in handover documentation at FM departments, which can also be used to generate as-is BIM for FM purpose.

With this vision in mind, this paper provides details of a case study in a 102 year old campus building. The BIM for this building is created by leveraging existing data sources available at the FM department. Two main challenges associated with the process are highlighted, and quantitative and comparative analysis at building component attribute level is conducted to further address these challenges. Finally, this paper concludes with a discussion of the requirements for a more systematic and automatic approach to extract and integrate information from existing data sources.

BACKGROUND RESEARCH

The main challenges defined in this paper are associated with information extraction and conflict resolution. This section provides an overview of prior work on these topics. Retrieval of information from documents is a well-known area of research in the computer science domain. Since the majority of the files stored at FM departments are in CAD and PDF formats, we have investigated approaches that extract data from such documents. In relation to CAD files, several researchers studied extracting geometrical information from 2D architectural plans to create 3D models. These approaches are semi-automatic and take digital architectural floor plans (CAD file) as an input. Using these input, they first recognize entities and spaces, and then use predefined libraries of 3D components (e.g. wall) to create 3D geometries (Se 1998; So et al. 1998). For scanned images, PDF files and photos from digital cameras, optical character recognition (OCR) is a widely used approach to convert written or printed text to computer-readable text (Campos 2009). In addition, text mining is another widely used method to retrieve information from text-based files. The goal is to turn text into data for analysis. It has been applied for many tasks, such as sentiment analysis (tracking people's sentiment by analyzing their tweets), word frequency distributions and pattern recognition (Massey et al. 2013; Li and Wu 2010; Rose et al. 2004). Such approaches can be used to retrieve useful information from existing documents for model generation.

Computerized approaches for conflict resolution in heterogeneous information sources have also been studied by many researchers. Most research studies defined a set of features (metadata) to describe properties of data. For example, Motro (2006) defined five features at data source level (i.e. timestamp, cost, accuracy, availability, clearance) in order to aid users to judge the suitability of each source for the intended use. Pradhan (2007) used three features (i.e. level of details, representation and reference system) to capture characteristics of data sources for construction productivity analysis. The developed data features were then used to handle inconsistency of attribute values extracted from different sources (Phan 2004; Veregin 1999; Motro 2006). In this study, five data features were defined based on the previous research, to describe both data sources and attribute values, which will be further used for conflict resolutions.

CASE STUDY

The objective of this case is to understand the challenges associated with generating BIMs from documentation accumulated throughout the life of existing facilities. The case involved the mezzanine floor of a 102 year old academic building. From FM department that operates and maintains this building, we were able to find totally 859 document sets related to the construction and FM activities that occurred in this building from 1911 to 2013. The documents are either hard copies or in electronic formats (e.g., PDF and CAD). The authors first used a commercially available modeling tool to create the BIM based on the information stored in these documents and encountered following issues. Manually extracting data from existing data sources was time-consuming. For example, it took almost 30 hours to create a model for one floor with 15477.12 square feet. In addition, in several cases, for a given attribute (e.g., height of a wall), there were multiple conflicting values, which means different values of attributes from different data sources.

In order to further address these issues, we conducted a comparative analysis at attribute level, which compares the values extracted for each specific attribute. Three specific components were selected, a wall, a space and a fan coil unit. These components represented four main types of building components, architectural (i.e. wall), structural (i.e. wall), spatial (i.e. room) and mechanical (i.e. fan coil unit). The attributes investigated in this study were the attributes defined in Industry Foundation Classes (IFC) and a modeling tool for these components. Applicable attributes were listed, and corresponding values were manually extracted from different sources for each attribute. The resulting comparative analysis matrix included total of 146 attributes analyzed for the selected components. The challenges encountered during this process were documented.

CHALLENGES ASSOCIATED WITH INFORMATION EXTRACTION AND INTEGRATION

Based on the comparative analysis, two main challenges were observed while generating the BIM of this existing building. Detailed findings are discussed below:

Challenge 1: Existing information for different components is stored in heterogeneous data sources with various formats, requiring different approaches to extract values.

Various file formats were encountered, scanned images from old hand-drawn drawings (.tif), scanned PDFs of recent drawings, hard copies of documents (.pdf), PDFs of printed drawings and documents (.pdf), CAD files (.dwg, .dwf), documents and spreadsheets representing reports, submittals, checklists (.doc, .xls). Figure 1 gives an example of difference among various formats. These three drawings show information for the same room, which are respectively in scanned image (.tif), PDF and CAD (.dwg) formats. The first drawing was drawn in 1915 by hand, and then scanned into an image file. Due to its age and drawing method, lines and texts on it are blurred, which makes data extraction challenging. The second drawing may be exported from a CAD file. It has fine lines and texts which are easy to recognize.

However, it does not have any measurement notation and drawing scale. The third drawing is a CAD file. Compared to other two formats, it has the most amounts of details. This is a typical example we observed that occurs frequently; various formats have different levels of readability, details for information and thus difficulty for extracting data from. As formats, information availability (i.e., directly reading from the source) and/or whether the value of the attribute can be obtained through measurements, calculation or deduction, the applicable data extraction methods will change.

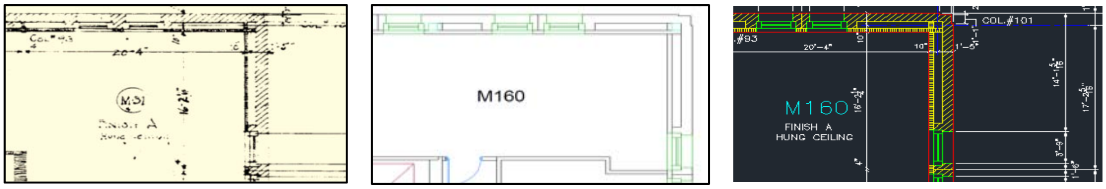

Figure 1. Difference among Various Formats: Scanned PDF, PDF and CAD

Table 1 provides percentages of available data in each format type of the available data sources for assigning values to 146 attributes associated with the three selected components. The sum of the percentage from different formats exceeds 100%, because there are redundant and contradictory attribute values (details discussed in *Challenge 2*) among various data sources. Though all of these data formats provide a certain percentage of the required information to model these components, not a single source contained all the information. Therefore, all of them are needed in order to generate a complete BIM.

Table 1. Percentage of Information Available in Various Data Formats for Modeling the Selected Components

Format	Corresponding Data Sources	Percentage of Available Data	Example Information Represented in Data Source
CAD	Ceiling Plan from 2011 Floor Plan from 2004	54.79%	Wall: dimensions, location Space: finish schedule
PDF Document	Submittals for FCU	43.84%	FCU: performance
Scanned Image	Drawing Set from 1915	31.51%	Wall: elevation, material, opening Space: elevation
PDF Drawing	Floor Plan from 2012 Drawing set from 2004	17.81%	Space: finish Schedule, boundary
Spreadsheet	Room Specification	9.59%	Space: occupancy
Scanned PDF	Drawing from 2012 Pre-commissioning Report	5.48%	FCU: served space, installation

Challenge 2: There are different types of conflicts among relevant attribute values extracted from different sources, but no formalized approach is available for conflict resolution and data fusion.

By building on the concepts developed for conflict resolution within the computer science domain, and synthesizing the observations experienced during model generation step, we have identified five data features (Table 2). These data features describe the characteristics of data sources, to enable systematic comparison of these data sources and attribute values, and used for conflict resolution. Features of the data sources (recentness, format and project phase) represent the general characteristics of a data source that can be inherited by all the attribute values extracted from this data source. Within a given data source, data values for various attributes may have different characteristics. Hence, two features have been defined for the attributes, namely availability and extraction methods.

Table 2. Data Features

	Feature	Description	Example Value
Data Source Level	Recentness	indicates when a data source was created	Time stamp
	Format	indicates the document/database file format, which impacts data quality and the methods for data extraction	Scanned Image, Scanned PDF, PDF, CAD, Doc
	Project phase	indicates the lifecycle phase a building during which the data source was created	As-built, As-designed
Attribute Level	Availability	indicates whether a value is available in the data source for a given attribute	Y: yes, N: no
	Extraction Methods	indicates how the data value is extracted, whether it can be directly read from the source, it needs measurements, calculations or can ben deduced from other attribute values	Directly read, Measured, Calculated, Deduced

These features have been used while existing documentation was being analyzed and their values were populated from existing data sources. After data extraction, we compared attribute values from heterogeneous sources and observed conflicts at the attribute level. A conflict is present in an attribute if it has multiple values extracted from different sources. If there is only one single value for the attribute, it is defined as conflict-free. We identified three types of conflicts, i.e. redundancy, complementary and contradiction. This section presents examples of these conflicts and provides the details of the conflict types identified.

Redundancy. If there are multiple attribute values extracted from different data sources, and the values are equivalent, those values are defined as redundant. As the example shows in Table 3, three drawings are all able to provide data for the width of a particular wall, thus we have a set of values {4, 4, 4}.

Complementary. Some attribute values cannot directly be extracted from data sources, but they can be derived from other attributes or combination of others. For example (see Table 4), we can extract wall height (13 feet) from Data Source 2, and wall footprint area from Data Source 1. There is no data source that directly provides value for wall volume. However, by multiplying both two values, we can calculate the value for the volume of the wall. Calculation is a main derivation method. Another one is deduction. For example, there are two attributes, Space Use and Is Occupiable, which are used to describe the function of a space. Once we know the space is used as an office, we can infer that it's occupiable.

Table 3. Example for Redundancy of Values

Wall Width		**Features**				
		Availability	Recentness	Format	Project phase	Extraction Method
Source 1	4 in	Y	2011	CAD	Null	Measured
Source 2	4 in	Y	2004	CAD	As-designed	Measured
Source 3	4 in	Y	1915	Scanned Image	As-built	Deduced

Table 4. Example for Complementary Values among Different Data Sources

	Source 1	**Source 2**
Wall Height	Null	13 ft
Wall Footprint Area	4.84 ft^2	Null
Wall Volume	Null	Null

Contradiction If an attribute has a set of values contains inconsistencies, it is defined as contradicting. Table 5 shows a multi-source contradiction example. For the room tag of a specific room, there is a set of values {M162, M162, M47}, which contains two different values. The difference is mainly because of renovations over time. Those values have different data features, which would be important to decide which value to use. A contradiction can also happen within a single data source. For example, from the same floor plan (a CAD file), we were able to extract two different values {115, 128.1} for the same attribute -- room area, by using different extraction methods. The first value is labeled on the drawing which can be directly read, while the second one is calculated based on the measurement of room width and length. The difference may be because of file flaw or human measurement error.

Table 5. Example for Contradiction among Different Data Sources

	Room Tag	**Features**				
		Availability	Recentness	Format	Project phase	Extraction Method
Source 1	M162	Y	2011	CAD	Null	Directly read
Source 2	M162	Y	2004	CAD	As-designed	Directly read
Source 3	M47	Y	1915	Scanned Image	As-built	Directly read

Given the 146 attributes associated with the selected components, we were able to find available values from existing data sources, and calculate the percentages for conflict-free and conflicting attributes (as shown in Table 6). The fan coil unit (FCU) has the most available information than other two components, and there is no contradictory attribute value. This is mainly because various documents in relation to FCU (e.g. technical submittal, commission report and drawing) have different scopes, which provide few overlapping information. Though the set of information for FCU is mainly conflict free, having only one source to indicate the value of an attribute might not always result in an accurate or up-to-date value.

Almost 40% of all the attributes are conflicting and decisions need to be made for which value to use. Among all the information available for wall attributes, half of them were conflict free and the other half in conflict. For space information however, the majority of the available information was in conflict. In this case, only three components were investigated and almost 60 conflicting attributes were found. A complete BIM contains thousands of components, which can result in a significant amount of conflicts. It will be time-consuming to resolve conflicts manually. Therefore, a formalized approach to handle these conflicts is necessary.

Table 6. Percentages for Conflict-free and Conflicting Attributes (%)

	Conflict -free	Conflict				Total
		Redundancy	Complementary		Contradiction	
			Calculation	Deduction		
Space	4.1	9.6	4.5	1.8	7.8	26.0
Wall	12.3	5.5	1.4	0	8.2	27.4
FCU	45.2	0	0	1.4	0	46.6
Total	61.6	15.1	4.1	3.2	16.0	

In summary, we observed that existing information is stored in different data sources with various formats and many types of information conflicts. Manually going through these sources to extract information and handling conflicts are time-consuming and may result in inaccuracies. Currently, we are developing a prototype to formalize a framework that is able to take existing data sources as inputs, automatically extract data for required attributes, resolve conflicts and populate attribute values into BIM. The framework will be able to address the challenges stated above by automating the procedures of selecting extraction methods based on formats of data sources and conflict resolution strategies based on data features.

CONCLUSION

Leveraging existing data sources is a promising way to generate accurate and complete as-is BIMs for existing buildings to support FM activities. This paper highlights two main challenges: 1) the explored existing information is stored in heterogeneous data sources with various formats, which requires different extraction methods 2) there are different types of conflicts among relevant attribute values extracted from various sources, but no formalized approach for conflict resolution. In the case study, attributes for three selected components (i.e. wall, space and fan coil unit) were listed, and values extracted from existing data sources for each attribute were compared. Existing data sources were explored and grouped into five formats, namely Scanned Image, Scanned PDF, PDF, CAD and Doc. We defined two-level data features to describe characteristics of data sources and data values, including Recentness, Format, Project Phase, Availability and Extract Methods. Through comparative analysis at attribute level, three types of conflicts have been identified: redundancy, complementary and contradictory. These challenges give motivation for further research towards development a more systematic and automatic approach for information extraction and integration.

REFERENCES

Bleiholder, D.I.J. 2010. *Data Fusion and Conflict Resolution in Integrated Information Systems*. (Doctoral dissertation), University of Potsdam. http://www.hpi.uni-potsdam.de/fileadmin/hpi/Forschung/Publikationen/Dissertationen/Diss_Bleiholder.pdf

Campos, T de, BR Babu, and M Varma. 2009. “Character Recognition in Natural Images.”. *In VISAPP,* Lisboa, Portugal, Feb 05-08.

Eastman, Chuck, Paul Teicholz, Rafael Sacks, and Kathleen Liston. 2011. *BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers and Contractors*. Wiley.

Kiziltas, Semiha, Anu Pradhan, and Burcu Akinci. 2007. “Fusing Data from Multiple Sources to Support Project Management Tasks.” *Proceed of Comp. in Civil Eng.* 411–418. Pittsburgh, USA, July 24-27.

Klein, Laura, Nan Li, and Burcin Becerik-Gerber. 2012. “Imaged-Based Verification of as-Built Documentation of Operational Buildings.” *Aut. in Constr.* vol. 21,161–171.

Li, Nan, and Desheng Dash Wu. 2010. “Using Text Mining and Sentiment Analysis for Online Forums Hotspot Detection and Forecast.” *Decision Support Systems,* 48 (2) 354–368.

Massey, Aaron K, Jacob Eisenstein, Annie I Ant, and Peter P Swire. 2013. “Automated Text Mining for Requirements Analysis of Policy Documents”: In *Requirements Eng.Conference (RE), 2013 21st IEEE International*, pp. 4–13.

McGraw-Hill Construction. 2012. The Business Value of BIM in North America Multi-Year Trend Analysis and User Rating (2007-2012).

Motro, Amihai, and Philipp Anokhin. 2006. “Fusionplex: Resolution of Data Inconsistencies in the Integration of Heterogeneous Information Sources.” *Information Fusion* 7 (2): 176–196.

Phan, V. 2004. “Data Quality Based Fusion: Application to Land Cover.” *In: 7th International Conference on Information Fusion, FUSION 2004,* Stockholm, Sweden, July.

Rose, S., Engel, D., Cramer, N., & Cowley, W. 2010. Automatic keyword extraction from individual documents. *Text Mining,* 1-20.

Se, Carlo. 1998. “Generation of 3D Building Models from 2D Architectural Plans” *Computer-Aided Design*, 30(10), 765-779.

So, Clifford, George Bach, and Hanqiu Sunt. 1998. “Reconstruct Ion of 3D Virtual Buildings from 2D Architectural”: *In Proceedings of the ACM symposium on Virtual reality software and technology*, pp. 17-23. ACM.

Tang, Pingbo, Daniel Huber, Burcu Akinci, Robert Lipman, and Alan Lytle. 2010. “Automatic Reconstruction of as-Built Building Information Models from Laser-

Scanned Point Clouds: A Review of Related Techniques." *Automation in Construction* 19 (7) (November): 829–843.

Teicholz, Paul. 2013. *BIM for Facility Managers*. 1st ed. Wiley.com.

Veregin, H. 1999. "Data Quality Parameters." *Geographical Information Systems*: 177–190.