LOST IN TRANSLATION

Achieving semantic consistency of name-identity in BIM

Y.R. HUANG,
University of New South Wales, Sydney, Australia
BLAIR1217@outlook.com

Abstract. Custom room naming in architectural projects can vary considerably depending on the user in building information models (BIM). Having multiple and diverse names for the same room is particularly problematic for information retrieval processes in BIM-based projects. Team agreement on naming labels for rooms in BIM models can avoid unnecessary misunderstanding and aim to more efficient collaboration, but current methods to achieve this remain laborious and flawed. One way of overcome inconsistencies of room naming is to rename them manually to align with an office-wide standard, yet it leads to compounding errors. This research explores how an automated naming-standardization workflow can enhance the interoperability of object-based modeling in a BIM environment and make information retrieval more reliable for a project life cycle. The novelty of this research involves (1) building a custom corpus specialized for architectural terminology to fit into the BIM environment and (2) devising a standard-naming system titled WuzzyNaming to save manual work for BIM users in maintaining room-name consistency. This was achieved by applying the natural language processing (NLP) technique and Fuzzy logic to perform the semantic analysis and automate the BIM room-name standardization. This research contributes to eliminating laborious technical tasks and directing architects back to design rather than fixing repetitive BIM error.

Keywords. Building information modeling; Natural Language Processing; Data interoperability; Naming convention; Fuzzy logic

1. Introduction:

In the architecture profession, building information modeling (BIM) is a parametric representation of shared-data management for the design, delivery, and production of building projects. BIM works through a data repository that consists of semantic information of object identification to facilitate digital data-sharing and design decision-making along the project. However, semantic inconsistencies in BIM entity labelling can significantly impact the efficiency and preciseness of data management, particularly on large architecture projects. Parameterized design objects in BIM models, require precise naming in order for the system and its users to navigate to the correct entities. (Pratt 2004). Information retrieval in existing BIM software, more generally, suffers from issues of low efficiency and poor accuracy because of the error-prone human naming and heterogeneous naming protocols, which has hampered the wider uptake of BIM platforms in the AEC industry.

In accordance with the overarching research methodology of action research (AR) that is adopted in this project, consultations with the research industry partner Cox Architecture have identified the BIM entity naming problem as an unresolved issue that impacts the AEC industry. For a host of reasons, the industry has been reluctant to adopt a standard BIM entity naming protocol, meaning that inconsistences in BIM entity naming continue to compromise seamless data coordination and team and stakeholder collaboration. And according to the ABAB (2018, p.226) with the "growing uptake of BIM, Australia has a timely window of opportunity to develop a common framework for BIM process consistency". But the question remains, what is the most appropriate and reliable method to address the problem of BIM entity naming inconsistencies?

Following the core principles of action research, that include iterative or repeated cycles of planning, acting, evaluating and reflecting, the research project outlined in this thesis investigates and develops a computational workflow to overcome semantic differences in BIM entity naming. More specifically, the research project implements and explores the NLP technique to drive an automated BIM room re-naming workflow. The standard process needs to work in hybrid automated and manual processes to avoid human natural reluctance in accepting machine automation during work process. (Ruikar et al. 2005). To encourage the utility of system, it is wrapped as a web-based API and can be conveniently carried out in BIM platforms. Consequently, the established workflow harnesses NLP and the technique of Fuzzy logic to address the non-uniform BIM entity naming in architectural projects. By developing a workflow that overcomes the barrier of language meaning conflicts, data reliability and interoperability in BIM

projects can be maintained which ultimately contributes to advancing multi-disciplinary collaboration in architectural projects.

2. Research Aims

The overarching goal of this research project is to apply semantic knowledge to enhance the data representation in collaborative design projects. More specifically, the project aims to develop a workflow to mitigate inconsistent BIM entity identification naming in architectural projects.

The project will develop techniques to extract entity labels for room names in a Revit model and utilize the algorithmic methods of NLP and approximate string matching in the Python programming language. Finally, hosting the system in a web-based interface for user's access. The completed language-standardization workflow proposed aims to assist designers by automating the task of making BIM entity naming consistent across a project which is currently undertaken manually. Ensuring naming consistency and accuracy in BIM data-management contributes to fostering the robustness of BIM project management in the AEC industry.

3. Research Question(s)

Based on the research aims summarized above, the questions are raised to be solved by the research:

How can Natural Language Processing be applied to develop a workflow that automates BIM room naming inconsistences to align with established organizational standards?

4. Methodology

With the definition of 'contribute both to the practical concerns of people in an immediate problematic situation and to further the goals of social science simultaneously' (ABL Group 1997), this research project adopts the overarching methodology of Action Research. Action Research (AR) is characterized by bridging the relationship between practice and concepts ideas to satisfy industry identified problems. AR is oriented towards collaboration amongst stakeholders involved and aims to incorporate multiple perspectives form industry to enhance the validity of the research (Hearn & Foth 2005). AR follows an action-reflection cycle that is organized iteratively from reconnaissance to action-execution (Lewin 1946) and this organization is reflected in this research project.

In the research project, the problem has been collaboratively identified between the researcher and industry partner (Cox Architecture). With the collaboration of industry partner, the knowledge-gap between clients and researcher can be alleviated through the intervention of action research,

which enable to comprise ‘tactic and codified knowledge’ in this research (Hearn & Foth 2005). The research issue is been investigated by developing the language-processing workflow, this procedure will involve 3 iterations. (Figure 1) Beginning with the inquiry of naming convention in BIM working-environment, action research advocates the appropriate evaluation of current expectations towards the research object for gathering potential actions. Specifically speaking, the invalid room’s name result in inefficient synchronous exchange of documentation and conflict between stakeholder involved in design team, this leads to the call for established organizational standards regards to existing naming system in BIM context.

Following by the diagnosing stage, the research demands corresponding plans to discover the rational solution. For addressing the issue caused by inconsistent names, an appropriate language prototype should be developed in the action-taking phase. Suitable data and language models must be collected and organized as the fundamental components for executing name-standardization along future iterations. Same process should be conducted and evaluated continually to gain comprehensive apprehension of research topic and improve the system accordingly. Through applying the cyclical process of action research, each iteration will be gauged based on the efficiency and accuracy test by which it makes the BIM naming standard be rational in workplace. Consequently, the iterative design process guarantees the effectiveness of research outcome and offer potential industry-practice in future.

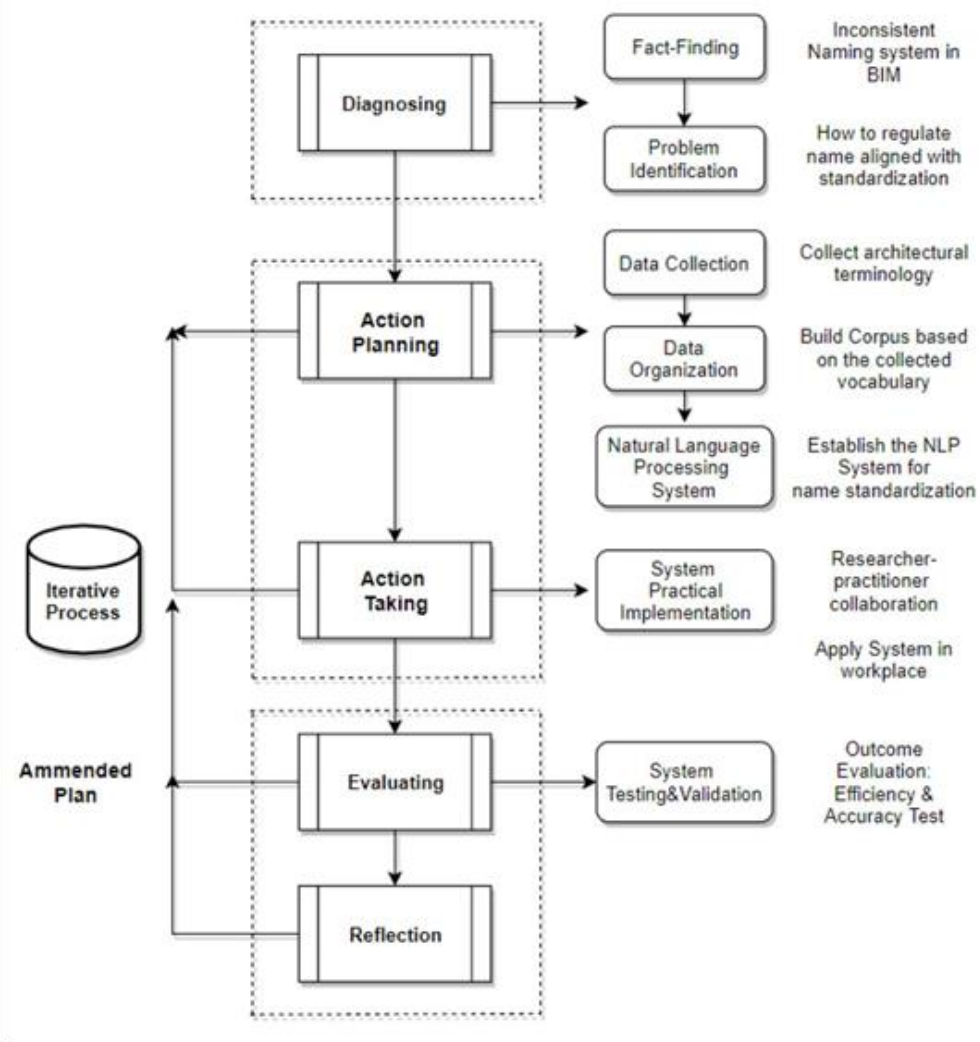


Figure 1. Action Research Cycle of the Research Project

5. Literature review

BIM STANDARDIZED NAMING CONVENTIONS

The term of Building Information Management, defined by NIBIMS (National Building Information Modeling Standard), is an advanced process which covers multiple aspects (planning, design, construction, operation and maintenance) associated with architecture project via implementing a

standardized machine-readable information model (NIBS 2007, p.25). In BIM project, the predefined naming standard is considered as the fundamental identifier of prefabricated elements at the preliminary manufacturing stage (Chen et al. 2015). The prevailing BIM standards allow for precise object recognition and remedy of ambiguous definitions which maintain the shared vision and common agreement amongst stakeholders (Barbosa et al. 2016).

Although BIM processes have been widely adopted in the architecture, engineering, and construction (AEC) industry globally, entity-handling remains problematic. The various naming conventions adopted by practitioners involved in AEC industry lead to semantic differences of entity labels. Many scholars argue that inefficiencies in architectural project organization and production related to entity naming misunderstandings between project team members, could be overcome through a standardized BIM naming system (Taylor 2007; Yang & Zhang 2006).

Nevertheless, developing a standardized BIM naming system is a challenging task. It is estimated that across the 20,000 architecture organizations in Australia each have developed their own unique in-house BIM library (Duddy et al.,2013). Lee et al (2012) identifies four key factors that need to be addressed to develop a standardized BIM entity naming process including: taxonomy (informal description), manual classification and deficiency of name uniqueness. Consequently, the current BIM environment requires a harmonization of entity names to limit the complexity across different naming documentations and eliminate onerous tasks in project management.

Several existing methodologies and guidelines have been imposed to solve the naming conflicts, ranging from numerous versions of official BIM standards (e.g. NATSPEC 2011, NBIMS 2012) to industrial specifications (e.g. IFC -Industrial Foundation Classes) for keeping the name conformance. Before exploring the potential of name compliance in BIM platform, Eastman (2005) introduced a Design Review tool to adopt the naming conflicts between Departmental space and individual base spaces. Chen et al (2017) developed a semiautomatic naming system accordance with the properties acquired from BIM objects in Revit, it enables simultaneous update and revision of inconsistent space names to eliminate the misinterpretation and identify human error. While it applied an auto-naming tool to ensure semantic consistency, this complicated system requires extra time to classify components. Additionally, it is difficult for non-expert users to retrieve the object based on the name generated in this unacquainted system. These examples put emphasis on accomplishing compliance checking in BIM models, but rather than impose a stringent BIM-naming standard on thousands of organizations, an automated process could identify

and rectify multi-named BIM entities. Duan (2011) attests to this, arguing that instead of re-adapting a new formal language standard, delivering a semantic mechanism to formalize the natural-language representation is more considerable and efficient.

In the architectural workplace, it is time-consuming for organizations to train employees to follow new naming identifications imposed by the unified name schema. To more effectively address the naming conflicts related to the use of different BIM libraries, and to overcome issues of ambiguity and diversity in entity naming, a generic semantic representation can be applied to represent, organize and regulate this information in real time.

DATA RETRIEVAL IN NATURAL LANGUAGE PROCESSING (NLP)

In order to construct a system that can extract, identify, organize, and rename BIM entities accurately, this research proposes the application of natural language processing (NLP). Kumar (2011) defines NLP as "a field of computer science and linguistics concerned with the interactions between computers and human(natural) languages" (p.65). NLP has been applied frequently on information retrieval, data extraction and data access with the goal of accomplishing 'human-like language processing' (Liddy, 2001). It should be noted that several tools involving NLTK (Bird et al. 2009) and spaCy (Honnibal & Montani 2017) are available for executing natural language processing.

In the context of this research, the NLP focuses on the semantic measure. These involve semantic similarity, string matching and keyword search. Natural language, with the inherent property of ambiguity and vagueness, is likely to provoke confusion and imprecise outcomes for computational processing. The language uncertainty can be ameliorated through evaluating the semantic similarity. Dating back to the DISTANCE algorithm (a search technique) developed by Rada and Bicknell (1989), the conceptual distance between different terms has been measured to support the information retrieval in artificial intelligence. On that basis, Resnik (1999) introduced the IS-A taxonomy approach to unravel semantic ambiguity and syntactic ambiguity by measuring the word-relatedness exist in corpus.

With the rapid development of computer science, NLP has become a transdisciplinary technique to assist semantic analysis. WordNet, an online lexical system developed from Princeton University (Miller et al. 1990), organizes and encapsulates word ontology in Synset (Synonym set) for word identification and differentiation. Applying this approach, Palkovskii et al (2011) presented an external plagiarism detection system by employing WordNet for Word sense disambiguation. In the medical field, Kim et al (2003) synthesized all the medical terminology from biological literature

into a comprehensive corpus with the help of NLP. But in the architecture, engineering and construction field (AEC), and in relation to design technology there are limited examples of NLP implementations. The NLP method raises the likelihood of organizing the information exchanged in BIM project and helps to standardize the natural language term of object name into specialist language concept.

COMMUNICATION OF COLLABORATIVE BIM SYSTEM WITH NLP

NLP is an appropriate fit for addressing BIM entity naming problems because according to Succar (2009), semantic richness plays an important role in object-based modeling of BIM collaboration. In the AEC industry, the interoperability of data exchange is vital to facilitate the efficient sharing of valuable information synchronously between stakeholders. But insufficient semantic representation leads to errors and the inability for model exchange.

While organizations who adopt BIM approaches employ different ways of working and regulations, Lorio et al (2016) proposed a semantic system to alleviate information dilemmas across different platforms, stakeholders, disciplines, and organizations. This system includes natural-language understanding techniques, which interprets design issues into statements with formal language use. The uniformed outcomes avoid designers suffering from misunderstanding heterogeneous problem definitions in collaborative design projects. Through applying NLP, an automated design-review database developed by Lee et al (2012) overcame string-matching issues in a name-based mapping process. Hence, similar ideas can be applied from formalizing design issues to standardizing room names.

The NLP interface in CAD tools presents the advantage of replacing specialized terminology into natural language expression, which avoids the contradiction in language use and allows easy recognition for users (Samad et al. 1985). Compare with traditional interface, an experiment proved that NLP receives better evaluation of user-preference and efficiency (Biermann et al. 1983). Furthermore, some studies discussed the implementation of NLP on facilitating the project management in Building information modeling (Lin et al. 2016; Lee et al. 2012).

In conclusion, the existing research shows that NLP has found applications in AEC field particularly in the domain of model-exchange or automated document classification (Venugopal et al. 2012; Jung & Lee 2019; Salama et al. 2016). However, there are limited examples of its use in relation to natural language processing for BIM naming convention as NLP is considered as the mainstream tool in Computer Science area in recent years. But it should be acknowledged that there is growing interest in

bridging the gap between Computer Science and Architecture discipline. Thus, as an essential branch of Computer science and Artificial intelligence, Natural language processing could be worth exploring to benefit the collaborative BIM system.

6. Case Study

To investigate the application of NLP in a workflow to automate room naming inconsistencies in a BIM project, this research project has adopted the case study of a finished stadium project from Cox, which incorporates the pragmatic advice from industry. The project proceeded in two key stages that involved firstly developing a custom architectural corpus and secondly programming a workflow to create a web-based naming standardization system. The architectural corpus was built first to specify the architectural terminology, this supports the semantic normalization with more built-in knowledge of room names. The system is wrapped as a python module and delivered as the Web-API interface for standard consultation in architectural projects. Hosting the application on the web makes the workflow more accessible across an organization. The workflow is capable of operating on diverse platforms for the convenience of users such as Dynamo or Grasshopper in the design workplace.

6.1 CHOOSING THE NAME DOCUMENTS FOR ANALYSIS: THE STADIUM PROJECTS

For realizing the possibility of a controlled language system in the architectural workplace, an iterative script is programmed to host the standardization process. With the ability to recognize and process design terminology, it is necessary to train machines to be aware of the jargons and direct them into correct classification. Due to the time limitation, the research scope confined to room names of Stadium projects. The training material was chosen from the existing room-name documents of Cox stadium projects. Through integrating the existing room names, the time spent on collecting specialized vocabulary will be saved. Employing the room names used in ongoing projects will prevent the system from being impracticable in the workplace.

6.1.1 Pre-analysis of the original room-names

To ensure the quality of the name-standardization approach, common problematic expressions during design collaboration have been revealed by analysing the features of received room names. Accordingly, the following system put emphasis on resolving the frequent language hurdles of room

naming listed in Table 1, for delivering explicit representation of room semantics.

TABLE 1. Category of common language conflicts in Room-Names

Name Category	Definition	Representation
Polysemantic word	Same meaning with different expressions	Toilet, Bathroom, Loo, Restroom..
Word contains excess content	Name with non-string characters (number, punctuation…)	PROPERTY-ROOM 3E.
Abbreviation	Name with unrecognizable abbreviation	COMMS ELEC.
Misspelled words	Name with spell mistake	HOME_MASSAGE

6.2 ESTABLISHING DOMAIN-SPECIFIC CORPUS, A WORD-DATABASE OF ARCHITECTURAL TERMINOLOGY

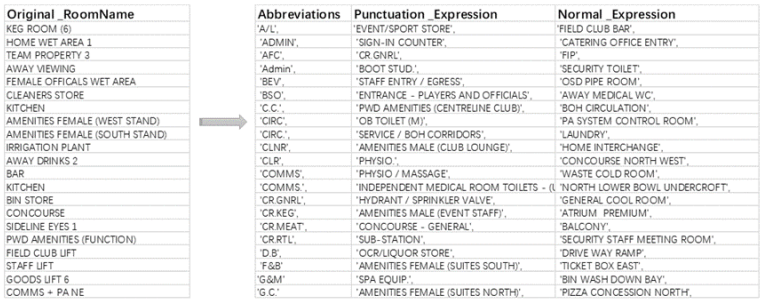
"Structured collections of annotated linguistic data are essential in most areas of NLP" (Bird et al. 2009). The domain-specific corpus is of the essence in achieving the accurate evaluation for the NLP system. Nevertheless, due to the limited application of NLP in the architecture discipline, the current ubiquitous dictionaries are unavailable for achieving the research aim. Since unsuitable language database hinders language analysis, building specialized corpus should be considered as the core precondition to assure the validity of linguistic processing in the AEC industry.

6.2.1 Gathering and organizing the existing room name resources

The development of representative corpus in the architectural domain starts from deploying an NLP-based workflow to pre-process raw names in existing BIM documents. The procedure involves several steps: cleaning, tokenization, and classification of names. Since unconventional phrases occurred frequently in room identification, the heterogeneous room names are categorized as follows to avoid missing those non-standard collocations (Figure 2):

1. Abbreviations.
2. Expression with punctuations.
3. Normal expression.

These name groups will be processed independently to expand the richness and diversity of architectural expressions that constitute the corpus.



Original _RoomName
KEG ROOM (6)
HOME WET AREA 1
TEAM PROPERTY 3
AWAY VIEWING
FEMALE OFFICALS WET AREA
CLEANERS STORE
KITCHEN
AMENITIES FEMALE (WEST STAND)
AMENITIES FEMALE (SOUTH STAND)
IRRIGATION PLANT
AWAY DRINKS 2
BAR
KITCHEN
BIN STORE
CONCOURSE
SIDELINE EYES 1
PWD AMENITIES (FUNCTION)
FIELD CLUB LIFT
STAFF LIFT
GOODS LIFT 6
COMMS + PA NE

Abbreviations	Punctuation _Expression	Normal _Expression
'A/L',	'EVENT/SPORT STORE',	'FIELD CLUB BAR',
'ADMIN',	'SIGN-IN COUNTER',	'CATERING OFFICE ENTRY',
'AFC',	'CR.GNRL',	'FIP',
'Admin',	'BOOT STUD.',	'SECURITY TOILET',
'BEV',	'STAFF ENTRY / EGRESS',	'OSD PIPE ROOM',
'BSO',	'ENTRANCE - PLAYERS AND OFFICIALS',	'AWAY MEDICAL WC',
'C.C.',	'PWD AMENITIES (CENTREUNE CLUB)',	'BOH CIRCULATION',
'CIRC',	'OB TOILET (M)',	'PA SYSTEM CONTROL ROOM',
'CIRC.',	'SERVICE / BOH CORRIDORS',	'LAUNDRY',
'CLNR',	'AMENITIES MALE (CLUB LOUNGE)',	'HOME INTERCHANGE',
'CLR',	'PHYSIO.',	'CONCOURSE NORTH WEST',
'COMMS',	'PHYSIO / MASSAGE',	'WASTE COLD ROOM',
'COMMS.',	'INDEPENDENT MEDICAL ROOM TOILETS - (L	'NORTH LOWER BOWL UNDERCROFT',
'CR.GNRL',	'HYDRANT / SPRINKLER VALVE',	'GENERAL COOL ROOM',
'CR.KEG',	'AMENITIES MALE (EVENT STAFF)',	'ATRIUM PREMIUM',
'CR.MEAT',	'CONCOURSE - GENERAL',	'BALCONY',
'CR.RTL',	'SUB-STATION',	'SECURITY STAFF MEETING ROOM',
'D.B',	'OCR/LIQUOR STORE',	'DRIVE WAY RAMP',
'F&B'	'AMENITIES FEMALE (SUITES SOUTH)',	'TICKET BOX EAST',
'G&M'	'SPA EQUIP.',	'BIN WASH DOWN BAY',
'G.C.'	'AMENITIES FEMALE (SUITES NORTH)',	'PIZZA CONCESSION NORTH',

Figure 2. Category of original room-names in stadium projects

6.2.2 Capturing essential architectural expressions from name resources

Based on the former analysis of raw documents, the phrases need to be tokenized and segmented to acquire the individual term. To improve the efficiency of word collection, only frequent concepts are remained to confine the corpus into a manageable size.

To calculate the word frequency, there is a need to transform the word collection into document-term matrix (DTM). The DTM technique enables to convert the linguistic text into machine-readable content in a vectorial semantic environment, then allows for selecting the significant terms among all expressions. The frequency-counting scheme is developed to filter the negligible words and gather valuable words (Figure 3). Figure 4 summarizes the word-extraction procedure with the example of 'Sign-in Counter'. Hence, these shared terms used in BIM projects will be passed into the customized corpus as the basic lexical entries.

Valuable Words
'barbeque',
'basement',
'bathing',
'bathroom',
'battery',
'bay',
'bbq',
'beer',
'admin',
'bench',
'beverage',
'big',
'bike',
'bin',
'bins',
'blast',
'board',
'boardroom',
'carpark',
'carpenter',
'carrier',

	frequency
acces	1
access	218
admin	609
administration	6
admission	2
...	...
workroom	1
workshop	4
written	1
yard	1
zone	5

655 rows × 1 columns

Figure 3. Frequency Ranking of Important Terms

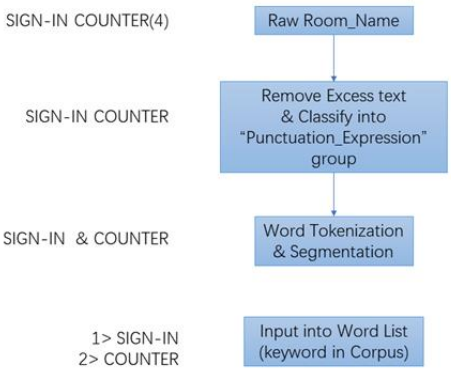


Figure 4. Preprocessing workflow of raw data- 'Sign-in Counter'

6.2.3 Defining Word Classifications and Building Class Hierarchy

Upon the extraction of valid terms, it was discovered that these words are chaotic and unrelated to each other, this results in difficulty for computer to recognize and standardize the name input based on their relations. To build a full-fledged database, the word list requires principled storage within a systematic structure. In the light of this prerequires, the corpus with explicit hierarchy must be established through computing the semantic similarity between terms and grouping them into individual sets under specified name class. Considering the complexity of room classification, the key terms (Figure 5) are specified by the industry partner to serve as the benchmark of name normalization.

Important_Terms
show_power
staff_room
stair
store_room
substation
substitutes_bench
suite
switch
tank
terrace
ticket_box
toilet
television_commentary_box
television_studio
uninterrupted_power_supply
valve
void
warm_up
waste_room
water_meter
workroom
workshop
...

Figure 5. Important room names from industry partner

For navigating the irrelevant words into determined name class, a python

script is built to detect the semantic relationship with respect to the word sense. The hierarchical structure of WordNet (Miller 1990) is suitable to compute the semantic relatedness according to the hypernym, hyponyms, and synonyms stored in its word-Synsets (Synonym-sets). Although WordNet performs well in decoding the semantic relations, the informal lexical entries with architectural specialized knowledge are absent in WordNet database. This bottleneck can be removed by seeking the synonyms of those terms manually to cover the deficiency of unconventional expression.

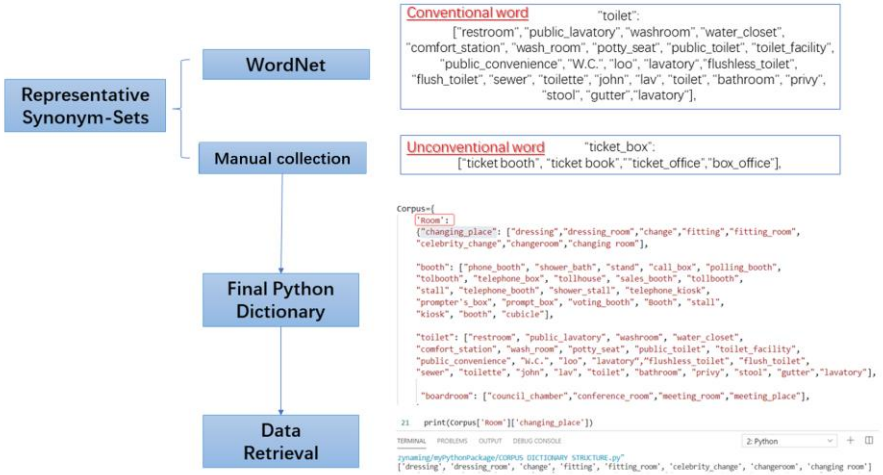


Figure 6. Synonym Sets & custom corpus in python dictionary

Accordingly, combining the lexical collections from both WordNet and manual word-acquisition, the synonyms of important terms are gathered as the fundamental components of the terminological corpus. Supported by adequate vocabulary, the corpus satisfies the accurate data retrieval (Figure 6).

6.2.4 Evaluation of final corpus

The final corpus is delivered to execute the computer-aided linguistic analysis for future name normalization. Commencing on evaluating and categorizing the room terms from stadium name schedules, the corpus is synthesized with accumulated synonym-sets under proper classification (Figure 7). During the preparation process, manual and computerized practice work in parallel for achieving the corpus rationality and remedy the insufficiency of technical phraseology in traditional dictionaries.

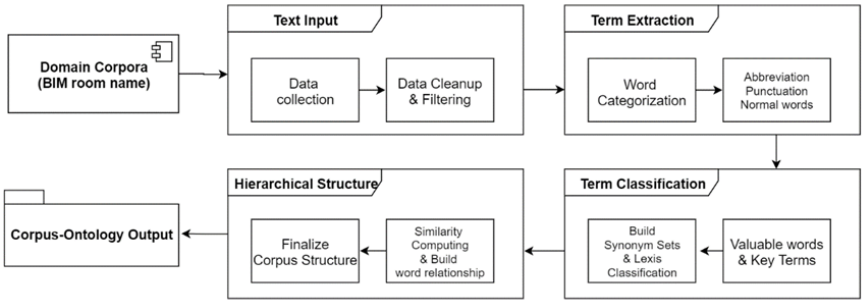


Figure 7. Workflow of domain-specific corpus development

Figure 8 depicts the lexical relations of room-names in final corpus, which gives the machine permission to accomplish the semantic understanding in accordance with the word hierarchy. (e.g.: Computer understands 'Changing_place' is synonymous with 'Fititng_room' as they locate in same class). Given the clear affiliation among vocabulary, the word network is capable of retrieving the term precisely without causing sense ambiguation.

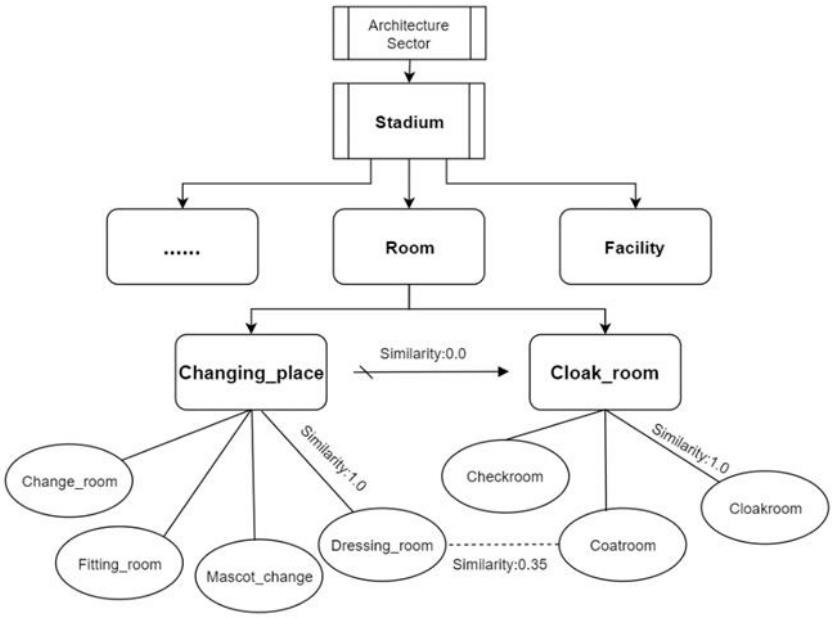


Figure 8. Fragment of the corpus hierarchy: the affiliation of room name.

6.3 BUILDING ‘WUZZYNAMING’, THE NLP PROCESSOR FOR NAME STANDARDIZATION IN DESIGN PROJECTS

6.3.1 Pre-processing of input data

The name standardization is started by the basic NLP procedure, involving punctuation removal and word segmentation. Recalling the pre-analysis of current inconsistent naming, the system gave priority to dispose spelling mistakes from input. Figure 9 illustrates the python script for spell-correction, which nicely rectify the errors. The Fuzzy logic- a computational algorithm assists in finding approximate strings- has been chosen to undertake the string-matching by comparing the pattern similarity between input text and words in corpus.

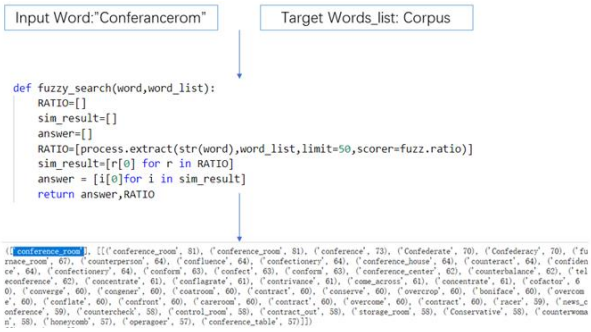


Figure 9. Fuzzy string-matching script

6.3.2 Developing the Name-Standardization Approach for Room-Names

In order to automate the naming standardization from the corrected input, the semantic understanding should be captured for each cleaned term. One major challenge in meeting this intention is how to solve compound expressions that are recurrently demanded in room naming. This synthetic language raises the difficulty for computer to comprehend the name connotation. To make sure the system is flexible and practical enough to standardize the random input from users, the issue of complex words is approached by establishing a comprehensive system to manage disparate scenarios showed in Table 2.

TABLE 2. Variety of random room naming

Random Name Input	Phrasal Structure	Example Name
Open Compound Word	Noun-Noun	ELEVATOR LOBBY
	Adjective-Noun	CLEANED ROOM
Hyphenated Compound Word	Adjective	FIRST-AID
	Noun	COLD_ROOM
Individual Word	Adjective	ELECTRICAL
	Noun	BATHROOM

In the standardized processor (Figure 10), the complex names that disobey the linguistic rules will be tokenized into separated words, the iterative system allows to conduct a thoroughly similarity calculation of each tokenized term compare with the pre-defined corpus. Once the computing completed, each word will be aligned with the standard class which has highest score in relevance ranking. The similarity measurement is carried out through evaluating both morphological and word-sense relationship to guarantee the preciseness of normalization. Similar with the spell correction, fuzzy-logic algorithm assists to examine the word-form similarity. For the word-sense relationship, the explicit word hierarchy in corpus eliminate the confusion of polysemy and associate the word-input to its standard classification if they share equivalence meaning. After verifying the success of normalizing individual lexis, joining them together and output the standardized room name for any intuitive input.

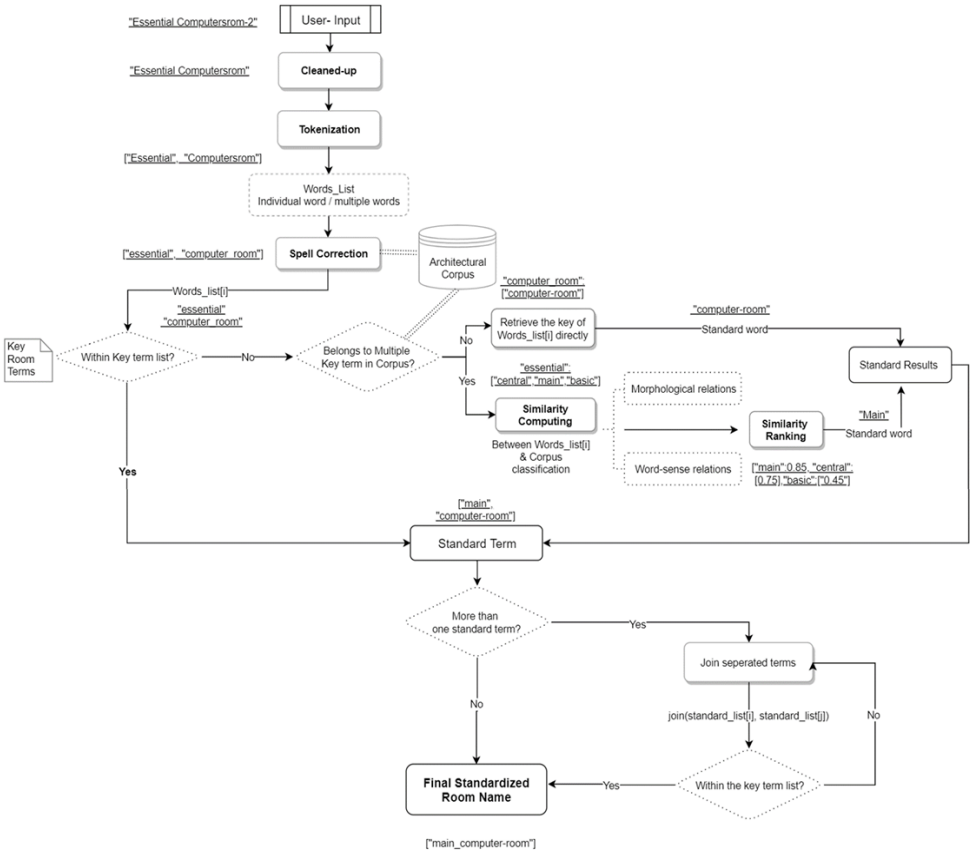


Figure 10. Workflow of standardized processor

6.4 USER-INTERFACE: DEPLOYING THE WUZZYNAMING MODULE INTO DESIGN PROJECTS

After compiling the python module of standardization processor, the completed script was packaged as the Web-API by Django. Django works as a python web framework to establish the programming interface. The delivered interface is through BIM software (e.g. Revit & Rhino) connecting to the API. To achieve this, a Grasshopper script is generated to interact with the API and then give access to WuzzyNaming module (Figure 11). Figure 12 demonstrates that the inconsistent room names in Revit floorplan will be standardized successfully by sending the Grasshopper output to Revit. Therefore, WuzzyNaming should be applicable to different BIM platforms, which permits the reusability of proposed system and eases user's access with simple operational interface.

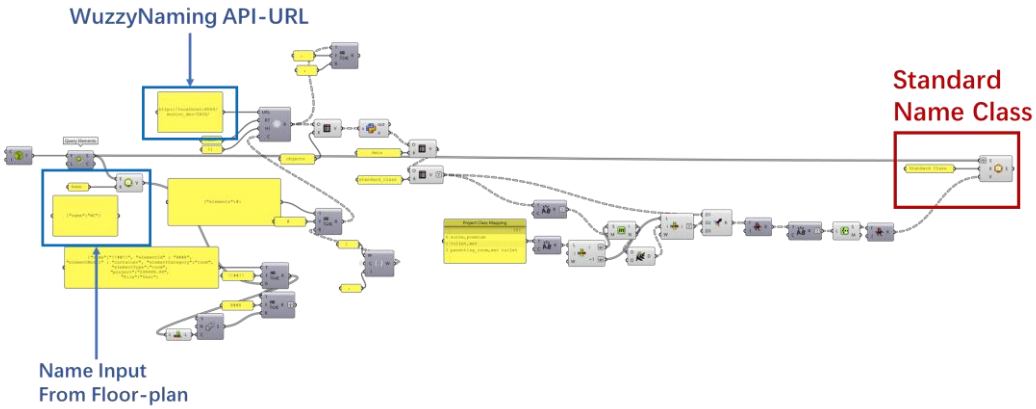


Figure 11. WuzzyNaming-API in Grasshopper workflow

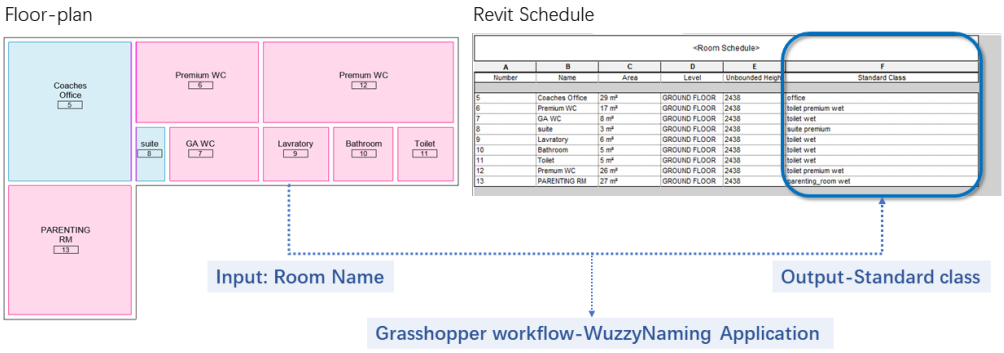


Figure 12. Standard result in Revit room schedule

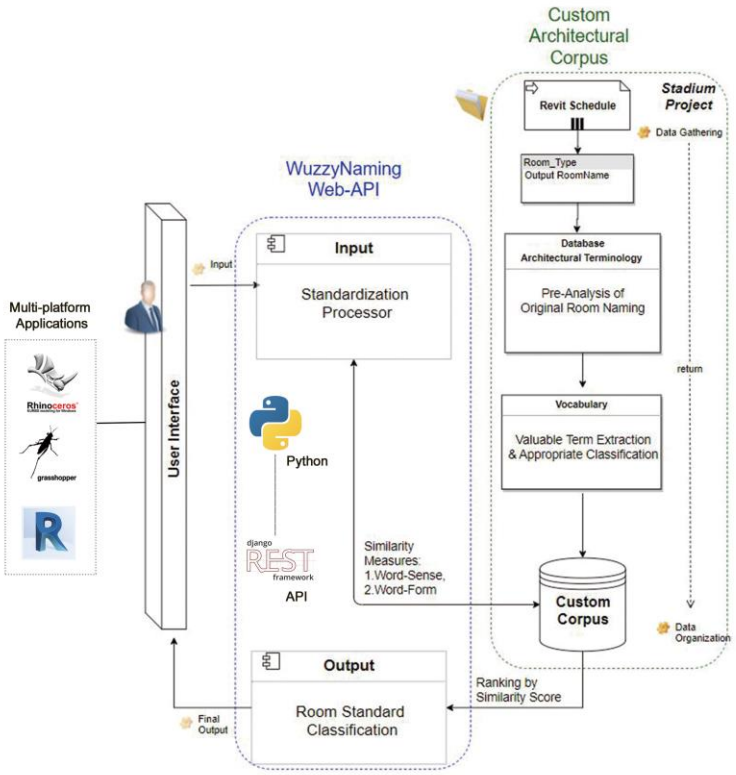


Figure 13. Overview of Standardization workflow

7. Discussion

The standardized classification generated from the normalization workflow demonstrates the success of applying Natural Language Processing technique to solve the naming discrepancy in BIM. To validate the efficacy of the proposed standardization approach accompanied with the customized architectural corpus, a series of room names are picked randomly from the ongoing stadium projects in Cox Architecture to be considered as the study objects for accuracy evaluation. The results shown in Table 3 indicate the viability of system with nearly 95% accuracy rate in acquiring standard name class. The results are calculated by dividing the number of correct room classifications by the total amount of room inputs. What is noteworthy is that after training the machine progressively, the computer's ability in processing complex expression has been optimized with the attempt to offer an enhanced understanding of NLP utilization in BIM, which becomes more responsive for user's requirement. As shown in Table 4, as iterations

progress Step from iteration 1 to the next iteration, the flexibility and accuracy of the workflow improves. (Table4).

TABLE 3. Evaluation outcome

Operating Speed(Second)			Accuracy Rate
Maximum	Minimum	Mean	94.52% 293/310 (total tested names)
0.51s	0.128s	0.28s	

TABLE 4. Reflections of Iterative testing outcome

Iteration1: Original Input	Standard	Reflection
cookroom	kitchen	1. Cannot process the complex expressions 2. Difficult to comprehend the word with abbreviation 3. Cannot direct to the correct classification when there are multiple words in expression
entrance	entry	
secondary wc		
podiam	podium	
rackss		
allergen prep		
comms	communication	
security_room	police	
Iteration2: Original Input	**Standard**	**Reflection**
therapy RM	physio	1. Improved system can deal with compound expressions. 2. Increased Result Accuracy
ticket booth	ticket_box	
drinks	drinks_station	
LIFT LOBBY	elevator_lobby	
ESSENTIAL COMPUTERSROOM	main_computer-room	

However, the limitations of the research project should also be acknowledged. The time constraints of a 10-week research project duration has meant that the project could produce a proof-of-concept demonstrating the basic utility of NLP technique for use in a BIM project for the AEC industry. A key limitation of the developed workflow uncovered during the research process is the vocabulary deficiency of architectural corpus, as in current stage, the devised corpus is the indispensable tool for computer to understand word sense. Given more time, this problem could be overcome with deeper investigation and collection towards the room expressions from different building projects (e.g. Office or Residential Building). Regarding the user interface, the system runs independent in web API at this stage, this leaves space for future development by exploring the user-friendly interface (Figure 14) to assist inexperienced designers in architectural workplace.

.

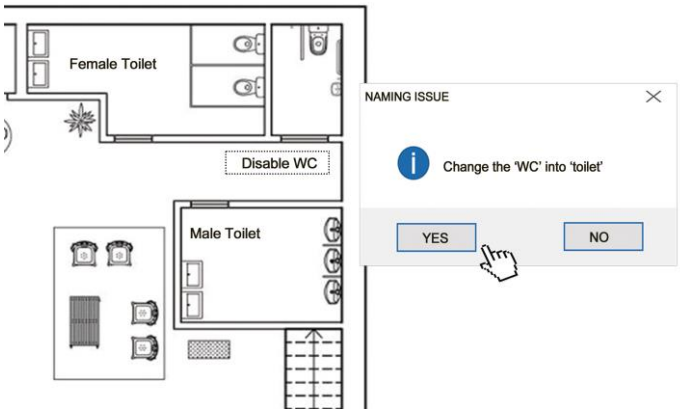


Figure 14. Mock Revit Interface-Popup Window

Furthermore, the present system faces the dilemma to process compound expressions by relying on the immature corpus created in early stage of research. The computational intelligence could be pushed further through introducing the machine learning technique. This shows more opportunities in which profound research could be explored to broaden the usability of system in a larger scale. Regarding the user interface, the system runs independent in web-API at this stage. This leaves space for future development and user testing to improve usability and accessibility across an organization and potentially the AEC industry more generally.

Despite the shortages outlined above, the system is capable of normalizing inconsistent-naming situations and deliver the standard class to Revit schedule (Table5). Along with the BIM popularization in architectural domain, the system showcases the possibility of reforming BIM system in a more intelligence way with respect to the human language understanding.

TABLE 5. Standard class result in Revit room schedule

<Room Schedule>					
A	B	C	D	E	F
Number	Name	Area	Level	Unbounded Height	Standard Class
5	Coaches Office	29 m²	GROUND FLOOR	2438	office
6	Premium WC	17 m²	GROUND FLOOR	2438	toilet premium wet
7	GA WC	8 m²	GROUND FLOOR	2438	toilet wet
8	suite	3 m²	GROUND FLOOR	2438	suite premium
9	Lavratory	6 m²	GROUND FLOOR	2438	toilet wet
10	Bathroom	5 m²	GROUND FLOOR	2438	toilet wet
11	Toilet	5 m²	GROUND FLOOR	2438	toilet wet
12	Premum WC	26 m²	GROUND FLOOR	2438	toilet premium wet
13	PARENTING RM	27 m²	GROUND FLOOR	2438	parenting_room wet

8. Conclusion

This research project has explored the application of NLP to develop the name standardization workflow based in BIM environment, the research result demonstrates that the proposed workflow can effectively delivers the standard room classifications to architectural stakeholders without requiring extra manual work. At the interaction of NLP and BIM, specific computer-aided strategies are put forward with the intention to mitigate risks resulted from non-standard room-naming. By finalizing the system with a web-API interface, the simple access intends to provoke the wide-spread utilization for this approach. Recalling the primary goal of the research in alleviating the inconsistent name, such system adopts the NLP-driven approach to resolve the current room-naming confusion in BIM and aims to increase the precision and effectiveness of normalization process, albeit with potential amendment.

Regardless, this research promotes further directions towards naming normalization. This leaves open exploration derived from this research, with the considerations of reducing the collaboration complexity from the view of semantic interoperability in BIM. Without compelling users to follow a specific naming convention, the system prevents people from struggling with the naming inconsistency in building projects. The naming-standardization system not only promotes the efficient collaboration across different design teams, but indeed lays the foundation of NLP application in BIM to offer a more humanized system. Overall, this research reveals the value of interdisciplinary pursuits that draw on methods in computer science discipline to innovate ways of working in the architecture discipline and AEC industry.

Acknowledgements

Thank you to the Cox Architecture for providing the research time and resources to the project and thank for the great help from Andrew Butler who contributes his idea and technical supports during the research process. Also thank all the teachers involved in this research for providing valuable assistants.

References

ABAB 2018, ‘BIM Process Consistency: Towards a Common Framework for Digital Design, Construction and Operation’, *Australasian BIM Advisory Board*.

ABL Group 1997, *Future Search Process Design*, York University, Toronto.

Azhar, S., Ahmad, I., and Sein, M.K., 2010.’ Action Research as a Proactive Research Method for Construction Engineering and Management’, Journal of Construction Engineering and Management, vol.136, no.1.

Barbosa, M, Pauwels, P, Ferreira, V and Mateus, L 2016. ‘Towards increased BIM usage or existing building interventions’, *Structural Survey*, vol.34, no.2, pp.168-190.

Biermann, A.W, Ballard, B.W and Sigmon, A.H 1983, 'An experimental study of natural language programming', *International Journal of Man-Machine Studies*, vol.18, no.1.

Bird, S, Klein, E and Loper, E 2009, *Natural Language Processing with Python*, O'Reilly edia Inc.

Chen, K, ASCE, S.M, Lu, W.S, Wang, H.D, Niu, Y.H and Huang, G 2017, 'Naming Objects in BIM: Convention and a Semiautomatic Approach', *Journal of Construction Engineering and Management*, vol.143, no.7.

Chen, K, Lu, W. S, Peng, Y, Rowlinson, S and Huang, G. Q 2015. 'Bridging BIM and building: From a literature review to an integrated conceptual framework', *International journal of Project Management*, vol.33, no.6, pp.1405–1416.

Duan, Y and Cruz, C 2011, 'Formalizing semantic of natural language through conceptualization from existence', *Management and Technology*, vol.2, no.1, pp.37-42.

Duddy, K, Beazley, S, Drogemuller, R and Kiegeland, J 2013, 'A platform - independent product library for BIM', *Proceedings of the 30th CIB W78 International Conference: International Conference on Applications of IT in the AEC Industry*, Tsinghua University press, China, pp. 389-399.

Eastman, C 2005, 'Automated Assessment of Early Concept Designs', *Architectural Design*, vol.79, no.2, pp.52-57.

Fellbaum. C., 1998. 'WordNet: An Electronic Lexical Database', MIT Press.

Hearn, G, and Foth, M 2005, 'Action Research in the Design of New Media and ICT Systems'.

Honnibal, M and Montani, I 2017, 'spaCy 2: Natural language understanding with Bloom beddings, convolutional neural networks and incremental parsing'.

Jung, N and Lee, G 2019, 'Automated classification of building information modeling (BIM) case studies by BIM use based on natural language processing (NLP) and unsupervised earning', *Advanced Engineering Informatics*, vol.41.

Kim, J.D, Ohta, T, Tateisi, Y and Tsujii, J 2003, 'GENIA corpus—a semantically annotated corpus for bio-textmining', *BIOINFORMATICS*, vol.19, pp. i180-i182.

Kumar, E 2011, *Natural Language Processing*, I.K. International Publishing House, New Delhi.

Lai, Y.C., Carlsen, M., Christiansson. P., and Svidt, K., 2003, 'Semantic-Web Supported Knowledge Management System: An Approach to Enhance Collaborative Building Design'.

Lee, J.K, Lee, J, Jeong, Y.S, Sheward, H, Sanguinetti, P, Abdelmohsen, S and Eastman, C.M 2012, 'Development of space database for automated building design review systems', *Automation in Construction*, vol.24, pp.203-212.

Lewin, K 1946. 'Action Research and Minority Problems'.

Liddy, E.D 2001, *Natural Language Processing*, In Encyclopedia of Library and Information Science, 2nd edn, Marcel Decker Inc, New York.

Lin, J.R, Hu, Z.Z, and Zhang, J.P 2013, 'BIM Oriented Intelligent Data Mining and Representation', *Proceedings of 30th CIB W78 International Conference on Applications of IT in the AEC Industry*.

Lin, J.R, Hu, Z.Z, Zhang, J.P and Yu, F.Q 2016, 'A Natural-Language-Based Approach to Intelligent Data Retrieval and Representation for Cloud BIM', *Computer-aided civil and infrastructure engineering*, vol.31, no.1, pp. 18-33.

Lorio, F, Cheong, H, Li, W, Shu, L.H, Tessier, A and Bradner, E 2014, 'Natural Language Problem Definition for Computer-Aided Mechanical Design', *ACM CHI 2014 -DSLI Workshop*, Toronto.

Miller, G.A, Beckwith, R, Fellbaum, C, Gross, D and Miller, K.J 1990, 'Introduction to WordNet: An On-line Lexical Database', *International journal of lexicography*, vol.3, no.4, pp.235-244.

NATSPEC 2011, 'NATSPEC National BIM Guide', Construction Information Systems.

NBIMS (National Institute of Building Science) 2012, 'National BIM Standard – United States V2', NBIMS-US.

NIBS (National Institute of Building Science) 2007, 'National building information model standard version 1.0-part 1: overview, principles, and methodologies', *NBIMS-US*, pp.71-87.

Palkovskii, Y, Belov, A and Muzyka, I 2011, 'Using WordNet-based semantic similarity measurement in External Plagiarism Detection', *CEUR Workshop Proceedings*.

Pratt, M. J 2004. 'Extension of ISO 10303, the STEP standard, for the exchange of procedural shape models', *IEEE*, pp.317–326.

Rada, R and Bicknell, E 1989, 'Ranking Documents with Thesaurus', *Journal of the American Society for Information Science*, vol.40, no.5, pp.304-311.

Resnik, P 1999, 'Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language', *Archives*, vol.11.

Ruikar, K, Anumba, C.J, and Carrillo, P.M 2005, 'End-user perspectives on use of project extranets in construction organisations', *Engineering Construction & Architectural Management*, vol.12, no.3, pp.222-235.

Salama, D.M, El-Gohary, N.M and ASCE, A.M 2016, 'Semantic Text Classification for Supporting Automated Compliance Checking in Construction', *Journal of Computing in Civil Engineering*, vol.30, no.1.

Samad, T and Director, S 1985, 'Towards a Natural Language Interface for CAD', *Proceedings of the 22nd ACM/IEEE Design Automation Conference*, IEEE Press, pp. 2-8.

Succar, B, 2009. 'Building information modeling framework: A research and delivery foundation for industry stakeholders', *Automation in Construction*, vol.18, no.3, pp.357-375.

Taylor, J.E 2007, 'Antecedents of successful three-dimensional computer-aided design implementation in design and construction networks', *Journal of Construction Engineering and Management*, vol.133, no.12, pp.933-1002.

Venugopal, M, Eastman, C.M, Sacks, R and Teizera, J 2012, 'Semantics of model views for information exchanges using the industry foundation class schema', *Advanced Engineering Informatics*, vol.26, no.2, pp.411-428.

Yang, Q.Z and Zhang, Y 2006, 'Semantic interoperability in building design: Methods and tools', *Computer-Aided Design*, vol.38, no.10, pp.1099–1112.