IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 29, 2019, accepted August 19, 2019, date of publication August 23, 2019, date of current version September 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2937219

Matching Real-World Facilities to Building Information Modeling Data Using Natural Language Processing

QINGSHENG XIE[1], XIAOPING ZHOU[1], JIA WANG[1], XINAO GAO[1], XI CHEN[2], AND CHUN LIU[3]

[1]Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing University of Civil Engineering and Architecture, Beijing 1000044, China
[2]School of Humanity and Law, Beijing University of Civil Engineering and Architecture, Beijing 1000044, China
[3]Hunan Vocational Institute of Safety Technology, Changsha 410000, China

Corresponding authors: Xiaoping Zhou (lukefchou@gmail.com) and Chun Liu (2207835512@qq.com)

This work was supported in part by the Natural Science Foundation of China under Grant 71601013, in part by the Youth Talent Support Program of Beijing Municipal Education Commission under grant no. CIT&TCD201904050, in part by the Beijing Natural Science Foundation under Grant 4174087, in part by the Scientific Research Project of Beijing Educational Committee under Grant SQKM201710016002, in part by the Youth Talent Project of Beijing University of Civil Engineering and Architecture, and in part by the Fundamental Research Funds for Beijing University of Civil Engineering and Architecture under Grant X18010.

ABSTRACT Building Information Modeling (BIM) is a promising technology for building informatics. Currently, an increasing number of applications adopt BIM to improve the building operations and facility management. In these applications, matching real-world facilities to the corresponding BIM items is a fundamental yet challenging task. This study addresses this issue using Natural Language Processing. Firstly, a novel BIM hierarchy tree (HiTree) is proposed to model the original spatial structure relationships of a BIM. Then, the locations of facilities are extracted from natural language through processes of word segmentation, keyword extraction, and semantic disambiguation. Thirdly, an algorithm that matches real-world facilities to the BIM data is developed using the HiTree and the extracted locations. Finally, a concrete case for a 35,000 m^2 library is presented to verify the effectiveness of the proposed solution. BIM has become a common paradigm in the construction industry, and our scheme can facilitate more applications of BIM in building operations and facility management. One of the most representative applications is integrating the BIM data and information within IoT (Internet of Things) system intelligently by matching the BIM data to real-world facilities.

INDEX TERMS Building information modeling (BIM), facility, natural language processing (NLP), facility management.

I. INTRODUCTION

Building Information Modeling (BIM) is a digital representation of physical and functional characteristics of a facility [1]. A BIM records reliable information of a building throughout its life-cycle, including planning, design, construction, operations and facility management phases [2]. By providing interoperable construction data, BIM enables collaboration among stakeholders [3], and provides a reliable basis for decisions during the building life cycle [4]. The past decade has witnessed an avalanche of studies on BIM [5]. Currently, an increasing number of companies from architecture, engineering, construction, operations and facility management (AECO/FM) have embraced BIM as a common paradigm [6], [7].

Building operations and facility management are procedures that integrating people, facilities, technology, and management [8]. During the operations and maintenance phases, there exist many restrictive features, including a large time span, long maintenance cycle, complex content, and a large number of stakeholders, which lead to the relatively low efficiency of traditional operations and facility management [9]. By matching BIM data to building facilities, building information accumulated in the phases of design, construction, operations, and maintenance can be effectively integrated [10], and this can be used to solve the issue of information

The associate editor coordinating the review of this article and approving it for publication was Monjur Mourshed.

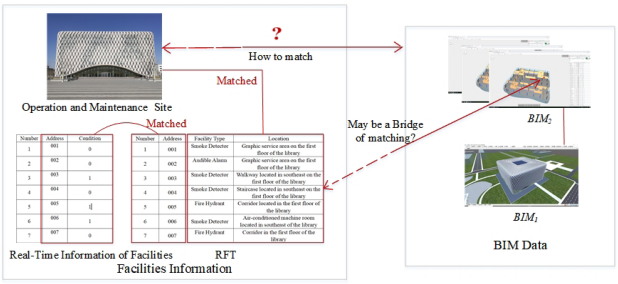


FIGURE 1. The problem of matching real-world facilities to BIM data.

silos [11]. Moreover, BIM provides a comprehensive facility management model for all operators to improve the efficiency of building operations and facility management [12].

The studies and applications of BIM used in the building operations and facility management are still in the initial stages [13]. Studies [14] have noticed that the real-time coordination between the real-world facilities and BIM model remains to be solved. The traditional method of fusing facilities and BIM data basically relies on manual association, which is time-consuming, labor-intensive and inefficient [15]. In order to maximize the value of BIM during the building operations and maintenance stage, it is necessary to break the information barrier between the building operations and facility management system and the BIM model [16]. Therefore, developing an effectively solution that matches real-world facilities to a BIM model is a significant work for BIM servicing in the building operations and facility management stage [17], [18].

A Real-world Facility installation information Table (RFT) records information such as the names, installation locations, and numbers of facilities [19]. Naturally, the RFT is obtained from the construction unit. Information of facilities in a building environment, which is recorded in the RFT, is consistent with the information in the building operations and facility management system. Usually, a unique code is used to represent a facility in the building operations and facility management system. However, currently, as a building was being designed, the designers were not asked to adopt a code to arrange the facility. Therefore, the BIM model would have no available code system for matching the BIM data to the information within the building operations and facility management system. This study addressed this task using a novel perspective. By matching the facilities information in the RFT to the BIM model, the information coordination between the building operations and maintenance system and the BIM model could be realized. Therefore, as can be seen from Figure 1, a RFT is effectively a bridge for matching real-world facilities to BIM data.

In order to solve the above problems, this study proposes a method that matches real-world facilities to BIM data based on NLP. The contributions of this paper include:

1. The definitions involved in the matching scheme are formally defined, and accordingly, a novel solution that matches the real-world facilities to the BIM data is given. The matching solution proposed in this paper can be divided into three steps: the approach of building the HiTree; a model that extracts the natural language, which includes three sub-steps of segmentation, keyword extraction, semantic disambiguation; the matching solution between location information of real-world facilities and BIM data.

2. Systematically present the adopted algorithm. Namely, the algorithms of building the HiTree, keyword extraction, semantic disambiguation, and matching solution.

3. Concretely demonstrate the effectiveness of proposed scheme. This was achieved by taking the BIM model of a library located in Beijing University of Civil Engineer and Architecture (BUCEA) as the experimental model, and combining it with a part of the RFT of the library. Information of the facilities was then represented by the natural language, and matched to the corresponding component information in the BIM model. The experiment shown that the method of matching real-world facilities to BIM data based on natural language processing is effective.

The remainder of this paper is organized as follows. Section 2 introduces related works; Section 3 provides the problem definitions and overall framework adopted in this paper; Section 4 puts forward the approach of building the HiTree; Section 5 conducts the processing procedure of natural language information in the RFT, and Section 6 proposes the solution of matching real-world facilities to BIM data, and demonstrates the effectiveness of the proposed solution by concrete experiments. The last section is a conclusion.

II. RELATED WORKS

A. NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is a science that integrates linguistics, computer science, and mathematics. NLP technology studies the theory and method of realizing human-computer communication through natural language [21]. NLP mainly includes three processes, which are lexical analysis, syntactic analysis and semantic analysis. Among them, lexical analysis could be divided into three steps, namely, word segmentation, tagging, and named entity recognition [6]. NLP has a broad range of applications for knowledge acquisition and retrieval in the construction industry [22]. Al-Qady and Kandil [23] have used NLP to develop ontologies from construction contractual documents. They have used NLP-based conceptual relationship recognition and use the shallow analysis method to automatically extract conceptual relationships from the text of contract documents. The Kappa score and F-measure have significantly improved knowledge acquisition, while constructing legal ontology. The works in [24], [25], and [26] have proposed an NLP-based information extraction system for automated compliance checking with construction regulatory documents. A set of pattern-matching and conflict resolution rules

has been developed that employs syntactic (syntax/grammar-related) and semantic (meaning/context-related) text features during NLP processing [24]. A technique for tagging, separation, and sequencing of regulatory document elements has been proposed to generate high-quality ontology [25]. The proposed algorithm has been tested on the regulatory documents, retrieved from the International Building Code, and the results are promising with high precision [26].

In this paper, NLP technology is introduced into engineering applications, and NLP technology is used to obtain natural language information on the RFT, so as to match real-world facilities to BIM data.

B. INFORMATION EXCHANGE BASED ON IFC AND IFD

BIM encompasses information throughout the life-cycle of a building, and it supports multidisciplinary collaboration and decision making [27]. Industry Foundation Classes (IFC) is a digital protocol used to share information between different software [28]. Through years of development and improvement, IFC standards have been widely recognized in the field of information management. Lee et al. have proposed an IFC-based design information management system [29]. Researchers have also developed some IFC-based BIM servers for information sharing, extraction, and integration [30], [31]. Based on BIMserver.org, an open query language for BIM called BIMQL was developed, which can provide a flexible data retrieval interface with domain-specific and platform independence [32]. An open IFC model analysis repository has also been presented to facilitate the interoperability of building information [33]. Meanwhile, the experience of model-based interoperability issues when exchanging BIM between various tools has been reported [3]. Extending IFC with Extensible Markup Language (XML) and exchanging information with model views has been discussed in two papers [34], [35]. Therefore, IFC has a feasible extension mechanism and many related tools, and this lays a solid foundation for BIM interoperability.

The International Framework for Dictionaries (IFD) is an internationally open library. In the IFD standard, concepts and terms are defined, semantically described and given a unique identification number, which is named a Global Unique Identifier (GUID) [36]. The GUID provided by the IFD is critical for supporting the accurate exchange of building information. Supported by multilingual terms, the IFD provides a mapping method from concept to IFC entities and attributes, supporting the distinction between concepts and specific language instances [34]. Shayeganfar et al. have conducted a case study on how to implement the IFD library using semantic web technology, which bridges the gap between BIM and web services [37]. Therefore, supported by the IFD library and semantic web, terms in a particular language or their synonyms can be mapped to entities in a data schema like IFC, to achieve the precise exchange of building information precisely.

TABLE 1. Notations.

Symbol	Description
$T=(b,s,r,f)$	HiTree
$b=\{s,r\}$	BIM b is a collection of vertical and horizontal spatial allocation
$s,\ s_i$	Vertical spatial distribution of the HiTree
$r,\ r_i$	Horizontal spatial distribution of HiTree
f,f_i	Facilities located in various sub-spaces in the building
n,m,k,i	Natural numbers used in this paper
Q^0	Untreated natural language information of real-world facilities
Q	Natural language word sequence after segmentation
$\bar{Q}$	Keyword sequence
W	Keyword sequence based on IFD semantic disambiguation
A_0	Reference vector
B_0， b_i	Matching path matrix
A， a_i	MP matrix A is a matching matrix, a_i is one of its vectors
C， C_i	Feature vector C, C_i is an element
N， N_i	N is a matching value vector， while N_i is a matching value

III. PRELIMINARIES

A. PROBLEM DEFINITION

The matching between real-world facilities and BIM data can be achieved when the extracted locations are matched to the corresponding components in the BIM model. Since computers cannot recognize the natural language information directly, we need to adopt NLP to process the natural language information in the RFT. Based on the processes of natural language segmentation and keyword extraction, the locations of facilities are extracted. On the basis of the IFD semantic disambiguation, the previously extracted locations information is consistent with the representation of the corresponding data in the BIM model, which facilitates the accurately matching between the real-world facilities and the BIM data. In this paper, the BIM model was constructed using a tree structure according to the spatial structure relationship of the building, which was called the HiTree. Therefore, the matching between the real-world facilities and the BIM model was equivalent to the matching of the extracted locations of the facilities and the HiTree.

In order to define the matching issue fluently, we firstly have the following definitions. The notations frequently used in this paper are listed in Table 1.

Definition 1 (RFT): A Real-world Facilities installation information Table (RFT) is a table that encompasses the facility information, including names, installation locations, numbers and the like.

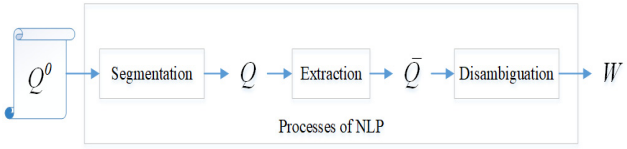


FIGURE 2. Illustration of the procedure of the LIEModel.

Definition 2 (BIM): A BIM b is a digital representation of the architectural geometric information and semantic information, which is the root node in the HiTree, and $b = \{s, r\}$.

Definition 3 (Spatial Allocation): Spatial allocation, which is the spatial distribution in the BIM model, can be divided into two types: vertical spatial allocation and horizontal spatial allocation.

The vertical spatial allocation, denoted as s, is the spatial separation of the building in the vertical direction, and includes information such as the elevation; the horizontal spatial allocation, denoted as r, is the spatial separation of the building in the horizontal direction, and forms several subspaces, such as rooms, halls.

Thus, in this paper, we have $s = \{s_1, s_2, \ldots, s_n\}$, $s_i (1 \leq i \leq n)$ denotes the i-th layer in a multiple storey building; $r = \{r_1, r_2, \ldots, r_m\}$, $r_i (1 \leq i \leq m)$ represents the i-th sub-space in the building.

Definition 4 (Facility): A facility f is a building asset within the facility management domain. A facility f is distributed in the building environments to facilitate something. Thus, we have $f = \{f_1, f_2, \ldots, f_k\}$, $f_i (1 \leq i \leq k)$ represents the i-th facility in the building.

Definition 5 (HiTree): A HiTree T is transformed from the BIM model based on the spatial structure relationship of the building, and we have $T = (b, s, r, f)$. The depth of the HiTree is denoted by T_D, and its number of layers is denoted by T_k.

Definition 6 (LIEModel): The location information extract Model (LIEModel) is proposed to process the natural language in the RFT, which encompasses the locations of facilities.

The raw natural language information of facilities is represented by Q^0. Based on the natural language segmentation, a natural language word sequence Q is generated. By means of the natural language keyword extraction, a keyword sequence is obtained, which is denoted as $\bar{Q}$. Then, the keyword sequence is expressed as W after the process of IFD semantic disambiguation. Figure 2 shows the relationship of the natural language sequences.

Definition 7 (MP Matrix): A MP matrix A is a matrix that generated through the comparison of the reference vector A_0 and the matching path matrix B_0. While a reference vector A_0 is consisted of all the nodes of HiTree; a matching path matrix B_0 is a collection of all matching path.

Definition 8 (Feature Vector): A feature vector C is generated based on contradistinction between keyword sequence W and the reference vector A_0.

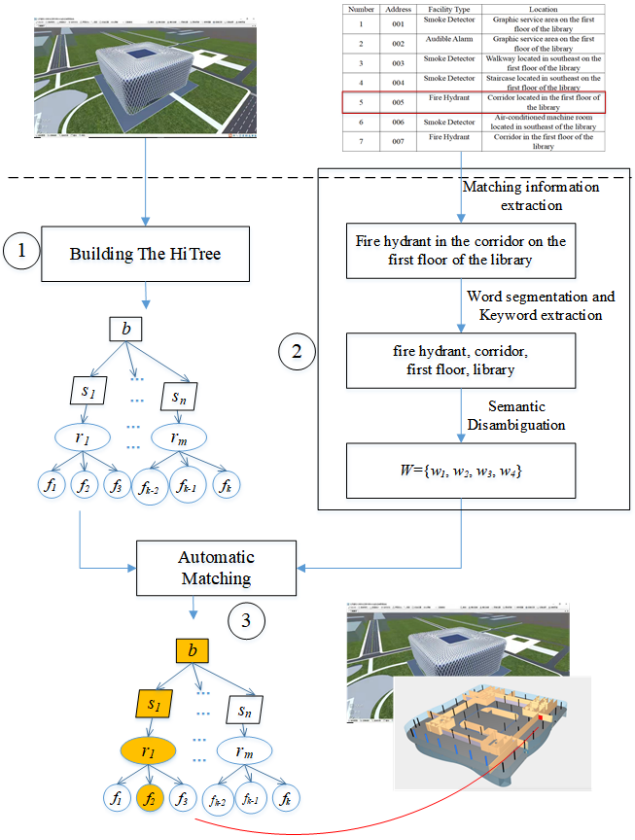


FIGURE 3. The overall framework of this study.

Definition 9 (Matching Value): The matching value N is the product of the matching matrix A and the feature vector C.

B. OVERALL FRAMEWORK

The matching solution proposed in this study can be divided into three steps, namely the construction of the HiTree; the LIEModel combined with natural language segmentation, keyword extraction and semantic disambiguation; and the matching scheme between the location information of real-world facilities and BIM data. The processes are shown in Figure 3

1. Building the HiTree. According to the original spatial structure relationship of the building, the BIM model is described as a tree structure with a depth of $T_D = 4$. The nodes in the first layer represent the whole model, the nodes in the second layer represent the vertical spatial allocations of the building, and the nodes in the third layer represent the horizontal spatial allocations of the building, and the fourth layer represent the facilities that are located in the sub-spaces of the building.

2. Obtaining the locations of facilities in the RFT based on the LIEModel. Since the locations information in the RFT is described by natural language, in order for the computer to understand this natural language information, it is necessary to first conduct NLP on it firstly, namely natural

language segmentation and keyword extraction. Due to the differences in natural language expressions, the same object always has multiple descriptive forms. In order to ensure that the same real-world facilities get the same understanding among stakeholders, the IFD was used to carry out semantic disambiguation for the key information that was processed by NLP, so as to obtain a standardized natural language word sequence.

3. Proposing a matching method. The number of matching paths of the HiTree is equal to the number of leaf nodes. In this study, all possible matching paths were described in a matrix, and the standardized natural language word sequences were described as column vector forms. Then the matching algorithm was given.

IV. HITREE CONSTRUCTION

A tree structure is a significant type of nonlinear data structures, which is a hierarchical structure defined by branch relationships. According to the IFC standard, the locations of building facilities in the building have an obvious spatial hierarchical relationship. Therefore, BIM data can be described as a tree structure. It is logical to describe the BIM as a tree structure that contains BIM models, floors, sub-spaces, and facilities. The number of layers in the HiTree is denoted as T_k.

The algorithm for constructing the HiTree is given below.

Algorithm 1 Building the HiTree

Input: b: BIM
Output: T
1: function BIMParser(b)
2: Obtain building model data
3: for each vertical spatial distribution
4: Collect all the vertical spatial nodes s
5: for each s
6: Collect all the horizontal spatial nodes r
7: for each r
8: Collect all the equipment nodes f
9: end for
10: Get all the nodes b, s, r, f
11: Build the HiTree with nodes b, s, r, f. *and* $T = \{b, s, r, f\}$
12: end for

As can be seen from Figure 4, BIM can be described as a tree structure with a depth of 4. The $T_k = 1$ layer represents the first layer of the HiTree, which is the root node in the HiTree and represents the BIM as a whole; The $T_k = 2$ layer represents the second layer of the HiTree, which is the spatial layer relationship of the first layer according to the vertical spatial distribution. The $T_k = 3$ layer represents the third layer of the HiTree, and the third layer node represents the spatial area divided by the horizontal space distribution in the building floor, and within this are sub-spaces in the HiTree, such as rooms, halls, corridors, etc.; The $T_k = 4$ layer represents the fourth layer of the HiTree, and the fourth layer nodes represent the building equipment distributed in the corresponding subspace.

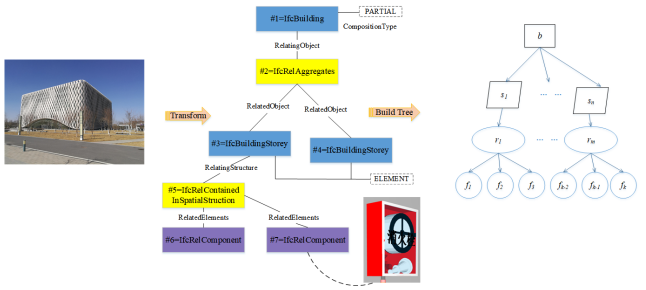


FIGURE 4. Conceptual diagram of HiTree.

V. KEYWORD EXTRACTION AND SEMANTIC DISAMBIGUATION

The information in the point table of the construction equipment is described in natural language. However, the computer cannot directly recognize the natural language information. Therefore, the initial information is first processed using NLP. Firstly, this section introduces the methods of word segmentation and keyword extraction based on NLP. Then, taking the ''fire hydrant in the hallway on the first floor of the library'' as an example, and the natural language segmentation and keyword acquisition were carried out. Finally, the IFD library was introduced to semantically disambiguate the keywords. Thereby, the processing of the natural language information on the RFT was completed.

A. NLP-BASED WORD SEGMENTATION AND KEYWORD EXTRACTION

In order to obtain information from natural language, the computer must first divide the natural language into a sequence of words with certain meanings. This process is called natural language segmentation. The main word segmentation algorithms include lexicographic matching and statistical machine learning. In this study, we adopted the LTP word segmentation module, and the algorithm used combines the two methods. The algorithm can not only obtain accurate word segmentation results by machine learning, but also flexibly introduce external resources such as a dictionary.

According to the previous definitions, the natural language statement which contains the building equipment information was denoted as Q^0. Based on the natural language segmentation process, Q^0 was divided into several natural language phrases with certain meanings, resulting in $Q = \{Q_1, Q_2, \ldots, Q_L\}$.

In order to accurately extract natural language keywords, we must firstly establish a thesaurus Q_0, which covers a large number of architectural concepts and descriptions. The IFD standard, as an international open library, contain almost all concepts and terms in the AECO/FM domain. In this study, the thesaurus Q_0 was established by extracting concepts and terms from IFD specification. Therefore, the thesaurus Q_0 was consistent with the information in the IFD specification.

Accordingly, the thesaurus Q_0 could be properly established properly, and it is suitable for all kinds of buildings.

Then, the construction equipment information word sequence Q, which is based on natural language word segmentation, was compared with the words in the thesaurus Q_0. If a word exists in the thesaurus Q_0, it will be taken out as a keyword to form a natural language keyword sequence $\overline{Q} = \overline{Q}_1, \overline{Q}_2, \ldots, \overline{Q}_n\}$. The keyword extraction algorithm is proposed below.

Taking the "fire hydrant in the hallway on the first floor of the library" as an example, based on the LTP natural language segmentation process, the input sentence is cut into word sequences. Specifically, English have an obvious boundary, which is different from Chinese. This boundaries are helpful for word segmentation of English. Thus, the result of the word segmentation with LTP is "fire hydrant in the hallway on the first floor of the library".

According to Algorithm 2, we firstly obtain the vector Q = {fire hydrant, in, the, hallway, on, the, first floor, of, the, library}, and following comparison between the elements in vector Q with thesaurus Q_0, we then acquire the natural language keyword sequence $\overline{Q}$. Therefore $\overline{Q}$ = {fire hydrant, hallway, first floor, library}.

Algorithm 2 Keyword Extraction

Input: Q
thesaurus Q_0
Output: $\overline{Q}$
1: Collect all the noun or noun phrase in Q
2: Let $Q = \{Q_1, Q_2, \ldots, Q_L\}$
3: for($i = 1$; i<= L; ++ i)
4: if $Q_i \in Q_0$,do
5: Put Q_i in $\overline{Q}$
6: else
7: return $\overline{Q}$
8: end for

B. IFD-BASED SEMANTIC DISAMBIGUATION

Due to the differences in language, culture, and expression of natural language, there are usually multiple expressions when describing the same objective entity. For example, in a building, both corridors and aisles represent horizontal traffic spaces within buildings, and they are also known as a "hallway". However, all these expressions all represent the same objective entity. The same object, which has multiple expressions, makes the exchange of building information difficult. And it is obviously a serious problem to not accurately make the same understanding of the same architectural concept. To this end, the IFD needs to be introduced to solve this problem and ensure the consistency of concepts in the construction field. The IFD defines building-related concepts and attributes with a GUID and links various descriptions of these concepts and attributes to the corresponding GUID. By transforming architectural concepts and attributes into a GUID, the differences in understanding, which come from differences in natural language expression, can be eliminated to achieve semantic disambiguation.

Consequently, this study adopted the IFD to semantically disambiguate the resulting keywords. Each extracted keyword could find its unique corresponding GUID in IFD dictionary to eliminate ambiguity, and finally get semantically disambiguated keywords, which denoted as $W = \{w_1, w_2, \ldots, w_n\}$. Algorithm 3 summarizes the whole process of semantic disambiguation.

Algorithm 3 Semantic Disambiguation

Input: keyword sequence $\overline{Q}$
International Framework for Dictionaries(IFD)
Output: Standardized keyword sequence W
1: for each keyword $\overline{Q}_i$ ($i = 1$; i<= L; ++ i)
2: if $\overline{Q}_i \in$ IFD, do
3: Put $\overline{Q}_i$ in W
4: return the GUID of $\overline{Q}_i$
5: else
6: return W
7: end for

VI. MATCHING METHOD OF FACILITIES AND BIM

A. MATCHING MATRIX

As can be seen from Figure 5, the building information model tree has $n + m + k$+1 nodes, and all nodes are represented by vector A_0, which is called the reference vector, and is given by:

$$A_0 = [b, s_1, s_2, \ldots, s_n, r_1, r_2, \ldots, r_m, f_1, f_2, \ldots, f_k] \quad (1)$$

From the root node b of the HiTree to the leaf node $f_i(1 \le i \le k)$, there are k non-repetitive paths, and the k paths are represented as a matrix B_0, namely:

$$B_0 = \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_k \end{bmatrix} = \begin{bmatrix} bs_1r_1f_1 \\ bs_1r_1f_2 \\ \ldots \\ bs_nr_mf_k \end{bmatrix}_{k\times 4} \quad (2)$$

where $b_i(1 \le i \le k)$ represents the i-th matching path in the HiTree.

Then, the elements of each path vector b_i in the path matrix B_0 are compared with the elements in the reference vector A_0, and a new vector a_i is formed in the process of the

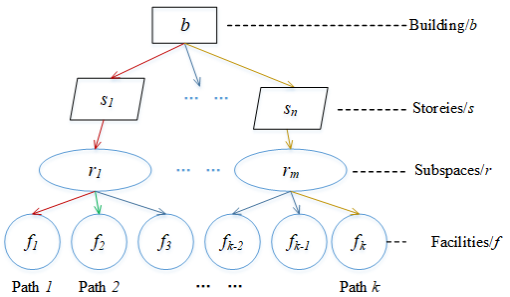


FIGURE 5. Matching paths in the HiTree.

comparison, and the dimension of the vector a_i is consistent with the reference vector A_0.

If an element in the vector b_i exists in the reference vector A_0, the corresponding position in the vector a_i is assigned a value of 1. When all the elements in the path vector b_i have been assigned in the vector a_i, the positions in the vector a_i that have no value are assigned as 0. The vector a_i forms a matrix A, named the matching matrix. Then, the matching matrix A is expressed as follows:

$$A = \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_k \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 1\,0\dots0 & 1\,0\dots0 & 1\,0\dots0 \\ 1 & 1\,0\dots0 & 1\,0\dots0 & 0\,0\,1\dots0 \\ \dots & & & \\ 1 & 0\,0\dots1 & 0\,0\dots & 1\,0\,0\dots1 \end{bmatrix}}_{n \quad\quad\quad m \quad\quad\quad k} \tag{3}$$

B. FEATURE VECTORS OF NATURAL LANGUAGE

Based on Definition 6, the natural language sentence Q^0, which contains the construction equipment installation information, is subjected to natural language processing and semantic disambiguation to obtain a standardized word sequence W, and $W = \{w_1, w_2, \dots, w_n\}$. The element $w_i (1 \le i \le n)$ in W is then compared with the element in the reference vector A_0, and we get the feature vector C. The feature vector C is defined, so as to have the same dimension as the reference vector A_0.

$$C_i = \begin{cases} 1 w_i \in A_0 \\ 0 w_i \notin A_0 \end{cases}, 1 \le i \le n \tag{4}$$

If an element w_i exists in the reference vector A_0, the corresponding element in the feature vector C is assigned a value of 1. If the element w_i does not exist in the reference vector A_0, the element is discarded and the position in the feature vector C, which has no value, is assigned as 0.

Taking the word sequence $W = \{b, s_1, r_1, f_1\}$ as an example, based on the comparison of W and the elements in the reference vector A_0, the feature vector C is obtained as follow:

$$C = \underbrace{\begin{bmatrix} 1 & 1\,0\dots0 & 1\,0\dots0 & 1\,0\dots0 \end{bmatrix}^T}_{n \quad\quad m \quad\quad k} \tag{5}$$

C. MATCHING METHOD

The determination method of the matching matrix A and the feature vector C is given above. Then the matching method is defined as follows:

$$A \times C = N \tag{6}$$

where, $N_i (1 \le i \le n)$ is the product of the i-th row vector of matching matrix A and the feature vector C. The value of N_i represents the degree of matching between natural language feature vector C and each matching path b_i in the HiTree. Then, the value of N_i is sorted, and the matching path, which corresponds to the maximum value, has the highest matching degree with natural language information. If $N_i = 4$, the natural language information has complete information about the building equipment, and the building equipment represented by the leaf node matches the natural language information. Algorithm 4 summarizes the process of matching method.

Algorithm 4 Matching Method

Input: Reference Vector A_0
HiTree T
Standardized thesaurus Q_0
$W = \{w_1, w_2, \dots, w_n\}$
Output: Matching facilities information to HiTree
1: Get all the matching paths
2: Let matching paths matrix $B_0 = \{b_1, b_2, \dots, b_n\}$.
3: for each matching path b_i
4: if $b_{ij} \in A_0$, do
5: $a_{ij} = 1$
6: else $a_{ij} = 0$
7: Let $A = \{a_1, a_2, \dots, a_n\}$.
8: for each matching path w_i
9: if $w_i \in A_0$, do
10: $C_i = 1$
11: else $C_i = 0$
12: Let $C = [C_1, C_2, \dots, C_n]^T$.
13: Get N using equation (6).
14: Get the maximum in N, do
15: Matching the corresponding HiTree information and keyword information
16: end for

D. EMPIRICAL STUDIES

In this subsection, we took the BIM model of a university library, which is located in the Beijing University of Civil Engineering and Architecture, as an experimental model. By matching part of construction equipment information to the corresponding BIM components, the feasibility and effectiveness of the method, which is proposed above, are demonstrated.

The experimental building is located in Beijing University of Civil Engineering and Architecture in Beijing, the capital city of China. It has a modern architecture with a total construction area of 35000 square meters and a total investment of 320 million RMB yuan. The main structure of the experimental building consists of seven floors above ground and one floor underground.

The detailed modeling and matching process is demonstrated below and in Figure 6.

1. The design BIM was established using Autodesk Revit 2016, and it includes models of the architecture/structure and construction equipment of machine, electric, and plumbing (MEP) systems. The design BIM was organized based on the spatial relationship in the IFC files. The experimental model contained 17669 facilities, and the quantity of fire facilities was 1191. Figure 6(a) is the front view of the experimental building. Figure 6(b) show the BIM model of the experimental building.

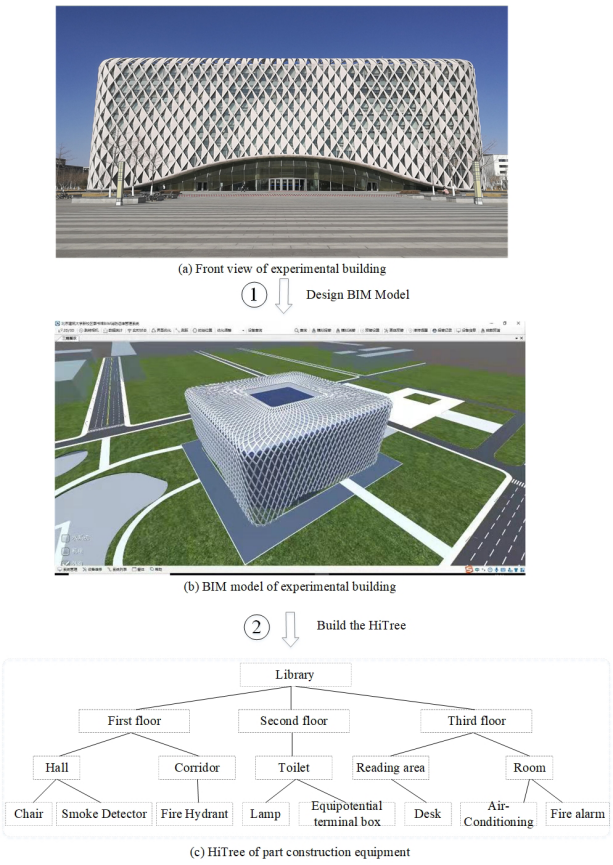


FIGURE 6. Process of modeling and building the HiTree.

2. Based on the original spatial structure of the BIM model, a HiTree, which contains part of the construction equipment in the experimental building, was built using Algorithm 1. This HiTree is shown in Figure 6(c). Then, we found the reference vector A_0 based on the HiTree, and $A_0 =$ [Library, First floor, Second floor, Third floor, Hall, Corridor, Toilet, Reading area, Room, Chair, Smoke Detector, Fire Hydrant, Lamp, Equipotential terminal box, Desk, Air-Conditioning, Fire alarm].

Assuming that the natural language information of the construction equipment to be matched is Q^0 ''fire hydrant in the hallway on the first floor of the library''. The Language Technology Platform (LTP) was adopted to make the word segmentation, and the results of the natural language segmentation were ''fire hydrant, in, the, hallway, on, the, first floor, of, the, library''. Then, based on the Algorithm 2, a keyword sequence $\overline{Q} =$ {fire hydrant, first floor, hallway, library}, was extracted.

In the extracted keyword sequence $\overline{Q}$, each element was a natural language description of a corresponding construction concept or description. Due to the diversity of natural language descriptions, the difference, which is caused by the diversity of understanding of the construction equipment information, will be the obstruction of the communication of construction equipment information. Therefore, the IFD library was introduced, and the global unique identifier (GUID) in the IFD library was applied to transform the keywords into a unique description of the concepts in the same architectural field. In this example, the word ''corridor'' and the word ''hallway'' in the pre-built building information thesaurus Q_0 should have the same building space description, so the two words had the same GUID. Since each node on the HiTree is named according to the pre-built building information thesaurus Q_0, in order to achieve an accurate matching between the natural language information from the RFT and the construction equipment information, the extracted keyword sequence is $\overline{Q} =$ {library, one layer, The aisle, fire hydrant} and it is converted into a consistent expression with the pre-built building information thesaurus Q_0. Then, a standardized word sequence $W =$ {library, one floor, corridor, fire hydrant} was obtained based on the Algorithm 3. According to the method proposed above, a feature vector C was generated, and the feature vector C is shown below.

$$C = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T \quad (7)$$

Since the experimental HiTree had eight leaf nodes, there were eight matching paths. They are shown in Table 2:

According to Algorithm 4, the MP matrix A can be obtained as:

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (8)$$

TABLE 2. Matching path.

	Passed Nodes of each Path			
	Nodes of First Layer	Nodes of Second Layer	Nodes of Third Layer	Nodes of Fourth Layer
P1	Library	First floor	Hall	Chair
P2	Library	First floor	Hall	Smoke Detector
P3	Library	First floor	Corridor	Fire Hydrant
P4	Library	Second floor	Toilet	Lamp
P5	Library	Second floor	Toilet	Equipotential terminal box
P6	Library	Third floor	Reading area	Desk
P7	Library	Third floor	Room	Air-Conditioning
P8	Library	Third floor	Room	Fire alarm

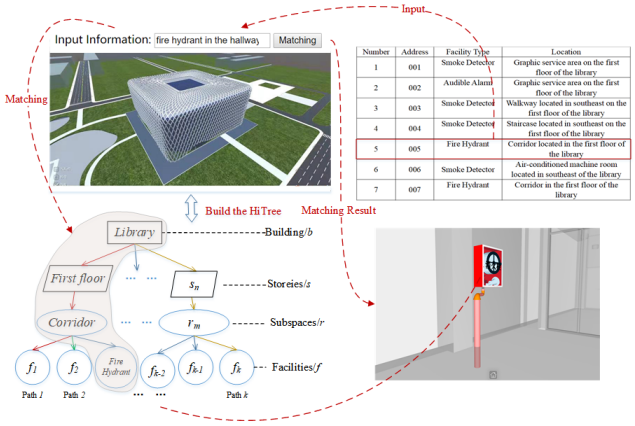


FIGURE 7. Illustration of the matching result of the experimental BIM.

TABLE 3. Matching results.

	# of facilities	# matched	Matching rate
Fire Facilities	1191	1089	91.43%

Then, based on equations (7) and (8), a column vector N was built.

$$N = A \times C = \begin{bmatrix} 2 & 2 & 4 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}^T \quad (9)$$

From the matching value vector N, it can be noticed that the product value of the a_3 vector and matching vector C is the largest. This meant that the building equipment information represented by the corresponding third matching path was matched to the natural language information. Therefore, the location information ''fire hydrant in the hallway on the first floor of the library'' matched to the third leaf node in the HiTree. The matching result is illustrated in Figure 7.

Based on a previous successful experiment on some specific facilities, we then evaluated the effectiveness of the proposed solution by further matching more facilities to the BIM data. In our experiments, the size of the library model was 41.65MB, and this model contained 17669 facilities. We extracted 1191 fire facilities of them to conduct the matching experiments. Table 3 shows the performance of the proposed scheme. Among 1191 facilities tested, 1089 facilities were matched successfully. The matching rate was 91.43%. In the meantime, there are still some facilities unmatched. Possible reasons are discussed below. The IFD library is adopted to build the thesaurus Q_0 and do semantic disambiguation. However, the IFD library may not contain enough concepts in the AECO/FM domain. This may cause some important information processed imprecisely. In conclusion, this experiments show that the matching solution proposed in this paper is effective and efficient.

VII. CONCLUSION

Building operations and FM tasks involve multiple stakeholders over a long time period, and they require comprehensive and high-quality data for correct and reliable operations and maintenance. However, the information needed for operations and maintenance, which is typically available from manufacturers, is time-consuming and laborious. In fact, collecting available sufficient information about the building facilities for any operations and FM tasks is considered a key challenge.

BIM has become an important part of building digitizing and informatics, and its applications are involved in the entire life of a building. In order to address the information collaboration issue of real-world facilities and BIM models, this paper presents a solution that matches real-world facilities to BIM data using NLP. The matching solution proposed in this study can be divided into three steps: namely the construction of the HiTree; a LIEModel combined with natural language segmentation, keyword extraction and semantic disambiguation; and the matching schema between extracted locations information of real-world facilities and BIM data. The experiments demonstrated the effectiveness and efficiency of the proposed matching method.

The proposed scheme can effectively improve the implementation capability of BIM of operations and facilities management systems and facilitate more applications of BIM in building operations and facility management. Moreover, this study may provide a reference for BIM cross-stage data fusion, which is a significant task for Internet of Things (IoT). However, the proposed matching solution does have the following limitations. (1) The process of extracting information from natural language relies on the words contained in the thesaurus Q_0. The thesaurus Q_0 is based on the IFD library. So the IFD library should contain enough concepts in the AECO/FM domain and be continuously enriched. (2) Currently, only simple sentences are supported. Complex sentences containing verbs and operators are not supported yet. Further improvement should be made to support processing sentences with verbs and operations properly. (3) Because all the nodes in the HiTree are represented in a vector, when the IFD document size exceeds a certain level, the dimensions of the matching matrix will become difficult to calculate. Thus, further improvement should be made to support matching larger size of BIM model.

REFERENCES

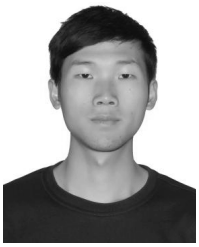
[1] National BIM Standard-United States. *Frequently Asked Questions About the National BIM Standard-United States*. Accessed: Apr. 20, 2019. [Online]. Available: https://www.nationalbimstandard.org/faqs#faq1

[2] X. Zhou, J. Zhao, J. Wang, D. Su, H. Zhang, M. Guo, and Z. Li, ''OutDet: An algorithm for extracting the outer surfaces of building information models for integration with geographic information systems,'' *Int. J. Geograph. Inf. Sci.*, vol. 33, no. 7, pp. 1444–1470, 2019.

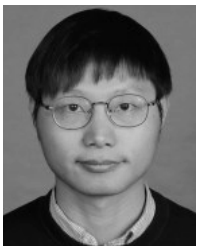
[3] J. Steel, R. Drogemuller, and B. Toth, ''Model interoperability in building information modelling,'' *Softw. Syst. Model.*, vol. 11, no. 1, pp. 99–109, 2012.

[4] R. Eadie, M. Browne, H. Odeyinka, C. McKeown, and S. McNiff, ''BIM implementation throughout the UK construction project lifecycle: An analysis,'' *Autom. Construct.*, vol. 36, pp. 145–151, Dec. 2013.

[5] Z. Pezeshki and S. A. S. Ivari, "Applications of BIM: A brief review and future outline," *Arch. Comput. Methods Eng.*, vol. 25, no. 2, pp. 273–312, 2018.
[6] J.-R. Lin, Z.-Z. Hu, J.-P. Zhang, and F.-Q. Yu, "A natural-language-based approach to intelligent data retrieval and representation for cloud BIM," *Comput. Aided Civil Infrastruct. Eng.*, vol. 31, no. 1, pp. 18–33, 2016.
[7] A. Jennifer. (2013). *Expected BIM Trends in*. Accessed: May 6, 2013. [Online]. Available: http://www10.aeccafe.com/blogs/aecsanjay/2012/12/20/expected-bim-trends-2013/
[8] R. A. Kivits and C. Furneaux, "BIM: Enabling sustainability and asset management through knowledge management," *Sci. World J.*, vol. 2013, Aug. 2013, Art. no. 983721.
[9] M. Marjani, F. Nasaruddin, A. Gani, A. Karim, I. A. T. Hashem, A. Siddiqa, and I. Yaqoob, "Big IoT data analytics: Architecture, opportunities, and open research challenges," *IEEE Access*, vol. 5, pp. 5247–5261, 2017.
[10] W. Shen, Q. Hao, H. Mak, J. Neelamkavil, H. Xie, J. Dickinson, R. Thomas, A. Pardasani, and H. Xue, "Systems integration and collaboration in architecture, engineering, construction, and facilities management: A review," *Adv. Eng. Inform.*, vol. 24, no. 2, pp. 196–207, 2010.
[11] I. Motawa and A. Almarshad, "A knowledge-based BIM system for building maintenance," *Automat. Construct.*, vol. 29, pp. 173–182, Jan. 2013.
[12] B. Becerik-Gerber and K. Kensek, "Building information modeling in architecture, engineering, and construction: Emerging research directions and trends," *J. Prof. Issues Eng. Educ. Pract.*, vol. 136, no. 3, pp. 139–147, 2009.
[13] S. Azhar, "Building information modeling (BIM): Trends, benefits, risks, and challenges for the AEC industry," *Leadership Manage. Eng.*, vol. 11, no. 3, pp. 241–252, 2011.
[14] R. Volk, J. Stengel, and F. Schultmann, "Building information modeling (BIM) for existing buildings—Literature review and future needs," *Autom. Construct.*, vol. 38, pp. 109–127, Mar. 2014.
[15] K. Orr, Z. Shen, P. K. Juneja, N. Snodgrass, and H. Kim, "Intelligent facilities: Applicability and flexibility of open BIM standards for operations and maintenance," in *Proc. Construct. Res. Congr. Global Netw.*, 2014, pp. 1951–1960.
[16] B. Hardin and D. McCool, *BIM and Construction Management: Proven Tools, Methods, and Workflows*. Hoboken, NJ, USA: Wiley, 2015.
[17] J. Teizer, M. Wolf, O. Golovina, M. Perschewski, M. Neges, and M. König, "Internet of Things (IoT) for integrating environmental and localization data in Building Information Modeling (BIM)," in *Proc. Int. Symp. Autom. Robot. Construct. (ISARC)*, 2017, pp. 603–609.
[18] T. Beach, I. Petri, Y. Rezgui, and O. Rana, "Management of collaborative BIM data by federating distributed BIM models," *J. Comput. Civil Eng.*, vol. 31, no. 4, 2017, Art. no. 04017009.
[19] G. D. Oberlender, *Project Management for Engineering and Construction*. New York, NY, USA: McGraw-Hill, 1993.
[20] C. S. Mellish, *Computer Interpretation of Natural Language Descriptions*, vol. 21. New York, NY, USA: Wiley, 1985.
[21] G. G. Chowdhury, "Natural language processing," *Annu. Rev. Inf. Sci. Technol.*, vol. 37, no. 1, pp. 51–89, 2003.
[22] M. Bilal, L. O. Oyedele, J. Qadir, K. Munir, S. O. Ajayi, O. O. Akinade, H. A. Owolabi, H. A. Alaka, and M. Pasha, "Big data in the construction industry: A review of present status, opportunities, and future trends," *Adv. Eng. Inform.*, vol. 30, no. 3, pp. 500–521, 2016.
[23] M. Al Qady and A. Kandil, "Concept relation extraction from construction documents using natural language processing," *J. Construct. Eng. Manage.*, vol. 136, no. 3, pp. 294–302, 2009.
[24] J. Zhang and N. El-Gohary, "Extraction of construction regulatory requirements from textual documents using natural language processing techniques," *Comput. Civil Eng.*, vol. 30, no. 2, pp. 453–460, 2012.
[25] J. Zhang and N. El-Gohary, "Automated regulatory information extraction from building codes: Leveraging syntactic and semantic information," in *Proc. Construct. Res. Congr., Challenges Flat World*, 2012, pp. 622–632.
[26] J. Zhang and N. M. El-Gohary, "Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking," *J. Comput. Civil Eng.*, vol. 30, no. 2, 2013, Art. no. 04015014.
[27] NBIMS. *National Building Information Modeling Standard V2—United States*. Accessed: May 20, 2013. [Online]. Available: http://www.nationalbimstandard.org/
[28] V. Thein, "BIM interoperability through a vendor-independent file format," Bentley Softw., Exton, PA, USA, White Paper, 2011. [Online]. Available: https://www10.aeccafe.com/link/BIM-Interoperability-Through-Vendor-Independent-File-Format/36550/link_download/No/IFC_WP%5B1%5D.pdf
[29] K. Lee, S. Chin, and J. Kim, "A core system for design information management using industry foundation classes," *Comput. Aided Civil Infrastruct. Eng.*, vol. 18, no. 4, pp. 286–298, 2003.
[30] J. Beetz, L. van Berlo, R. de Laat, and P. van den Helm, "BIMserver. org—An open source IFC model server," in *Proc. CIP W78 Conf.*, Nov. 2010, p. 8.
[31] X. Zhou, J. Wang, M. Guo, and Z. Gao, "Cross-platform online visualization system for open BIM based on WebGL," *Multimedia Tools Appl.*, pp. 1–16, Mar. 2018. doi: 10.1007/s11042-018-5820-0.
[32] W. Mazairac and J. Beetz, "BIMQL—An open query language for building information models," *Adv. Eng. Inform.*, vol. 27, no. 4, pp. 444–456, 2013.
[33] R. Amor and J. Dimyadi, "An open repository of IFC data models and analyses to support interoperability deployment," in *Proc. CIB W78 Conf.*, Istanbul, Turkey, Sep. 2010, pp. 14–16.
[34] J. P. Zhang, J. R. Lin, Z. Z. Hu, and F. Q. Yu, "Research on IDM-based BIM process information exchange technology," in *Proc. 14th Int. Conf. Comput. Civil Building Eng.*, 2012, pp. 1–9.
[35] C. M. Eastman, Y. S. Jeong, R. Sacks, and I. Kaner, "Exchange model and exchange object concepts for implementation of national BIM standards," *J. Comput. Civil Eng.*, vol. 24, no. 1, pp. 25–34, 2009.
[36] H. Bell, L. Bjørkhaug, A. Bjaaland, and R. Grant. (2008). *IFD Library White Paper*. Accessed: Jan. 2012. [Online]. Available: https://www.Ifd-library.org/images/IFD_Library_White_Paper_2008-04-10_I.pdf
[37] F. Shayeganfar, A. Mahdavi, G. Suter, A. Anjomshoaa, A. Zarli, and R. Scherer, "Implementation of an ifd library using semantic Web technologies: A case study," in *Proc. ECPPM eWork eBusiness Archit., Eng. Construct.*, 2008, pp. 539–544.



QINGSHENG XIE received the B.E. degree from the School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, in 2018, where he is currently pursuing the Ph.D. degree. His research interests include building information modeling and data processing.

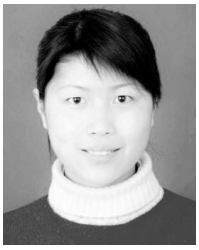


XIAOPING ZHOU received the B.E. and M.E. degrees from Beijing Information Science and Technology University, Beijing, China, in 2006 and 2009, respectively, and the Ph.D. degree in computer science from the Renmin University of China, Beijing, China, in 2018. He is currently an Associate Professor and leads the Smart City Digitization Laboratory, Beijing University of Civil Engineering. He is also the Chief Scientist with BIMWinner Company Ltd., where he founded the www.bos.xyz, in 2018. His research interests include construction big data, building information modeling, social big data, and artificial intelligence.

JIA WANG received the Ph.D. degree from Beijing Jiaotong University, China, in 2014. She is currently a Professor with the Beijing University of Civil Engineering and Architecture. Her research interests include building information modeling and fire information system.

XINAO GAO received the B.E. degree from the School of Science, Beijing University of Civil Engineering and Architecture, in 2018, where he is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering. His research interests include building information modeling and data retrieval.

XI CHEN received the Ph.D. degree from the Chinese Academy of Social Sciences, in 2014. She held a postdoctoral position with Johns Hopkins University, USA. She is currently a Lecturer with the Beijing University of Civil Engineering and Architecture. Her research interests include urban studies and English for scientific and techonological purposes.



CHUN LIU is currently an in-service graduate student with the Hunan Vocational Institute of Safety Technology. Her main research interests include computer communication and information security.