Evaluation of IFC4 for the GIS and Green Building Domains

Jack C.P. Cheng[1], Yichuan Deng[1], Moumita Das[1]and Chimay Anumba[2]

[1]Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong; PH (852) 23588186; email: cejcheng@ust.hk, ycdeng@ust.hk
[2]Department of Architectural Engineering, The Pennsylvania State University, 104 Engineering Unit A, University Park, PA 16802; PH 814-865-6394; email: anumba@engr.psu.edu

ABSTRACT

In the BIM domain, Industry Foundation Classes (IFC) is the major data exchange standard. In March 2013, IFC4 was released as the new IFC standard. IFC4 has many changes in scope, entities, and data representations. In particular, IFC4 has been enriched to support geographical information and green building design. In IFC4, a building design can be linked to a GIS coordinate system and vice versa. There are also improvements in IFC4 for energy analysis, building service systems, and lighting and shading components. In this paper, changes in IFC4 in the aspects of GIS and green buildings were identified and evaluated using the versioning technique and domain knowledge. CityGML and gbXML were selected as the representative data schemas in the GIS domain and the green building domain, respectively. The version differences were compared with the domain knowledge using linguistic-based method to discover the mapping between IFC4 and CityGML/gbXML. The missing elements in IFC4 for the two domains were also identified for further investigation.

INTRODUCTION

The Architecture, Engineering and Construction (AEC) industry is fragmented due to the participation of various stakeholders. Data exchange between stakeholders is thus a critical problem as they may use different software and data standard to create and represent building models. The integration among different data standards calls for a common public data exchange standard (Wang et al. 2007). Industry Foundation Classes (IFC) is one of the major efforts to facilitate interoperability among stakeholders. Developed by buildingSMART (formally known as IAI), IFC aims to cover the applications of the whole building lifecycle, such as architectural and structural design, construction 4D simulations for constructability analysis and facility management.

The schema of IFC expands as it updates to new releases, which could be shown in the scope of domains covered and number of entities defined in the schema. In IFC release 2.0, it covers six major domains which are architecture, facility management, building service, codes and standards, simulation, and scheduling. In IFC 2x3, which is nowadays the major release of IFC, it covers

nine domains, adding HVAC, electricity, construction management domains. The number of entities in IFC schema also keeps increasing. As shown in **Table 1**, the number of entities has increased dramatically since the beginning of IFC. In the newly released IFC 4, the number of entities is 766, which is 113 more than the previous version.

Table 1 Number of Entities in IFC History

Release	Year	No. of Entities	No. of Types
IFC 1.5.1	1998	186	95
IFC 2.0	1999	290	157
IFC 2x	2000	370	229
IFC 2x2	2003	623	311
IFC 2x3	2006	653	327
IFC 4	2013	766	391

As introduced by buildingSMART, the IFC 4 release contains major changes in scope of coverage. IFC 4 incorporates more features such as modeling of structural steel and timber, site planning and 5D modeling than IFC 2x3. IFC 4 also highlights the newly added entities that are related to green building design (energy analysis, environmental impact) and GIS (coordinate system transformation). The data integration of IFC with these domains has always encountered problems due to the lack of corresponding entities, which forced researchers to develop domain extensions for old versions of IFC (Cheng and Das 2011; Cheng et al. 2013). These newly added entities could benefit the interoperability of IFC with other domain schemas, such as Green Building XML (gbXML) in the green building domain and City Geography Markup Language (CityGML) in the GIS domain. As IFC keeps expanding, it is anticipated that it will add more entities to represent data from different domains in order to truly serve as a common data schema in the AEC industry (Amor and Ge 2002). However, the expansion of IFC also brings a critical research question about the evaluation of domain-specific changes in IFC. The following challenges must be tackled to answer this question: first, to identify domain related classes in IFC, such as *IfcMapConversion* which is related to GIS; second, to evaluate these domain-specific entities in order to know whether IFC covers all the required information for this domain; and finally, based on these findings, to develop mappings between IFC and domain schemas in order to enable interoperability between IFC and certain domain.

In this paper, we introduce a methodology framework to evaluate the schema of IFC4 against the domain-specific changes. The green building domain and GIS domain are chosen in this research as demonstration of the methodology framework. The designed framework serves for three purposes: to generate domain-specific terminology for evaluation of IFC 4; to find domain related classes using these domain terminologies inside IFC 4 and IFC 2x3; and to evaluate the domain-specific changes in IFC 4 as compared to IFC 2x3 using versioning techniques.

RELATED WORKS

Since the differences between each version of IFC is large, the data conversion between IFC versions has become a complex problem. For example,

in IFC 4, new enhancements of geometry resources which allows arbitrary sweeps and non-planar surface bounds were added, which create difficulties for down-grading of IFC 4 to previous versions. The mapping between versions of IFC has long been discussed by IFC developers and researchers. Amor and Ge (2002) analyzed the mapping categories between schema version of IFC and showed statistics of the correspondences between different versions of IFC models. Based on this, they developed a manual assisted version mapping framework for IFC. Wang et al. (2007) further detailed the categories of version changes of IFC, adding merged, split and attribute changes. They also investigated the structure changes of entities. Based on this work, they developed the automatic version matching approach called VMA. Wang et al. (2009) discussed the potential application of version differences in schema mapping. They argued that given a mapping between schema *A* and an older version of *B*, we could generate version mapping between new *B'* and *B*, then update the mapping of *A* and *B'* automatically using version differences. They validated this approach using IFC 2.0 and 2X against Building Code data model. However, these research efforts mainly focused on locating the version changes of IFC and generating statistics of each change categories. They did not consider domain knowledge (for example the domain knowledge from BC data model) and did not provide improvement advice for further development of IFC for specific domains.

RESEARCH METHODOLOGY

In this paper, we try to evaluate the changes in IFC 4 as compared to IFC 2x3 for some specific domains, namely the green building domain and the GIS domain. The domain-specific terminologies were first constructed from domain schemas using text-mining techniques. The domain related classes in IFC 4 and IFC 2x3 were located using these domain-specific terminologies. Then a program we developed to detect the version differences between IFC schemas scanned through these filtered entities and detected version changes for them (i.e. added, deleted, merged). The methodology framework is illustrated in **Figure 1**.

The Generation of Domain-Specific Terminology. The generation of domain-specific terminology is an important research area in natural language processing. According to Ahmad et al. (1994), domain terminologies are labels in the special language of a domain which designate a particular concept in the knowledge of that domain. The domain-specific terminologies will be used for filtering the entities in IFC schema and will have a great impact on the accuracy of the final results. We adopted a corpus-based method using schema of gbXML for the green building domain and CityGML for the GIS domain.

Corpuses were for generation of domain terminology in a number of studies (Drouin 2004). The data schemas of specific domain were chosen as corpus because of their well-defined structure and clear relationships between terminologies. For example, in CityGML, the concept of *Building* is the children of the abstract type *Site*, which is linked to *CityObjects*. By traveling inside the schema structure, we could get related terms "City" "Site" and "Building". In our research, rather than generating multiword terms, such as "Coordinate Reference System", we generated uni-terms such as "coordinate" "reference" and "system". This is because when filtering using these domain terminologies, IFC and domain

schema may use different multiword terms to represent the same concept (e.g. *WallSurface* and *IfcWallStandardCase*). Besides considering the entity names inside the corpus, the parent/super type name and attribute name/type were also considered to expand the possibility of extracting the domain terminologies.

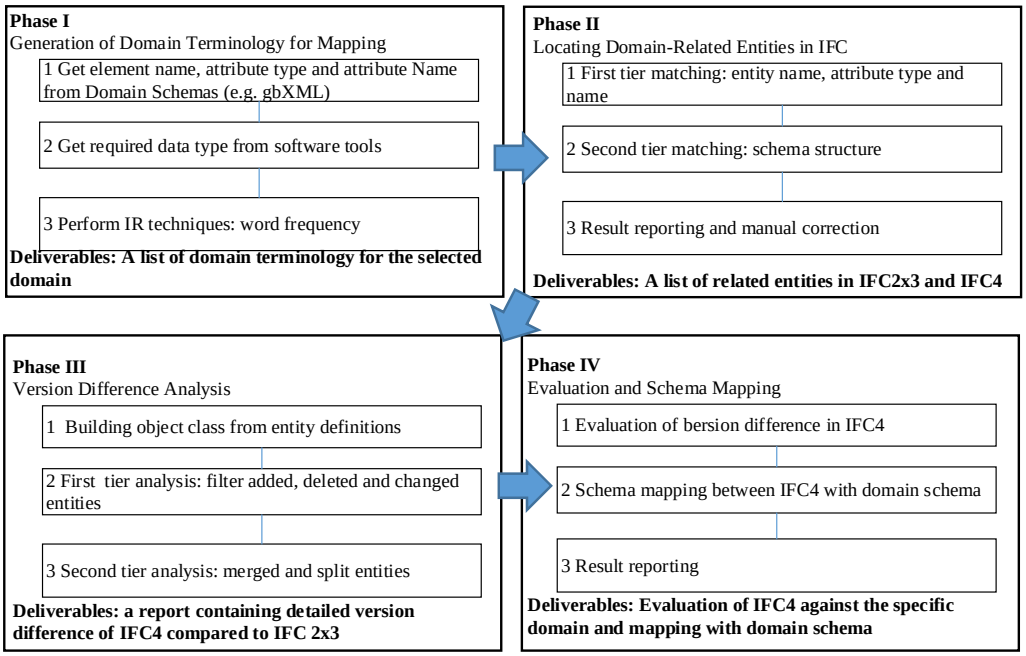

Figure 1 The Methodology Framework

One of the drawbacks of generating the uni-terms is over-generation, which could be 50%-80% of the terminology candidate population (Velardi et al. 2001). For example, when getting the term "City" from "*CityObjects*", the word "Object" is also reported which is irrelevant and invalid. The true terminology could be filtered using term frequency (TF) and inverse document frequency (IDF) combined (Velardi et al. 2001). The IDF is calculated as follows

$$\mathrm{IDF} = \log\left(|D| / f_{w,D}\right)$$

where $|D|$ is the size of the document pool and $f_{w,D}$ is the number of documents in which *w* appears.

A list of stop words, which are commonly seen in the corpus, could be generated from term frequencies. As shown in **Table 2**, some of the words with large frequency and small IDF were considered stop words. After generation of the terminology candidates, a domain expert can inspect the results and perform a manual filtering of the candidates. The generation of domain terminologies also considered the data requirements for software in the domain.

The Filtering of Entities. After generation of domain terminologies, we used these terms to filter the domain related classes in IFC schema. The evaluation consists of three components: the entity name, the entity super type and the attribute type/name. The super type name was also taken into account because if a super type is domain-related, the possibility of a children is also domain-related is

higher. The attribute type and name were also considered as it unveils some of the domain-related entities that do not have a domain related name or super type. For example, the IFC 4 entity "*IfcCoordinateOperation*" is related to GIS, but as "coordinate" and "operation" are common stop words, they are not included in the domain terminology, and it has no super type. The relationship could only be discovered by its attribute name "*Source CRS*", where CRS is a domain terminology to represent the world coordinate system. The scores to evaluate the relatedness of an entity to a certain domain is calculated as follows

$$R_E = \frac{\sum_{1}^{N_{name}} S_{E.namesplit.i}}{N_{name}} + \frac{\sum_{1}^{N_{Es.name}} S_{Es.namesplit.i}}{N_{Es.name}} + \frac{\sum_{1}^{N_{att}} S_{\text{att}.namesplit.i}}{N_{att}}$$

Where N_{name}, $N_{Es.name}$, N_{att} denote the number of hits in name, super type and attribute type of entity *E*, respectively. *S* denotes the score of one hit, which is set to be 10 in our research. The maximum score of R_E is thus 30.

After getting all the scores for each entity, the results were filtered by a domain expert to make sure all the domain related classes are included and all the unrelated candidate eliminated. Our tests show that this approach generated a precision of up to 78%, of which the details will be described in the next section.

Table 2 Some common words generated from IFC that are removed

Term	Frequency	IDF
Define	595	0.4515
Type	535	0.4977
Element	387	0.6383
Value	319	0.7222

The Version Difference Analysis. What distinguished our version difference detection approach from previous IFC versioning approaches is that we only tested the domain-related entities. As shown in **Figure 1**, we first built a dictionary of data model using JSDAI for both IFC 2x3 and IFC 4. This dictionary contains entity names, hierarchy structure of schema, and entity attributes. Then each filtered domain related class was tested against entities in older/newer version. The test contained entity level comparison to detect addition or deletion, attribute level comparison to detect modifications and structural level comparison to detect super type changes. We adopted and modified the version change category reported in Wang et al. (2007), which are shown in Figure 2.

RESULTS AND DISCUSSIONS

Validation of Domain Specific Terminology Generation. In the generation process of domain terminologies, we used TF IDF as the automatic filtering measure. We neglected the first 100 words with higher frequency reported in IFC documentations and also neglected the words with IDF<1 (i.e. appear in more than 10% of IFC documentation). Before the manual filtering, the corpus from the green building domain and the GIS domain automatically generated 634 words and 271 words, respectively. After manual filtering, the numbers of valid domain terminologies for these two domains were 481 and 128, which means an accuracy

of 75.7% and 47.4%, respectively. This result also complies with the observations made in Velardi et al. (2001).

The low accuracy of the terminology generation for corpus in the GIS domain is due to the schema definitions of CityGML. In CityGML, many entities are used for representing the basic geometry features. Words such as “linear”, “geometry”, and “Cartesian” are observed in the generated candidate list. Moreover, words that truly represent domain specific knowledge of GIS such as “map” and “CRS” are not as much as that of green building design. The result of the GIS domain shows that we have to carefully select the corpus that is used for generating the domain-specific terminologies.

Validation of Filtering Domain Related IFC Classes in IFC Schema. The results of filtering domain related classes in IFC 2x3 and IFC 4 are shown in Table 3. The total precision for filtering entities for the green building domain and the GIS domain is 72.55% and 78.20%, respectively. The filtered entities were validated manually based on the following principles: first, keep entities that are representing real world concepts and are domain related, such as *IfcPump* or *IfcMapConversion*; second, keep enumeration and selected types that are related to specific domains, such as *IfcEnergyConversionDeviceType* and *IfcHeatExchangerTypeEnum*; and third, neglect entities that are representing values of objects, which are common for every domain, such as *IfcLoop*. The high precision of the result shows the effect of the filter design.

Table 3 Results of filtering domain-related entities

	IFC 4 Located	IFC 4 Validated	Precision	IFC 2x3 Located	IFC2x3 Validated	Precision	Total Precision
Green Building	484	329	67.98%	314	250	79.62%	72.55%
GIS	227	180	79.30%	163	125	76.69%	78.20%

The scores of the filtering has also been reasoned to show the correctness of the filter design. It is observed that with higher R_E, the entity is more likely to be a domain-related entity. With higher scores, the precision of the filter results tend to be higher.

Results of Version Differences of Domain-Related IFC Classes. A detailed report of the version differences of domain-related IFC classes identified in the previous phase was generated using the version difference comparison program we developed. This report contained all the detailed changes of IFC classes, including: (1) the entity level changes, such as identical, addition and deletion, (2) changes of schema structure, such as changes in super types, and (3) changes inside entities, such as changes in attributes, inverse attributes, where rules and content changes in enumerations and select types. The statistics of each categories in version changes for domain related entities were shown in Figure 2.

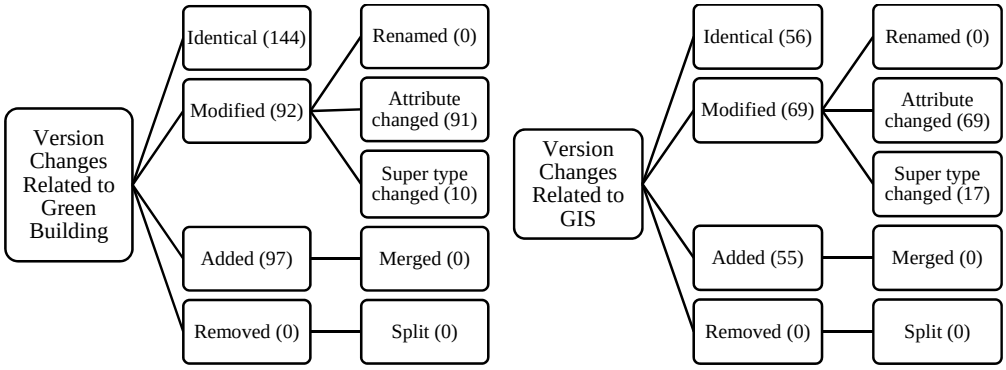

Figure 2 Categories of version changes and statistic of GB and GIS

Figure 2 shows that besides adding a large number of entities related to green building (97 added) and GIS domains (55 added), buildingSMART did not remove IFC classes that are related to these two domains in IFC 4. This proves the statement about version changes of IFC 4, which shows IFC's consideration of both the green building domain and GIS domain.

The Missing Concepts in IFC 4 Regarding Green Building and GIS. Since a domain terminology list was generated for both the green building domain and GIS domain, it is possible to identify not only what are contained in IFC 4, but also the missing items to represent the domain concepts in the schema. With reference to the domain terminology list, we identified a number of concepts that were not hit by any entity definitions in IFC 4 for both domains. For example, for all the 128 domain terms in GIS, there are 69 terms that were not hit by any definitions of IFC classes. It shows that the newer version of IFC cannot fulfill the data modeling requirement for green building design and GIS as compared to the domain schemas. Analyzing these zero-hit terms could provide a clue for further development of IFC towards complete modeling of these two domains. Table 4 shows some of the zero-hit terms, which mainly refer to weather and indoor environment for green buildings and to road, railway and tunnels for the GIS domain for further consideration of IFC development.

Table 4 Examples of missing terminologies in IFC 4 for green building and GIS

GB	Weather	Humidity	Carbon	Ventilation	Fan	R-value
GIS	Road	Railway	Bridge	Traffic	Tunnel	Vegetation

CONCLUSIONS

This paper evaluates domain related entities of IFC 4 for the green building domain and GIS domain. Version changes of IFC 4 with previous version were analyzed in order to under standard whether IFC 4 is capable of modeling the two domains. A list of domain terminologies was generated using a corpus and text mining techniques. These terminologies were used to find the domain-related classes in IFC 4 using a filter system that we designed. Finally, the filtered results were analyzed using versioning techniques. The methodology framework was implemented and tested using domain schemas from green

building and GIS domains. The results show that our developed methodology could achieve a high accuracy for generating domain-specific terminologies and filtering the domain-related classes. The version change analysis reveal that a large amount of entities regarding the two domains were added in IFC 4 while no deletion had been reported. Some missing entities in IFC 4 for modeling these two domain were also discussed based on the results. There will be a need for IFC to be expanded in order to cope with civil infrastructure modeling, indoor environments and weather information. The framework we developed so far, however, is still semi-automatic. The generation of domain terminology and the filtering of domain-related entities still require manual input. In the future, we aim to adopt more sophisticated Natural Language Processing techniques to facilitate these processes.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support by the Hong Kong Research Grants Council, Grant No. 622812. Any opinions and findings are those of the authors, and do not necessarily reflect the views of the Hong Kong Research Grants Council.

REFERENCES

Ahmad, K., Davies, A., Fulford, H., and Rogers, M. (1994). "What is a term? The semi-automatic extraction of terms from text." *Translation studies: an interdiscipline*, 267-278.

Amor, A., and Ge, C. "Mapping IFC versions." *Proc., Proc of the EC-PPM Conference on eWork and eBusiness in AEC, Portoroz, Slovenia*, 9-11.

Cheng, J., and Das, M. "Ontology-based standardized web services for context aware building information exchange and updating." *Proc., Proc., 2011 ASCE International Workshop on Computing in Civil Engineering, June 19, 2011-June*, 649-656.

Cheng, J. C. P., Deng, Y., and Du, Q. (2013). "Mapping Between BIM Models and 3D GIS City Models of Different Levels of Detail " *13th International Conference on Construction Applications of Virtual Reality*London, United Kingdom.

Drouin, P. "Detection of Domain Specific Terminology Using Corpora Comparison." *Proc., LREC*.

Velardi, P., Missikoff, M., and Basili, R. "Identification of relevant terms to support the construction of Domain Ontologies." *Proc., Proceedings of the workshop on Human Language Technology and Knowledge Management-Volume 2001*, Association for Computational Linguistics, 5.

Wang, H., Akinci, B., Garrett, J. H., Nyberg, E., and Reed, K. A. (2009). "Semi-automated model matching using version difference." *Advanced Engineering Informatics*, 23(1), 1-11.

Wang, H., Akinci, B., and Garrett Jr, J. H. (2007). "Formalism for detecting version differences in data models." *Journal of Computing in Civil Engineering*, 21(5), 321-330.