Automation in Construction 102 (2019) 230–244

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Automated management of green building material information using web crawling and ontology

Sim-Hee Hong[a], Seul-Ki Lee[b], Jung-Ho Yu[a,*]

[a] *Department of Architecture Engineering, Kwangwoon University, South Korea*
[b] *ICEE, Seoul National University, South Korea*

ARTICLE INFO

Keywords:
Green Building Material Information (GBMI)
Information collection
Information classification
Web-crawling
Ontology

ABSTRACT

Various green building certifications have been discussed as a part of efforts to realize sustainable development. In some countries, it is mandatory to acquire certifications for buildings above a certain scale. As a result, the demand for green building certifications has increased. Various studies have been conducted on the efficient performance of green building certification tasks.

To improve the tasks for material information management, the following problems should be addressed: 1) Evaluation of material selection is difficult because of limitations on the amount and quality of the collected information; 2) Unnecessary duplication of work occurs because the important information created at each stage of a project is not delivered efficiently to the next step; 3) Information management for material information, which requires continuous updating, is not sufficient.

Therefore, this study proposes an automated process of collecting and classifying Green Building Material Information (GBMI) using "web crawling" and "ontology" to improve the work efficiency of material information management. The proposed process is verified for interior finishing materials, which are a part of green building certification tasks. The proposed process can reduce the time required for the information management of building materials and eliminate human errors.

1. Introduction

Environmentally Sound and Sustainable Development (ESSD) has become a hot topic since the Rio Declaration of 1992. In the Rio Declaration on Environment and Development, sustainable development should be discussed around humans and humans should enjoy healthy and productive lives harmonized with nature (UN, 1992). Accordingly, various efforts were made to realize sustainable development in construction, and environmental certification systems such as LEED of the United States and BREEAM of the United Kingdom were established. In Korea, the Green Building Certification System (G-SEED) was newly implemented in 2003 by integrating the green building certification system with the housing performance rating system.

Certification systems in each country have different priorities depending on their situation and their environment. However, the fundamental purpose of the system is to reduce energy and build sustainable buildings, so certification items of energy, materials, and resources are common factors. Also, items for indoor environments are included as common factors of systems. This is probably due to increased interest in the indoor environment, as the causes of directly threatening users' health, such as the new home syndrome have been identified. Therefore, it is recommended to use materials applied indoors as green building products and the demand for green building certifications is increasing.

The works of green certification are characterized by the need to perform simple and repetitive tasks as many times as design changes. The simple tasks such as calculating the used material area and collecting information related to green building certification of applied materials are required as design change. Thus, the efficiency of certification evaluation work is reduced.

Especially, construction materials are the essential elements for carrying out the project. Information for construction materials should be managed, such as delivery time, quantity, quality, form, color, and price [17]. Due to the increased demand for green building certification, it is necessary to manage information, such as green material certification, test certification, expiration dates of certifications, etc. Therefore, if construction material information is continuously managed along with the project process, it is expected that more effective project execution is possible. Information management of green building materials for obtaining a green material certification requires

* Corresponding author.
E-mail addresses: hshgl9786@kw.ac.kr (S.-H. Hong), sklee1128@snu.ac.kr (S.-K. Lee), myazure@kw.ac.kr (J.-H. Yu).

https://doi.org/10.1016/j.autcon.2019.01.015
Received 13 April 2018; Received in revised form 18 January 2019; Accepted 22 January 2019
Available online 05 March 2019
0926-5805/

general material information and Green Building Material Information (GBMI) for the entire construction period. As the demand for certificating green buildings has increased, the need for getting certification has also increased.

Recently, various studies have been conducted to develop an integrated project process by applying Information Technology (IT). The application of IT improves the coordination and collaboration among stakeholders in fragmented construction projects [19]. IT provides a better environment for project management. The Internet, an IT, provides a collaborative environment, enabling more effective execution for accomplishing a project [4,6]. Various studies to develop a material information management system have been done, but it has been difficult to operate the system because it does not reflect the material information that needs continuous updating.

Also, there is a catalog format [7] using IT technology of Korea; however, GBMI on the green-certification works currently exists in various types of documents that are distributed or managed on the personal computers of workers. Designers have to spend a lot of time and effort in the material selection stage. There is difficulty in evaluating candidate materials because of the limited amount and quality of information [18]. Also, important information for each project step is not transferred to the next step. This results in unnecessary duplication of work with lost information [28]. To improve this problem, the development of an integrated system for managing information throughout the project is ongoing, but it is still difficult to apply it [12]. Also, the traditional catalog-based method, which is built in Korea, is not updated continuously and is not utilized [12].

Also, efforts to using BIM into various applications such as support for green building certification are actively carried out. Information generated from the BIM design environment can be integrated management and used throughout the construction process. However, it is difficult to obtain the latest information periodically in the traditional catalog-based method, because separate efforts to upload new information must be preceded to maintain the newest product information. Moreover, additional work is required to modify the material name and product name to use the collected material information for BIM design and so on.

Therefore, this study suggests an automated process for managing the construction material information of green certificated products by utilizing web crawling and an ontology to improve the collection and classification steps. In this research, we used Anaconda 4.1.1. as the Python distribution and Protégé v5.2.0 for building an ontology. We propose an ontological approach that enables automated information collection of interior wall finishing materials, ceiling materials, and flooring from four product manufacturers and classifies them into ontologies.

We discuss the trends of research and element technologies for GBMI management in Chapter 2. We then propose an automated process for the management of GBMI using web crawling and ontology in Chapter 3. In Chapter 4, we present a case study that confirms that the proposed process can work in real-world situations.

The proposed process periodically extracts product information from a manufacturer's webpage and automatically recognizes the information for a standard classification system with a standardized product name. It is expected that the efficiency of work can be improved by reducing the duplicated work.

2. Preliminary studies

2.1. *Current practice of Green Building Material Information (GBMI) management*

A green building uses environmentally responsible and resource-efficient models throughout the building's lifecycle, from planning to demolition. Green building certification is a system that evaluates the environmental impact of a building, for example, Leadership in Energy and Environmental Design (LEED) in the United States, BRE Environmental Assessment Method (BREEAM) in the UK, Green Standard for Energy and Environmental Design (G-SEED) in Korea, and the Comprehensive Assessment System for Built Environment Efficiency (CASBEE) in Japan, etc. The main items to be assessed in a green building certification are energy and materials. For the management of harmful substances in an indoor environment, items related to the indoor environment are also included.

It is essential to use construction materials which have a green material certification. Emitted harmful substances are assessed for obtaining the scores of indoor environmental categories. There is an increasing usage of green building materials for the prevention of house sickness syndrome and health care for seniors. Green building material properties include the material codes, material names, product codes, standards, certification names, expiration dates, certificate documentation, manufacturer names, and company contacts. Product names and certificate documentation are necessary for green building certification.

The green building material information management process for certifying a green building is divided into the information collection step, information classification step, and information storage step (see Fig. 1).

2.1.1. *Information collection*

A worker manually collects information by calling or emailing a product manufacturer, using a paper catalog or a web page, etc. Before acquiring material information on the Internet, the worker collects information directly from the manufacturer or obtains information through a paper catalog, provided to promote products. The release of information on the Internet is a common method of providing information because new information is periodically generated according to product development.

2.1.2. *Information classification*

The collected information is identified according to the material classification system. Classification refers to the act of grouping information with similar properties and determining a name that can represent the group name, based on a standard classification system. This study uses the revised Commodity Classification System of Public Procurement (CCSPP) as the standard classification system, which is based on the United Nations Standard Products and Services Classification (UNSPSC). For example, a product corresponding to a

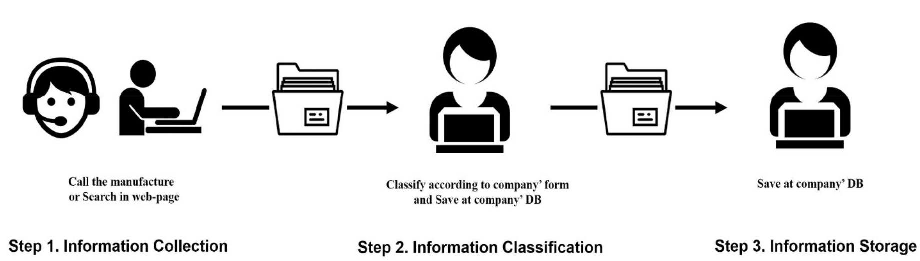

Fig. 1. General process of green building material information management.

Table 1
Limitation of existing management method.

Type	Process		Limitation
	Information collection	Information classification	
Tool based BIM	To extract information form BIM model or to find the information manually	To classify according to BIM-IFC standard	It needs to develop the library of each object for managing the GBMI information.
Standard form	To use the form set the sort of information	–	If information input is not enforced, the amount of information is small and if the form changes, it need to request information again
Database by manufactures	To call the manufacture or to find the information manually	To classify according to worker's standard	There is insufficient management of information expiration date and depending on who builds the database, the classification of information is different.

wallpaper as the interior wall finishing material and a product corresponding to ceiling panels are classified as a ceiling material. A product corresponding to wood flooring is classified as flooring. Because the information classification in the existing process does not have a standard system for classification, the classification was performed on a different basis, according to workers. Therefore, there is information that is missing because it is not classified, making it difficult to share information.

2.1.3. Information storage

The classified information is stored in a database. Information storage in the existing process was saved either on a worker's personal computer or in a company-specific database. Since there is no integrated storage system that can manage information, the performed work may be duplicated or the information cannot be used beyond a validity period. Accordingly, an integrated database for managing construction material information needs to be continuously updated.

However, information management using web pages does not provide comprehensive information about materials, such as quality, construction ability, etc. For comprehensive information, it is difficult to get the necessary information even with a lot of time and effort. In addition, some information is missing in web pages and is not well utilized because the information is only occasionally updated [12]. Accordingly, a worker has to spend a lot of time in the information collection step to obtain the needed information. Because the information is provided from various sources, the quality and quantity of the information depend on the skill level of the worker.

2.2. Previous studies on construction material information management

The main purpose of this study is to systematically manage the information generated by designers, construction engineers, and green certification consultants during a project. The types of management methods for managing material information are tool-based Building Information Modeling (BIM), standard forms (i.e. green construction material information system in Korea, GreenSpec in the U.K., Sweets Catalog in the United States), and manufacturers' databases.

First, tool-based BIM is a tool for the integrated management of information using building models. Salman et al. [22] analyzed articles that used the LEED system and developed a system that applied BIM to information management. Bae [2] presented guidelines for supporting the green building certification design process by analyzing green BIM according to the level of development (LOD). Kim [12] analyzed the green building certification process for sustainable construction and presented an integrated design process, in which designers participate in each stage, from the beginning of a project. Wei et al. [27] analyzed the literature and the best performing project for applying green BIM and proposed a business process map that can be used to implement a LEED project. This tool needs the building libraries of all objects for effective information management because BIM collects and manages information based on object libraries.

Second, standard forms are a relatively common information management tool. Standard forms can acquire the required information easily because a form is defined before the information collection step. For an example in Korea, the green construction material information system [7] of the Korea Environmental Industry & Technology Institute (KEITI) is a one-catalog-type site. It spreads information on the applications of green building materials for design and procurement. The KEITI page provides information and documentation, such as product names, product functions, environmental mark certifications, carbon certifications, field applications, specifications, and manufacturer information, etc. In the U.K., GreenSpec is a one-catalog-type site [8], which was launched in 2003 with government funds. This site provides information on green buildings. GreenSpec is designed to promote the development of sustainable building products, materials, and construction technologies. It also provides information on green products and building design for sustainable building design guidelines. In the United States, the Sweets Catalog [24] is a one-catalog-type site, which provides comprehensive information about products, CAD details, BIM objects, product catalogs, galleries, green building product information, etc. It is a web-based database and uses the MasterFormat as the standard classification system. This tool is limited to a small amount of collected information.

Finally, a manufacturer's database can provide information for users through web pages. Ryu [21] proposed a new green building information system to reorganize and improve the information system using the Internet. Kim [11] analyzed the present situation of domestic and overseas construction material information systems and presented an improved user model by finding the required information and the utilization method.

The limitations of existing management methods for green building material information management are shown in Table 1.

This study suggests that web crawling and an ontology improve the construction material information management process. When applying web crawling technology, the same quantity and quality of information can be collected from the same source, irrespective of the skill of a worker. The collection process is automatically performed by the same process, so information can be collected with the same working time. When applying ontology technology, it is possible to classify the information consistently, regardless of workers. This enables the effective sharing and management of information, classified on the same basis.

2.3. Web crawling in construction

Web crawling is defined as the act of collecting information on the Internet, starting with a SEED URL (Uniform Resource Locator). A program that performs crawling is called a web crawler. It is divided into a general web crawler that collects information along a hyperlink starting from a seed URL, and a distributed web crawler that collects information from a server and client environment [23]. Web crawling, which is a commonly used concept, is performed through a standard web crawler. Crawling is the act of collecting regular information, such as text, and is also referred to as web scraping.

We interviewed five experts in construction management with an average of more than 10 years of experience and found improvements for applying web crawling with automatic information collection. We suggest improvements, such as the standardization of classification codes, grouping with the control of construction management, standard information collection forms, management of product performance information, manual input sections that are not automatically collected, etc. As needed improvements, “standard information collection forms for ensuring information reliability” and “standardization of classification codes” are suggested.

2.3.1. Standard information collection forms for ensuring information reliability

Since the information provided on the Internet has various forms and is inaccurate because of unknown sources, it can be improved by standardizing the collection routes and storage forms for consistency and reliability. A reliable web page was designated as a Seed URL, and web crawling was performed for the required information for green building certification work.

2.3.2. Standardization of classification

It is an improvement to have standardization of material codes. A lot of money and labor must be invested for developing and establishing a new classification. In this study, the Commodity Classification System of Public Procurement (CCSPP), which is based on the United Nations Standards Product and Services Classification (UNSPSC), was adopted as the standard classification system to lower investment and labor costs.

2.4. Ontology in construction

Ontology refers to structured information that a computer understands, such as the data and information about daily life, in a certain form [5]. This is also called computational ontology, which refers to the metadata of a specific area, expressed by semantics that can be interpreted by a computer [16]. To build a semantic system, we need an ontology model written in a language that the software understands. We use the Resource Description Framework (RDF) and Web Ontology Language (OWL) as a programming language [5].

The purpose of using an ontology is summarized in four main areas: semantic interoperability, standardization, communication, and knowledge management [16]. The technology of an ontology is expected to standardize the information provided in the automation process and facilitate knowledge management. In the literature, the properties and classes of an ontology suggest information standardization structures and knowledge management structures that define the relations among classes and properties.

Tserng et al. [25] proposed an ontology that constructs a risk management framework by analyzing the risk factors that occur in a project and estimating the cause of a risk by weighting factors. Lee [14] proposed an ontology by the consistent reasoning of standard items of tile construction with inadequate design information and the subjective judgment of workers. Hexu et al. [9] proposed a quantity calculation ontology for more accurate extraction of complex quantity information of objects through the BIM model. As an extension of the use of BIM technology, the ifcOWL ontology, which combines Industry Foundation Classes (IFC) and ontology technology, is utilized. The ifcOWL ontology is a representation of IFC, which are a standard for building and construction data based on BIM, by OWL, an ontology language, and has the same structure as IFC EXPRESS and XSD schema [3]. The ifcOWL ontology is expected to utilize data such as linkages of building data, by interworking with GIS information and product data, by utilizing semantic web-based technology.

As a result of reviewing the existing literature, an ontology is used to infer new knowledge by the combining and reasoning of extracted information. An ontology should be constructed according to the purpose of use. In the ontology construction process, a problem is first analyzed to propose an improvement by applying the ontology. Next, the process structure is defined for establishing the ontology. For constructing an ontology, the structure of an ontology consists of classes, subclasses, object properties, and instances for constructing the ontology. Finally, the ontology structure is implemented using an ontology support program. The operation of the ontology is verified by entering actual values. We also applied an ontology to a case study to verify whether the ontology based on this process is an improvement, compared to the existing process.

Because the main purpose of this study is to systematically manage green building material information, we need to automatically recognize the same meanings with differently expressed information. This information is generated by designers, construction engineers, and green certification consultants. By constructing a green building material ontology that can classify information according to the standard classification system, it is expected that users can access the data needed for green building certification using the ontology and manage the data in the same classification system. Also, in terms of knowledge management, the problem of performing unnecessary repetitive tasks due to missing information is also expected to be improved.

3. Proposed process for automated management of GBMI using web crawling and ontology

3.1. Automated collection and classification

The proposed automation process uses a web crawler to automatically collect the required information and an ontology to automatically collect information for GBMI management. It is possible to consistently maintain the information quality as this can vary depending on the skill level of a worker, to reduce redundant tasks.

Web crawling can avoid adding unspecified and inaccurate information because the location of the information a user needs is specified. The standard form can be convenient for using information. However, for providing information, input tasks can be imposed as an additional single task. The duplication of effort (tasks) should be minimized. In this study, for non-standardized material information, the standard material name was deduced from the collected information to ensure the consistency of information by using the ontology (see Fig. 2).

It is possible to specify the web page or the web catalog of each manufacture which is frequently used for collecting information. In web crawling, the code should be applied to each web page or each web catalog to get information. Maintenance is relatively easy because the HTML structure does not frequently change. Similarly, in the ontology, the material names of standards can be inferred with effort, but the maintenance is relatively easy because it has more flexibility than the existing databases. A comparison of the proposed process in this research and existing processes is shown in Fig. 3.

Existing methods of information management are (A) calling manufacturers or searching for their web page and (B) searching standard web pages like the Green construction materials Information System in Korea or the Sweets Catalog in the United States.

The first method saves no working time because workers must manually search for information on each manufacturer's web page or call them individually; however, the other can reduce working time. An advantage of this method is that it can collect information quickly from standard web pages; however, it can only handle a limited amount of information. If an insufficient amount of information is found, then additional time must be spent gathering more.

By contrast, our method of using web crawling and ontology effectively saves working time because it gathers information in a standard form by designating web pages of each manufacturer that the user wants. This method does not needlessly repeat work, because it collects information from a web page that contains the targeted information. From the developer's point of view, this method may offer no

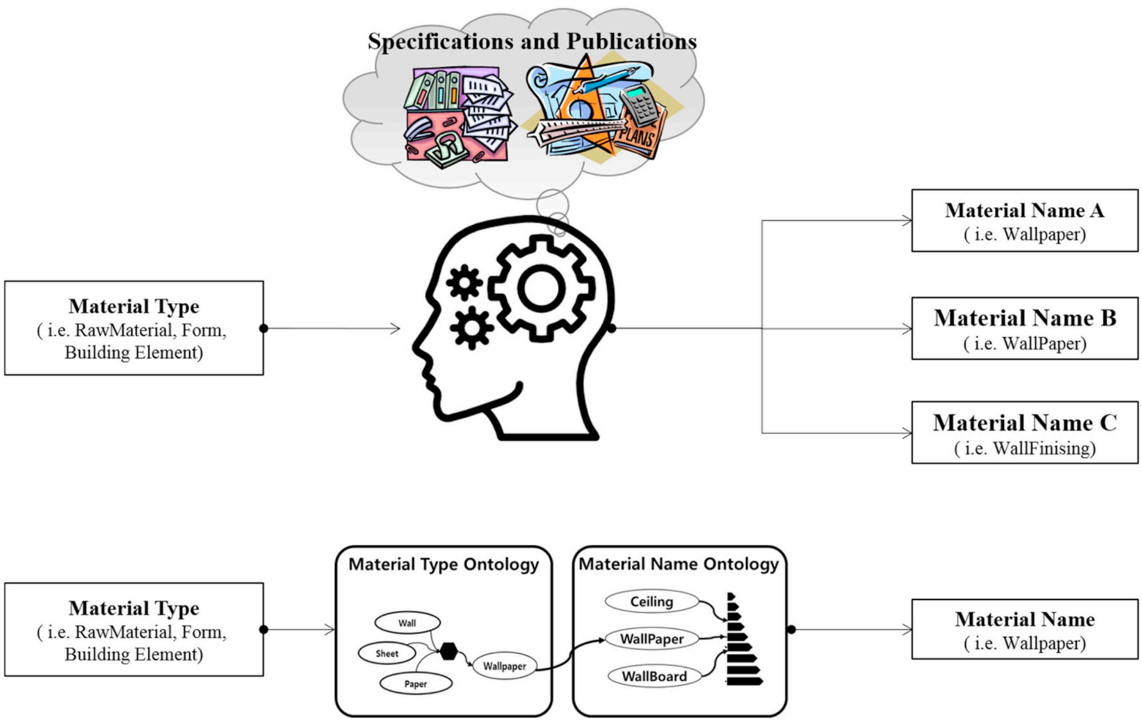

Fig. 2. Concept of ontology inference.

advantages regarding saving time, because the code of each website needs to be modified. However, from the user's point of view, it can save working time and improve the amount and quality of information (Table 2).

Advantages of process for automated management of GBMI using web crawling and ontology are expected to be:

GBMI can be managed in the desired form. The traditional catalog-based method obtains information from a database that is stacked prior to users accessing a specific site and entering information into a given form. However, when using Web-Crawling and Ontology, it

Table 2
Comparison of existing methods and our method.

	Time saving	Quality increase
Existing method A	X	△
Existing method B	O	△
Our method	O	O

A: Calling manufacturers or searching for their web page, B: Searching standard web pages.

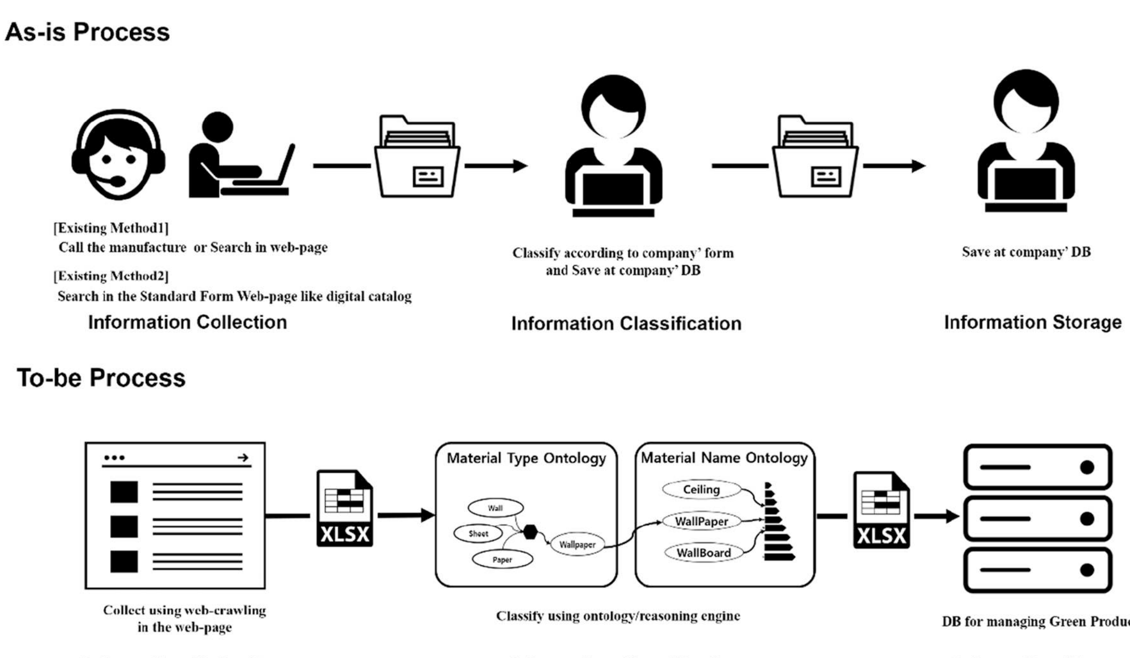

Fig. 3. Comparison of the as-is process and to-be process.

is possible to manage GBMI on websites that are not organized into a catalog because the prior process of entering information into a given form is automated. In addition, the validity of the products applied in the preliminary certification in the design phase expires in this certification in the construction-after phase and changes the products. However, the information collected through the process is managed by the DB, which reduces the inconvenience of having to change the product after the expiration date and allows the DB to be maintained with accurate and recently information.

GMBI can be automatically updated. GMBI should also be updated continuously because new products are continuously created due to the feature of building materials. However, the traditional catalog-based method continues to allow information updates if someone enters new information. Using the process for automated management of GBMI using web crawling and ontology, it is possible to automatically update information by setting up a period. In addition, it can contribute to improving the accuracy of information by reducing the human error caused by human intervention in the information management process.

GBMI can be compatible with BIM. GBMI collected by process for automated management of GBMI using web-crawling and ontology can be renamed to the material or product name entered in the BIM model using Ontology, so information can be compatible with the BIM model without further effort. In addition, the process presented in this study is applicable in such a way that GBMI is automatically selected in the GBMI DB collected for each material, so it is easier to compare product for alternatives.

3.2. Web crawling for automated information collection

The process of information collection through web crawling is shown in Fig. 4. The process of information collection consists of information analysis and web crawling. The information analysis prepares the information that is collected by web crawling. The phase of web crawling is the step of actually collected information using a web crawler. A web crawler identifies the HTML structure of a web page to collect information. The result of this phase is raw data of GBMI. The collected raw data of GBMI is classified using ontology into a data table.

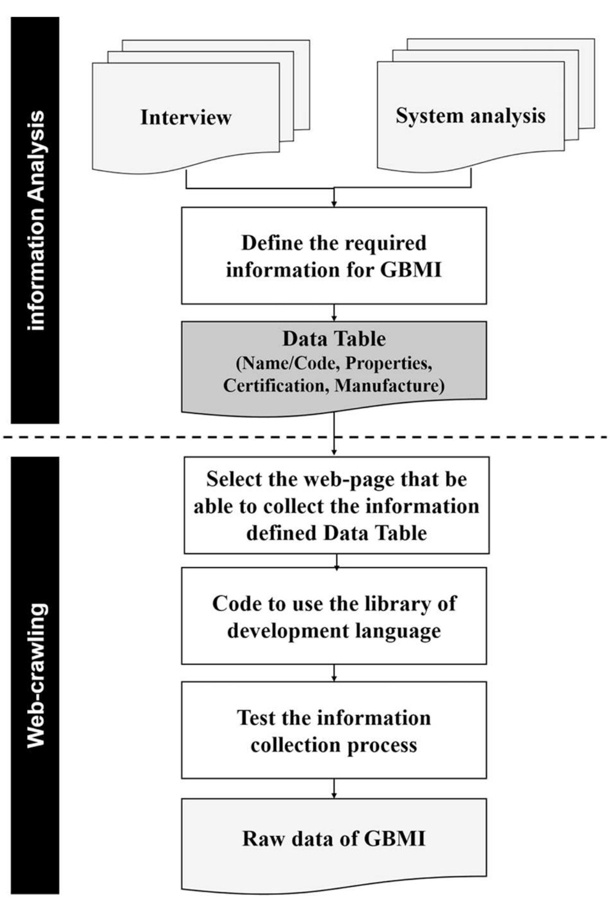

Fig. 4. Information collection process using web crawling.

3.2.1. Information analysis

This step analyzes the required information through interviews of workers that are preparing for green building certification and the analysis of the green building certification system and material information system. As a result, the required information consists of Name/Code (A: Material, B: Product), Properties (C: Element, D: Raw Material, E: Use, F: Form, G: Other), Certification (H: Name, I: Certification authority, J: Expiration date, K: Certificated document), and Manufacture (L: Name, M: Web-page, N: Tel.) (see Table 3).

3.2.2. Web crawling

To run a web crawler, the first step is to understand the HTML structure of the web page containing the information of the selected product group and to understand the HMTL syntax of the information to be collected. The manufacturer's web page that a user selected to collect product information determines the SEED URL to start web crawling. The web crawler sequentially analyzes the pages from the SEED URL for collecting information. The web crawler needs to determine the HTML structures to move from the SEED URL to web pages where information is collected. For example, the HTML structures that are used to move from the SEED URL to the web pages where information is collected include “ < a href > < img src:URL > < /a > ” that contains an image in < a > to represent an icon that goes to the next page. The HTML structures that contain the information include “ < td > information < /td > ” and “ < tr > information < /tr > ” that represent a table.

The next step is the coding for web crawling, using a Python library, such as Selenium, BeautifulSoup, and Pandas. For example, the function of Selenium and Beautiful Soup is collecting information on the Internet. The function of Pandas is documenting the information. In this research, the information collecting process is implemented by selecting Python as the development language. Python can easily understand the meaning of language and can be easily implemented by beginners.

3.3. Ontology for automated information classification

This study constructed the ontology on the basis of elements that define the material properties and that constitute the material classification system. The system automatically recognizes the information obtained from the web page of a manufacturer through web crawling. To enable such an automated system, the Green Building Material Type Ontology (GBMTO), consisting of the elements that define the material properties, and the Green Building Material Name Ontology (GBMNO), composed of the material classification system, should be established in advance.

In this research, the GBMTO, which is the material type ontology of interior materials, and the GBMNO, which is the material name ontology of interior materials, are used. The framework of GBMTO and GBMNO is shown in Fig. 5.

A continuous line (subclassof) expresses the hierarchical relationship between a class and a sub-class, and a dotted line (ObjectProperty and DatatypeProperty) expresses the relationship among classes and the relationship between a class and a data value.

The process of information classification using GBMTO and GBMNO is shown in Fig. 6.

Information that was web crawled from a manufacturer's web page is recognized as one of the material types through GBMTO, which

Table 3
Collected information from existing material information system.

Type		Name (code)		Properties					Certification				Manufacture		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N
Catalog type	Korea	O	O	O	O	O	O	O	O	O	O	O	O	O	O
	U.K.	O	O	O			O					△	O	O	O
	U.S.A.	O	O	O			O					△	O	O	O
Board type	K Co.	O	O	O			O	O	O	O	O	O	O	O	O
	L Co.	O	O	O		O	O	O	O	O	O	O	O	O	O
	A Co.	O	O	O			O						O	O	△
	T Co.	O	O	O			O	O					O	O	△

defines the material type. GBMNO recommends the standardized material name for the recognized material type. First, the collected GBMI is transformed by the RDF Data through the RDF transformation engine. Second, in the semantic reasoning system, the transformed data is automatically recognized by the instance of the subclass corresponding to the class defined in the GBMTO. This class is 'Material Type_1' with inferred information of the subclass of the material type class as a necessary and sufficient condition. The 'Material Type_1' class recommends a standard material name that has an instance of a subclass of the GBMNO. Finally, GBMI with the standard material name is classified according to the standard classification system.

3.3.1. Definition of classes

The class of GBMTO is an ontology for inferring the material name using the collected material information in the web crawling step. This class is composed of the material properties. To define the GBMTO, this study uses the subdivision of procurement commodity classification system (CCSPP) to represent the characteristics of each material [20]. Also, it refers to various web pages. On this basis, this study checked the minimum amount of provided information and most of the information related the product characteristics on manufacturers' web pages.

GBMTO contains six classes: *Element*, *RawMaterial*, *Form*, *Certification*, *Manufacturer*, *and MaterialType*. The '*MaterialType*' class contains semantic reasoning rules which are defined by combining *Element*, *RawMaterial*, *and Form* of GBMTO. The classes of the proposed ontologies are described in Table 4.

The class GBMNO is used to classify material information deduced from the standard material name of CCSPP. CCSPP has four step hierarchies in classification codes and provides an explanation of an item as basic information, explaining the characteristics of classification. Therefore, the GBMNO structure follows the same hierarchy as CCSPP.

GBMNO contains one class: *MaterialName*. The '*MaterialName*' class contains sub-classes for the material hierarchy used in public procurement. Instances of the lowest class are a subdivision of the material hierarchy, such as wall, ceiling, and flooring, which are standard classification names.

3.3.2. Definition of property

OWL properties consist of two types: *owl:ObjectProperty* and *owl:DatatypeProperty* [14]. The object property (owl:ObjectProperty) is used to express the relationship among objects, and the data-type property (owl:DatatypeProperty) is used to express the relationship between an object and a data value, such as a numerical value [14]. Properties of the proposed ontologies are shown in Table 5.

4. Verification of the proposed process

4.1. Overview

The automation process was applied to the manufacturing company web page to verify whether the information was collected and that the classification of the information was automated. To verify the process, four manufacturing companies that provided information on interior wall finishing materials, ceiling materials, and flooring were selected,

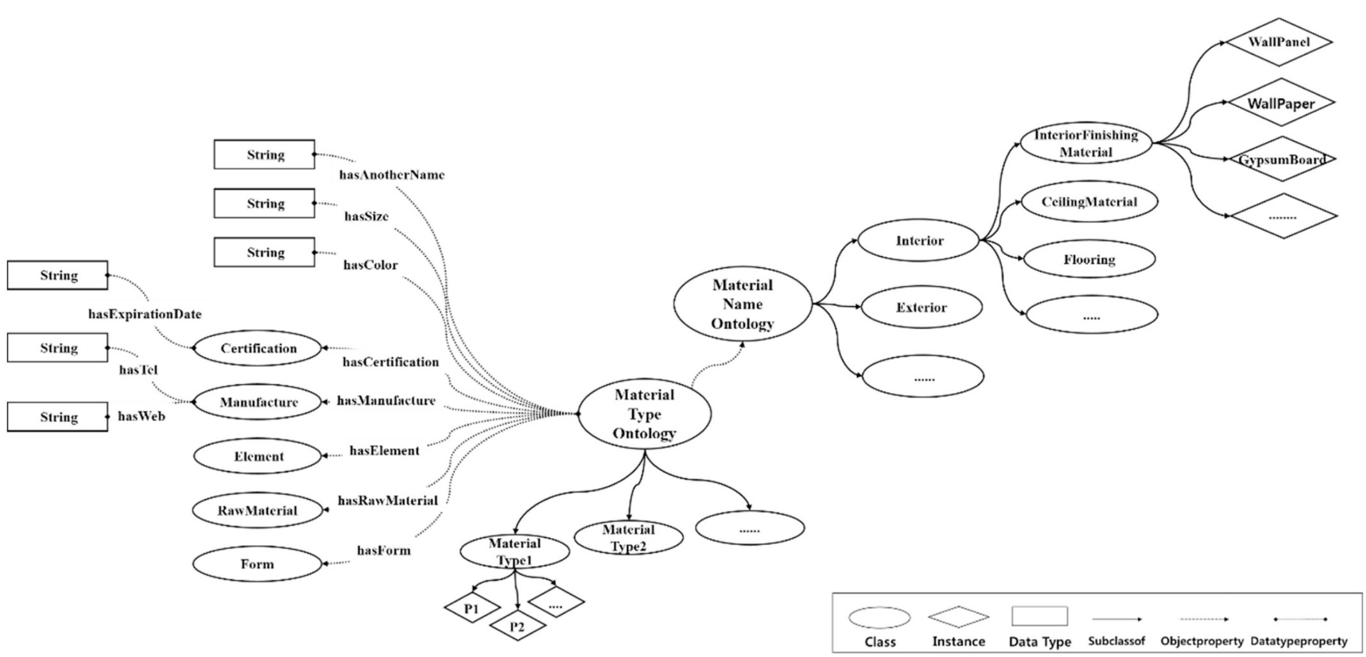

Fig. 5. The structure of green building material ontology.

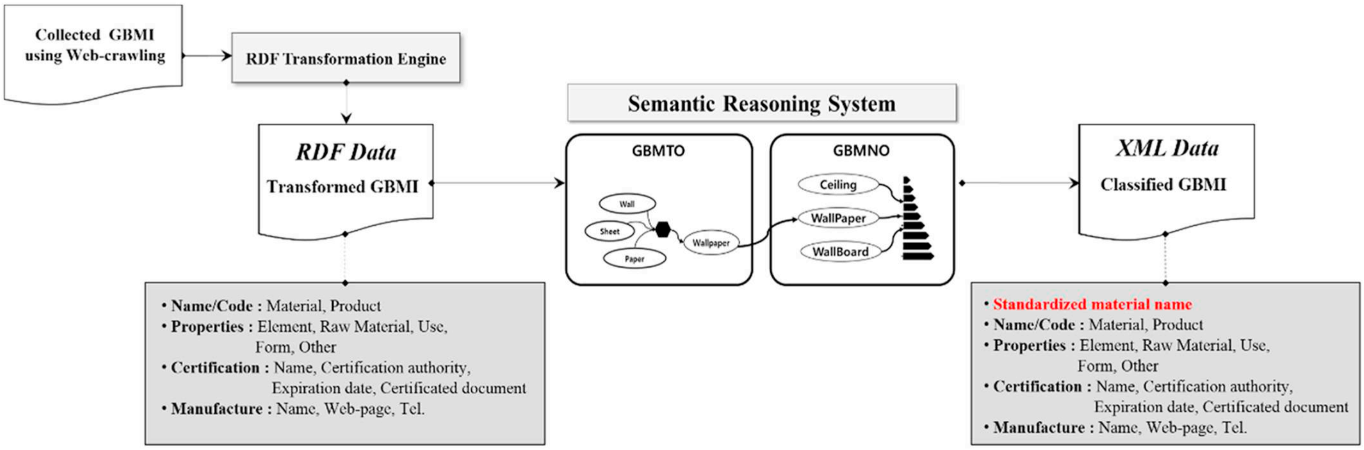

Fig. 6. Information classification process using ontology.

and the automation process was performed. For web crawling, this study uses Anaconda version 4.1.1, which is a Python distribution. For the ontology, this study uses Protégé version 5.2.0, based on the defined classes and properties. For verifying consistency, this study uses HermiT Reasoner v.1.3.8.413 and DL Query, built in Protégé v5.2.0.

4.2. Information collection using web crawling

4.2.1. Information analysis

The required items are divided into the name (code), properties, certification, and manufacture. The categories of names are material code, material name, product code, and product name. The categories of properties are the product standards defined in CCSPP: element, raw material, form, color, size, and a material name. These may be called by other names in other systems. The categories of certification are the certification name, the expiration date of certification, and the certificate documentation. The categories of manufacturing are manufacturer name, manufacturer web page, and manufacturer Tel. The information categories for getting the green building certification are "Product Name", "the expiration date of certification", and "certificate documentation". The information categories for supporting the information classification are "the categories of names". These categories use "the categories of properties" that use the GBMTO class. The information categories for providing the other links to collect the information are "the categories of manufacturing".

4.2.2. Web crawling

For verifying the information collection process, in this study, the interior finishing materials were selected as the range of GBMI. The web crawler selected web pages to collect information [1,10,15,26]. For a catalog web page, the web crawler is already provided the standard form for collecting information [13]. It also selected the web pages which are not provided the standard form like the individual manufacturer's web pages to help develop an enhanced web crawler. The selected web pages are K Co., L Co., A Co., T Co, which confirms the information provided in Table 3.

For developing an enhanced web crawler, first, the HTML structure of the web page containing the information of the selected product group was identified by using the developer tool in Google Chrome. The web crawling was performed by separating the product information and the certification. Second, the automated process runs commands from the Python library. It blocked unnecessary image information and collected information using the Chrome driver and the Python library. The automated process can collect information, such as the material name, product name, type (size, weight), construction method, packing unit, unit price, usage, color, name of certification, organization of certification, certification expiration date, certification, manufacturer's name, and manufacturer's web page. Finally, the automated process develops a prototype web crawler by organizing the product information codes with a program written in Python (see Fig. 7). Product information is collected sequentially from the web pages of each manufacturer and saved as csv files. When all the pages have been collected, each file is converted to an xls. file, and the web crawler automatically deletes the csv files generated during the information collecting process.

Table 5
Properties of green building material ontology.

Type	Property	Domain	Range
Object property	Has Element	Material type	Element
	Has raw material	Material type	Raw material
	Has form	Material type	Form
	Has certification	Material type	Certification
	Has manufacture	Material type	Manufacture
	Has material name	Material type	Material name
Property	Has another name	Material type	String
	Has color	Material type	String
	Has size	Material type	String
	Has expiration date	Material type	String
	Has tel	Manufacture	String
	Has web	Manufacture	String

Table 4
Classes of green building material ontology.

Class	Description	Instance
Element	Part of building which installed green building material	Wall, ceiling, floor and so on
RawMaterial	Raw material of green building material	Chemical, paper, gypsum, wood, marble and so on
Form	Form of green building material	Sheet, panel and so on
MaterialType	Combination of the green building material properties	Define separately based on the classification
Material name	Material name of standard classification system	WallPaper, WallPanels, GypsumBoard, CornerGuards and so on
Certification	Certificate authorities	Keiti and so on
Manufacture	Product manufacture	LGHausys, KCC, Tuscan and so on

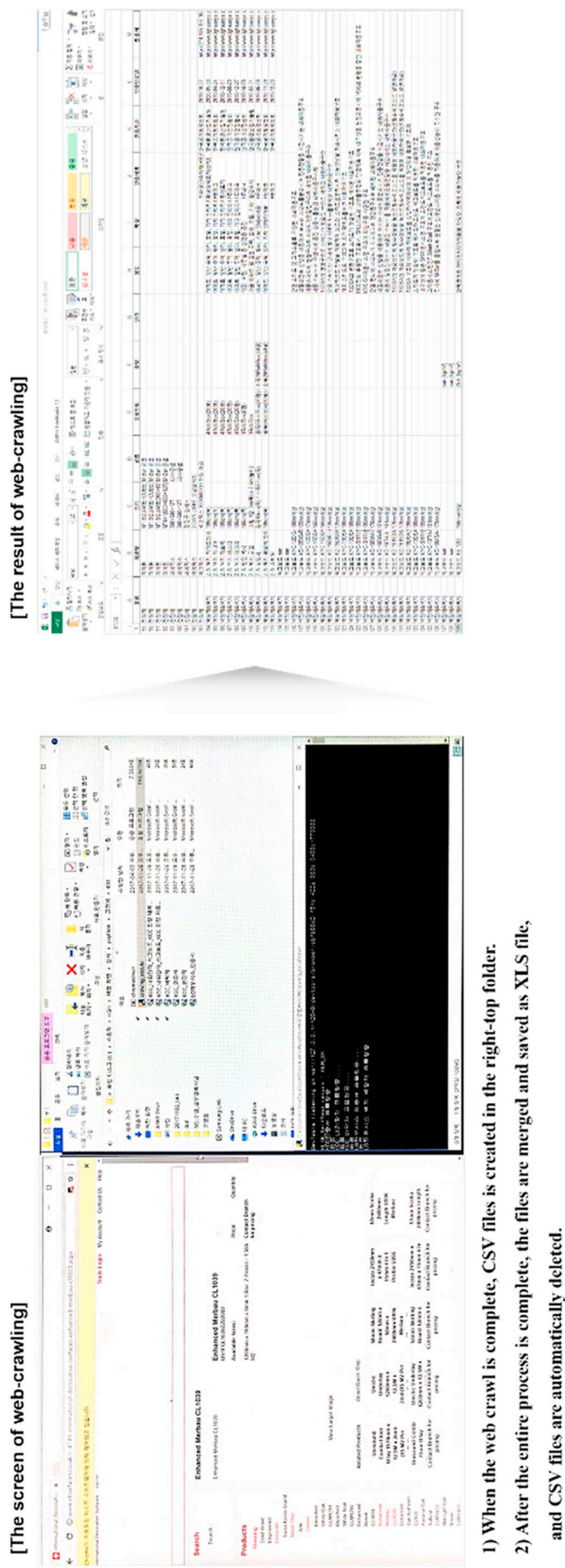

Fig. 7. Running screen of web crawling.

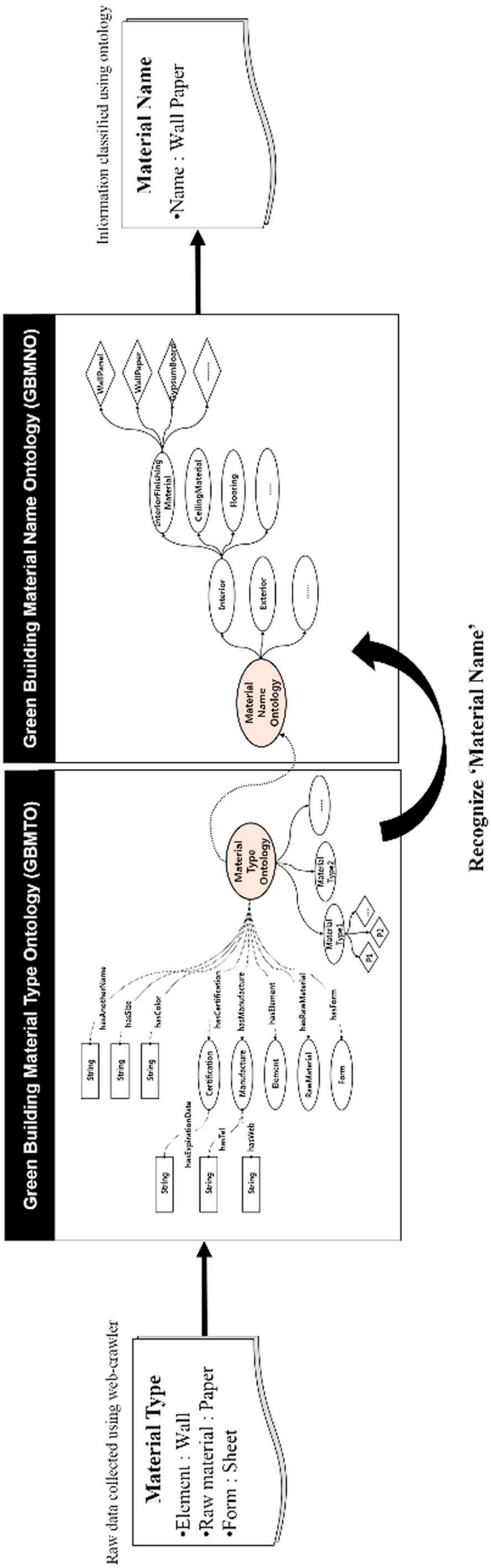

Fig. 8. Process of semantic reasoning for a material name.

4.3. Information classification using ontology

The information, based on the product properties, is automatically classified using an ontology. The element, raw material, and form information, parsed by crawling a manufacturer's web page, are recognized as a material type through the GBMTO, which defines the material types. Material names, suitable as a material type, are recommended through the defined GBMNO.

For example, information such as 'wall, paper, and sheet' is parsed and automatically recognized by instances of subclasses corresponding to Element, Location, Raw Material, and Form classes. Material Type_1, which is a subclass of the GBMTO, has a "Necessary" that 'the element is a wall, the location is a load-bearing wall, the raw material is a paper, and the form is a sheet'. This 'Material Type_1' class has 'Wallpaper', which is an instance of the subclass of the GBMTO, as the most suitable material name. As a result, the standard material name is recommended (see Fig. 8).

4.3.1. Semantic reasoning

The semantic reasoning rules for the automated inference of product codes are as follows.

- **GBMTO_1** ≡
  - Necessary & Sufficient
  - ∋ hasElement has ***wall***
  - ∋ hasRawMaterial has ***Paper***
  - ∋ hasForm has ***sheet***
  - Necessary
  - ∋ hasGreenBuildingMaterialName has ***Wallpaper***
- **GBMTO_2** ≡
  - Necessary & Sufficient
  - ∋ hasElement has ***wall***
  - ∋ hasRawMaterial has ***Paper***
  - ∋ hasForm has ***sheet***
  - Necessary
  - ∋ hasGreenBuildingMaterialName has ***Wallpaper***

The reasoning of the ontology is based on HermiT Reasoner version 1.3.8.413, included in Protege v5.2.0. The reasoning result shows that the material properties and the material name are recognized, as an individual product, defined as an instance of the material type, as shown in Fig. 9. The result of the GBMI obtained through the proposed automation process is shown Table 6.

4.3.2. Query

In addition, the reasoning results for wallpaper, ceiling panels, and wood flooring as classified in the example were searched by DL Query to confirm consistency, as shown in Fig. 10. The collected information by web crawling and classified information by the ontology were matched. It was confirmed that the reasoning is consistent.

4.4. Validation

We developed the program and measured accuracy of classification to validate accuracy of proposed process. And we measured expected reduction of working time by interviewing certification worker to validate efficiency of proposed process.

4.4.1. The program development

The developed program is a BIM-based prototype program for information extraction, and utilization process is as follows. First, upload the BIM model. If any of the materials applied in the model have no information in the material database, a pop-up window will be loaded to add the required information to the material database. When you

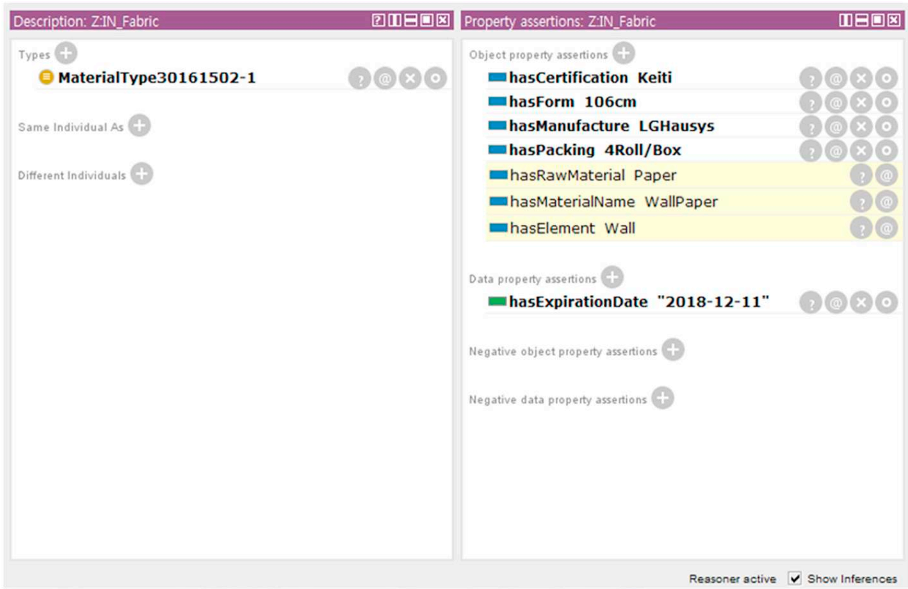

Fig. 9. Results of reasoning information using ontology.

upload a BIM model, the area information is automatically extracted. In the BIM, information about floor area, floor width, floor length, wall area, wall area applied green building product, ceiling area, window area, and ceiling length is extracted by each room (Fig. 11).

The applied product information can be checked by materials. The used product information is finishing materials, adhesives, and background material, and the application ratio is automatically calculated by comparing the wall area and wall area utilized green building product (Fig. 12).

Next, the Excel file is uploaded to the program as a result of web crawl performance. The web crawler was tested on catalog-type web pages, such as Green construction materials Information System, which currently has the most substantial amount of information on construction materials in Korea. We collected 586 items information using web crawler on the board-type web pages of four major domestic and overseas manufacturers. The uploaded file verifies the operation of the material classification function reflecting the material properties of the ontology. The uploaded file is the result of collecting information from the manufacturer of the product about the internal finishing materials, ceiling materials, and flooring materials. The program carried out information classification on 586 items.

4.4.2. The accuracy of proposed process (quality improvement of collected information)

This study validated accuracy through well-classified numbers of collected information. As a result, the accuracy was 96.8%. 567 out of 586 items were correctly classified according to the inference criteria. However, 18 items were not classified due to missing properties information defined for inference, and 1 item was not classified due to a missing product name. Thus, 96.8% could be confirmed to be classified according to ontology properties. For unclassified 4% of the information, an additional definition of properties is required to sort it. To define the ontology classification system, we referred to various manufacturers' webpages and checked the minimum amount of provided information and most of the information related the product characteristics on manufacturers' webpages. On this basis, we defined an ontology classification system. Therefore, similar accuracy can likely be obtained from other manufacturers (Fig. 13).

4.4.3. The efficiency validation of proposed process (time saving of certification work)

This study interviewed workers who are currently performing green building certification work in the field about the expected working times when using the process for each stage of the certification process. The result of interview is shown Table 7.

4.4.3.1. Step 1 selection of interview targets. We interviewed green building consulting companies that have over 5 years of experience in green building certification. The selected interview targets were the green building certification teams. They were asked about the detailed tasks that occurred during the process of green building certification using GBMI and the required time for performing each task.

4.4.3.2. Step 2 interviews of the detailed task and the required time. Based on the G-SEED in the green building certification system, this study surveyed the time required for each material task during the green building certification. The tasks are divided into 4 categories: (A) checking the applying materials, (B) checking the list of applied materials and certification, (C) selecting the applicable material list

Table 6
GBMI obtained through the proposed automated process.

Categories	Name	Format	Proposed process	Existing systems
Name/code	Material name	Text	Ontology	Manual sorting
	Product name	Text	Web crawling	Manual collecting
Properties	Product standard	Text or file	Web crawling	Manual collecting
Certification	Certification name	Text	Web crawling	Manual collecting
	Expiration date of certification	Date	Web crawling	Manual collecting
	Certificate documentation	File	Web crawling	Manual collecting
Manufacture	Manufacture name	Text	Web crawling	Manual collecting
	Manufacture web-page	Hyper-link	Web crawling	Manual collecting
	Manufacture tel.	Text	Web crawling	Manual collecting

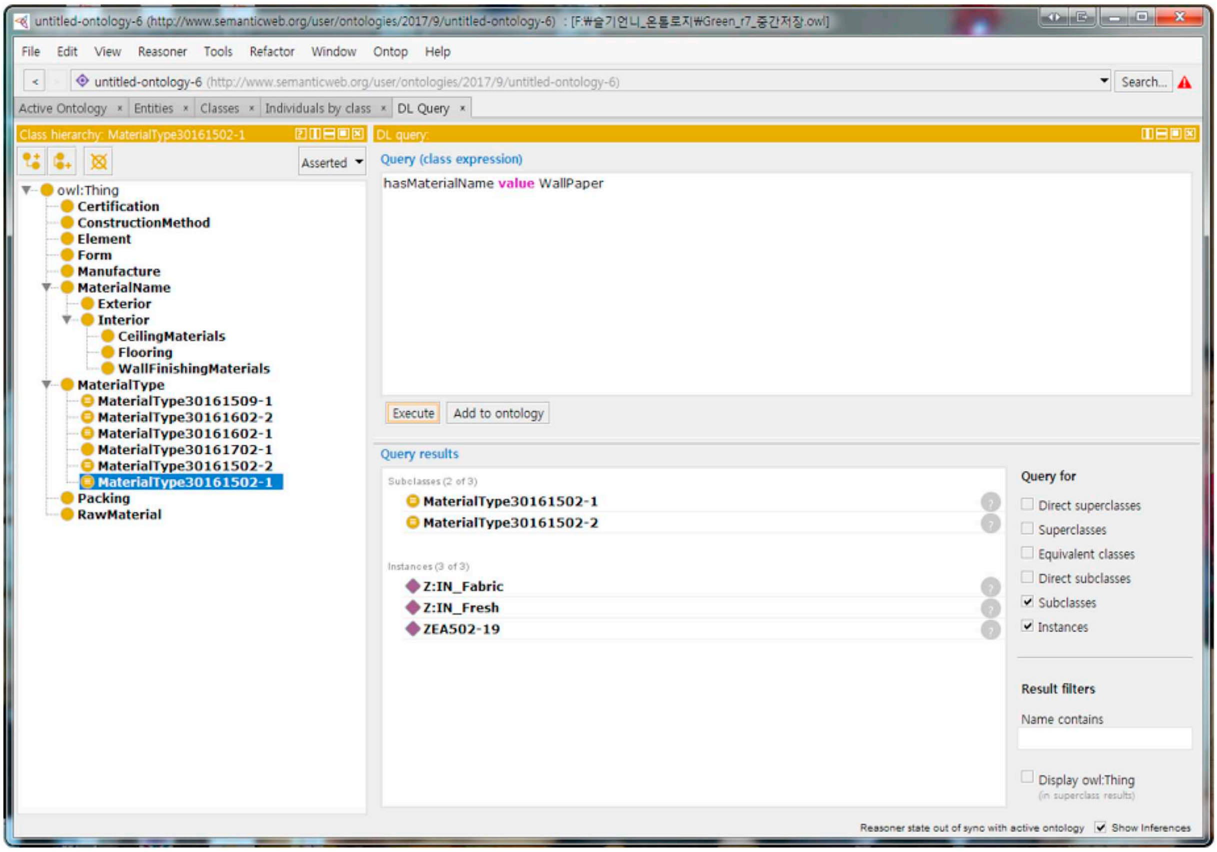

Fig. 10. Results of searching material information.

and calculating the applicable area, and (D) preparing the attached documents. This process estimated all the times like a real certification process. It surveyed the three steps of a real project: initial step, primary design changes, and secondary design changes. Then, it compared the required time for the existing process and the improved process.

4.4.3.3. Step 3 the result of an interview. According to the Green Building Certification System (GBCS) in Korea, green construction material related work averages 36.5 h for checking the applied materials, checking the list of applied materials and certification, selecting the applicable material list, calculating the applicable area, and preparing the attached documents. When the improved process is applied, it is possible to reduce the work time by about 65% for checking the list of applied materials and certification, selecting the applicable material list, and calculating the applicable area. In addition, this consistent automated process is expected to minimize human intervention, eliminate human errors, and improve the quality of work.

5. Conclusion

There is a lack of processes for integrally managing construction material information. Duplication of unnecessary work occurs with a lack of information sharing. Therefore, this study proposes an automated process for green building product information management using web crawling and ontology technology, to improve the efficiency of the information management system. If an integrated management process of construction material information is established, first, the loss of information can be prevented. Second, information can be periodically updated. Third, it is easy to change and determine products for material repair. Fourth, it is possible to manage the project information more effectively because the project owner can easily participate in the decision making of the project.

The main contents of this study are as follows.

First, web crawling technology improves the information collection step so that the quality and quantity of information are not dependent on the skill level of a worker. Second, the ontology improves the information classification step so that the classification of GBMI information is not dependent on the decisions of a worker. Third, for validating the automated process, this study interviewed green building consulting companies that have over 5 years of experience in green building certification and compared the working time of workers and the automated process in 4 categories.

The present state of the IT utilized in the management of green building certification information was examined through a literature review. The information management process is divided into the information collection step, information classification step, and information storage step. It is pointed out that the quality and quantity of information depend on the skill level of a worker in the information collection step. Information is not classified according to a standard classification in the information classification step. Therefore, web crawling technology is proposed to improve the information collection step, with ontology technology to improve the information classification step. For applying web crawling in the information collection step, this study interviewed experts on web crawling technology. For applying the ontology in the information classification step, the literature of applied ontology in construction is reviewed and an application method is suggested. Finally, this study presents a process flowchart based on the above discussion. For verification, this study collected information about interior wall finishing materials, ceiling materials, and flooring from four product manufacturers and classified them into ontologies. The Python distribution is Anaconda 4.1.1. The ontology building tool is Protégé v5.2.0, and the ontology verification tool is Reasoner v1.3.8.413 and DL Query in Protégé v5.2.0. First, 14 SEED URLs were selected to carry out the information collection step.

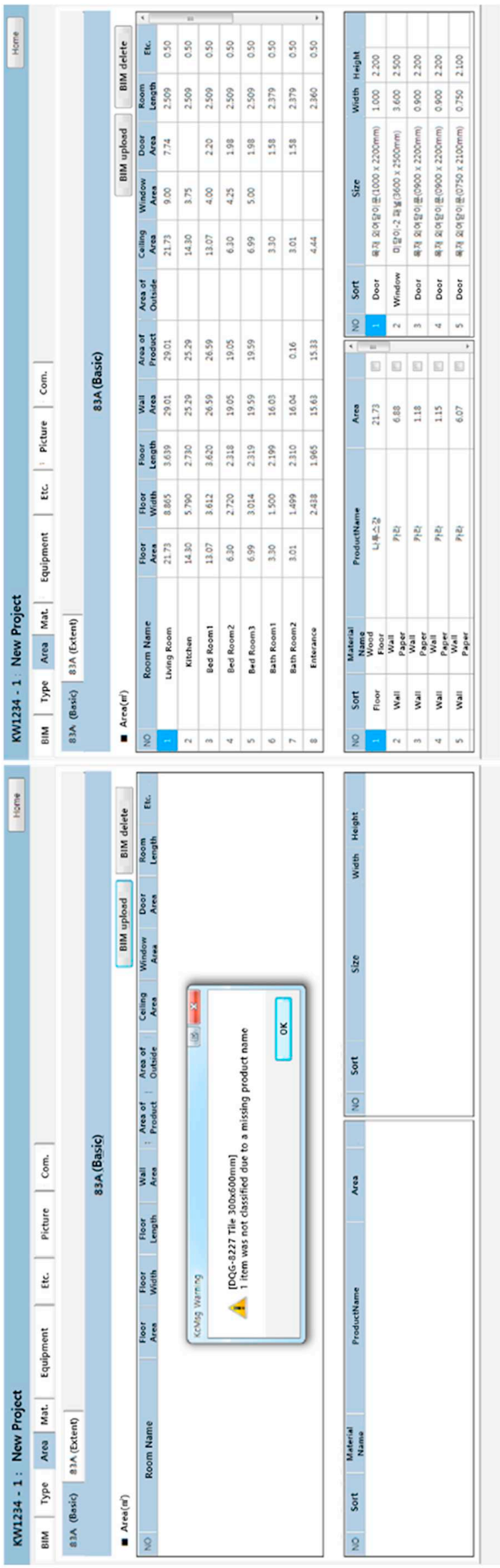

Fig. 11. Screens of uploading BIM Model.

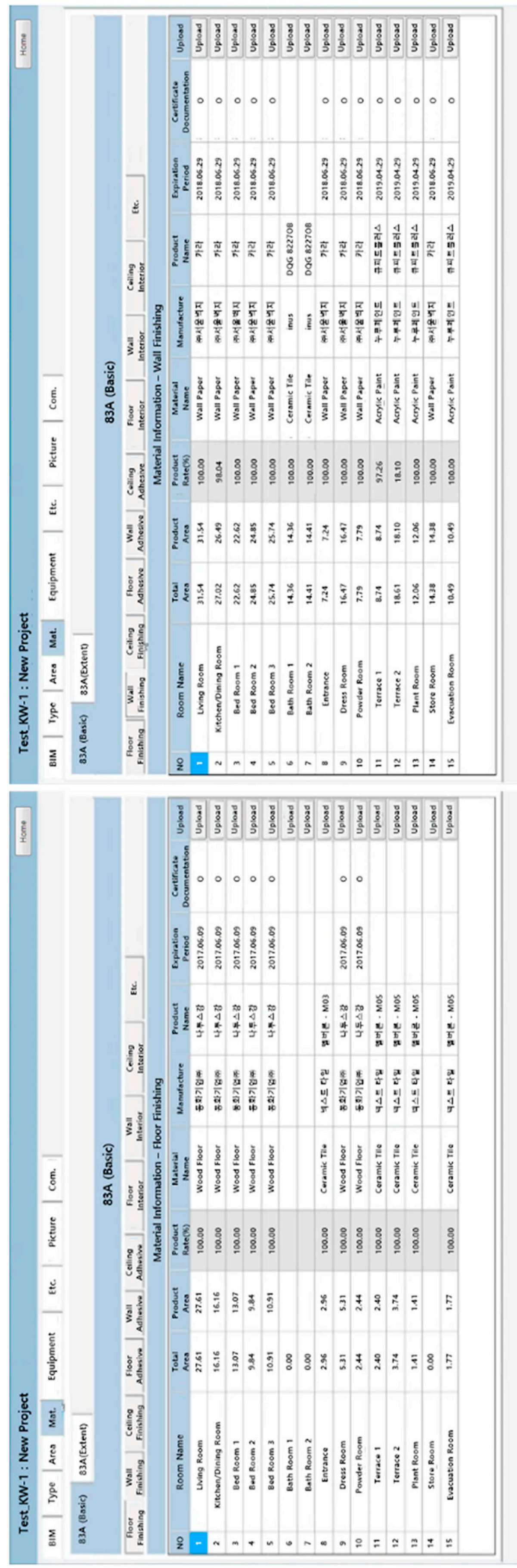

Fig. 12. Screens of matching GBMI.

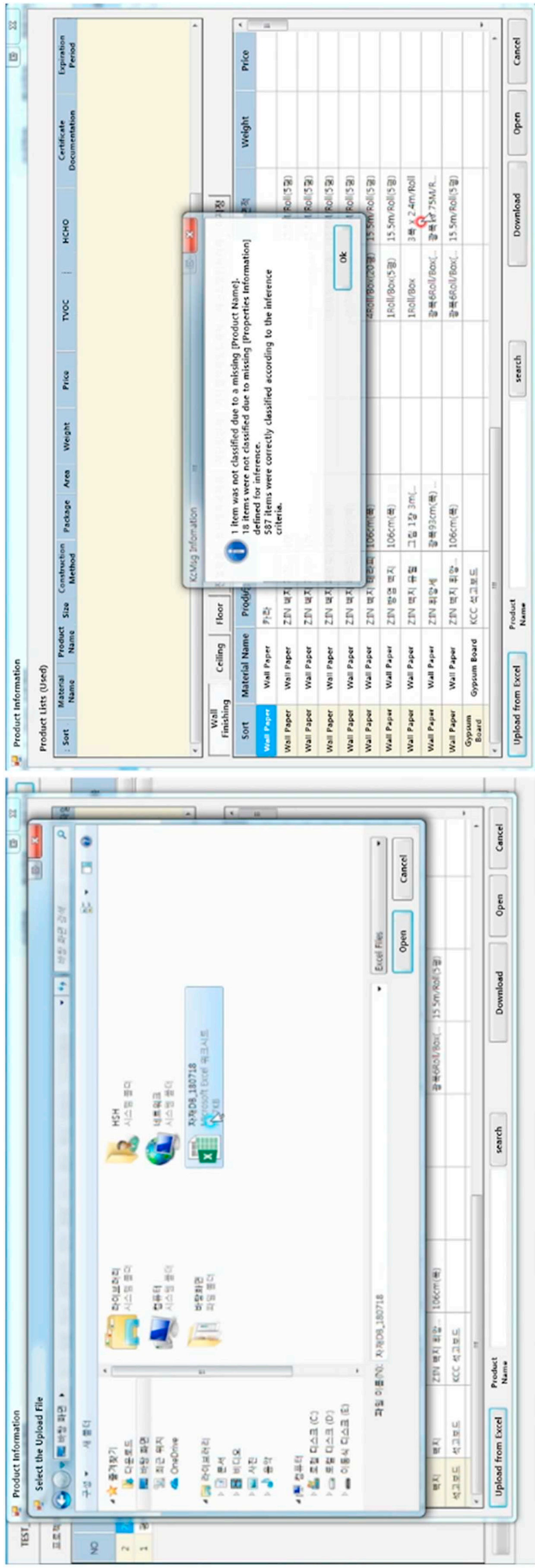

Fig. 13. Screens of uploading the result of web-crawling.

Table 7
Expected reduction of working time by using proposed process.

		A	B	C	D	Total (A + B + C + D)	The degree of improvement
Project 1	Existing	4.7	2.7	10.8	4.0	22.2 h	69.6%
(The unit of house: 6)	Improvement	0	1.4	5.4	0	6.8 h	
Project 2	Existing	5.7	8.1	24.3	8.0	46.1 h	64.9%
(The unit of house: 30)	Improvement	0	4.1	12.2	0	16.2 h	
Project 3	Existing	5.7	5.4	21.6	6.0	38.7 h	65.1%
(The unit of house: 14)	Improvement	0	2.7	10.8	0	13.5 h	

A: Checking the applying materials.
B: Checking the list of applied materials and certification.
C: Selecting the applicable material list and calculating the applicable area.
D: Preparing the attached documents.

Information collecting was performed with a web crawler that made the program statements (written in Python) and put them into an executable file. Based on the collected information, the process classifies the information by changing the class and object properties of the ontology, constructed through the public procurement classification system. As a result, it was confirmed that the information was classified consistently, and it was verified that the proposed construction information management process was automated.

The proposed automated process suggests that web crawling and the ontology can improve construction material information management processes. However, this study is limited to the interior-finishing construction materials that are applied to walls, ceilings, and floors. Therefore, in future research, the process scope should be extended to other materials and different fields.

The expected effects of the proposed automation process are as follows.

First, the information collecting process can be simplified by providing a database in which the latest information is updated periodically by the green building consulting company. Second, it is possible to improve the quality of information by eliminating the human error by minimizing human intervention. Third, it is possible to recommend materials according to the requirements of building users or clients by providing a management system that effectively manages material information. For example, the proposed automated process can recommend a product that has a certain level of eco-performance and that is frequently used by people. Lastly, the proposed automated process is applicable to overseas construction fields where information is collected from various sources and standardized material names are different according to users. Therefore, it is expected that the reliability of information utilization in the construction field will be improved by automatically managing and supplying information according to the standards with the proposed automation process.

Acknowledgment

The present Research has been conducted by the Research Grant of Kwangwoon University in 2018.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2017R1A2B4012228).

References

[1] Amstrong World Industries, https://www.armstrongceilings.com/, (2017) (U.S.A.).
[2] Ji-Heay Bae, Sung-Woo Shin, Byeong-ho Lee, LOD(Level of Development) Standard for Development of Sustainable BIM Guideline, Design Convergence Study 47, 13(4) 2014.08 (ISSN(Print): 2287-4089), http://210.101.116.18/kiss9/viewer.asp .
[3] BuildingSMART EXPRESS to OWL conversion routine, http://openbimstandards.org/wp-content/uploads/sites/2/2016/04/20151014_LDWG_EXPRESS-to-OWLrecommendation.pdf, (2015).
[4] Carrie Sturts, Makoto Sakagami, Implementing Web-Based Project Management Systems in the United States and Japan, J. Constr. Eng. Manag. 134 (3) (2008) 189–196, https://doi.org/10.1061/(ASCE)0733-9364(2008)134:3(189).
[5] Dean Allemang, Jim Hendler, Semantic Web for the Working Ontologist : Effective Modeling in RDFS and OWL, Second Edition, Morgan Kaufmann, 978-0-12-385965-5, 2011.
[6] Derek H.T. Walker, Martin Betts, Information Technology Foresight: The Future Application of the World Wide Web in Construction, Information Technology Support for Construction (1997) 399–407 http://itc.scix.net/paper%20w78-1997-399.content.
[7] Green Construction Material Information system, Korea Environment Industry & Technology Institute (KEITI), http://gmc.greenproduct.go.kr/main.do, (2017.01).
[8] GREEN SPEC, http://www.greenspec.co.uk/, (2017) (U.K.).
[9] Hexu Liu, Ming Lu, Mohamed Al-Hussein, Ontology-based semantic approach for construction-oriented quantity take-off from BIM models in the light-frame building industry, Adv. Eng. Inform. 30 (2016) 190–207, https://doi.org/10.1016/j.aei.2016.03.001.
[10] KCC, http://www.kcc.in/product/about/proudctAbout1.asp, (2017).
[11] Jung-Kyu Kim, A Study on Problems and Improvement to the Information Systems of Construction Materials (MS thesis), Chung-Ang Univ, 2013, http://dcollection.cau.ac.kr/public_resource/pdf/000000055353_20181022125242.pdf.
[12] Yoong-Sil Kim, A Study on the Integrated Design Process for Sustainable Architecture Certification (Doctor of Philosophy), Kook-Min Univ, 2014, http://kookmin.dcollection.net/public_resource/pdf/000001742487_20181022125541.pdf.
[13] Korea Public Procurement, Korea e-Catalog Information system, http://www.g2b.go.kr:8100/onTechCat01L.do?lefttab=techdic, (2017).
[14] Seul-Ki Lee, Ka-Ram Kim, Jung-Ho Yu, BIM and ontology-based approach for building cost estimation, Autom. Constr. 41 (2014) 96–105, https://doi.org/10.1016/j.autcon.2013.10.020.
[15] L.G. Hausys, http://www.z-in.co.kr/rn/productcategory/category3/list.jsp?sup_category=A02&mid_category=A0201&fin_category=A020102, (2017).
[16] Shang-Kyu Noh, Jin-Su Park, Ontology : The Key of Internet Evolution, from Web 2.0 to Web 3.0, gods Toy business, First Edition, (2007) 978-8-99-591911-8.
[17] Kun-Soo Oh, Jeong-Hwa Song, Ju-Hyun Lee, So-Hyun Park, Information Integrated Management through Life Cycle of Construction Materials in Apartment Housing, Korean J. Constr. Eng. Manag. 1226-9107, 18 (3) (2009.08) 175–183 http://www.aik.or.kr/html/page04_01_list.jsp?ncode=p002c.
[18] Hyung-Jin Park, Kyo-jin Kook, Metadata based Information Management Prototype System of Building Material, J. Archit. Inst. Korea Struct. Constr. 1226-9107, 30 (5) (2014) 109–116 http://www.aik.or.kr/html/page04_01_list.jsp?ncode=p002c.
[19] Pollaphat Nitithamyong, Miroslaw J. Skibniewski, Wed-based construction project management systems: how to make them successful? Autom. Constr. 13 (2004) 491–506, https://doi.org/10.1016/j.autcon.2004.02.003.
[20] Product Classification System, Korea Instituted of Procurement, http://www.kip.re.kr/mall/business/listSphere.asp, (2017).
[21] Soo-Hoon Ryu, A Study on the Improvement of Green Building Certification Information System, J. Digit. Archit. Interior Assoc. 11 (3) (2011.09) 5–16 (ISSN (Print):1738-0790).
[22] Salman Azhar, Wade A. Carlton, Darren Olsen, Irtishad Ahmad, Building information modeling for sustainable design and LEED rating analysis, Autom. Constr. 20 (2011) 217–224 https://doi.org/10.1016/j.autcon.2010.09.019.
[23] Dong-Min Seo, Han-Min Jung, Intelligent Web Crawler for Supporting Big Data Analysis Services, J. Korea Contents Soc. 13 (12) (2013) 575–584, https://doi.org/10.5392/JKCA.2013.13.12.575.
[24] Sweets Catalog, http://sweets.construction.com/, (2017) (United States).
[25] H.P. Tserng, Y.L. Samuel, R.J. Yin, D. Wou, M.D. Tsai, W.Y. Chen, A study of ontology-based risk management framework of construction projects through project life cycle, Autom. Constr. 18 (2009) 994–1008, https://doi.org/10.1016/j.autcon.2009.05.005.
[26] Tuscan, http://www.idsurfaces.co.uk/, (2017) (U.K.).
[27] Wu Wei, R.A. Raja, BIM Execution Planning in Green Building Projects: LEED as a Use Case, J. Manag. Eng. 31 (1) (2015) 01, https://doi.org/10.1061/(ASCE)ME.1943-5479.0000314.
[28] Seo-Kyung Won, Sun-Kuk Kim, An Application of Finishing Material Information System for Building Finishes, J. Archit. Inst. Korea Struct. Constr. 1226-9107, 24 (1) (2008) 179–186 http://www.aik.or.kr/html/page04_01_list.jsp?ncode=p002c.