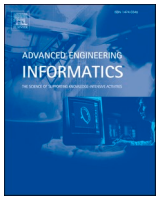
Advanced Engineering Informatics 62 (2024) 102653

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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



Full length article

A platform-based Natural Language processing-driven strategy for digitalising regulatory compliance processes for the built environment

Ruben Kruiper [a], Bimal Kumar [b,*], Richard Watson [c], Farhad Sadeghineko [d], Alasdair Gray [a], Ioannis Konstas [a]

[a] Department of Computer Science, Heriot-Watt University, Edinburgh, UK
[b] Department of Architecture, University of Strathclyde, Glasgow, UK
[c] Department of Architecture and Built Environment, Northumbria University, Newcastle-upon-Tyne, UK
[d] School of Computing, Engineering and Built Environment, Glasgow Caledonian University, UK

ARTICLE INFO

Keywords:
Digital Regulatory Compliance
Natural Language Processing
Semantic Web
Machine Learning
Knowledge Graph
Automated Compliance Checking

ABSTRACT

The digitalisation of the regulatory compliance process has been an active area of research for several decades. However, more recently the level of activities in this area has increased considerably. In the UK, the tragic incident of Grenfell fire in 2017 has been a major catalyst for this as a result of the Hackitt report's recommendations pointing a lot of the blame on the broken regulatory regime in the country. The Hackitt report emphasises the need to overhaul the building regulations, but the approach to do so remains an open research question. Existing work in this space tends to overlook the processing of actual regulatory documents, or limits their scope to solving a relatively small subtask. This paper presents a new comprehensive platform approach to the digitalisation of the regulatory compliance processing. We present i-ReC (intelligent Regulatory Compliance), a platform approach to digitalisation of regulatory compliance that takes into consideration the enormous diversity of all the stakeholders' activities. A historical perspective on research in this area is first presented to put things in perspective which identifies the challenges in such an endeavour and identifies the gaps in state-of-the-art. After enumerating all the challenges in implementing a platform-based approach to digitalising the regulatory compliance process, the implementation of some parts of the platform is described. Our research demonstrates that the identification and extraction of all relevant requirements from the corpus of several hundred regulatory documents is a key part of the whole process which underlies the entire process from authoring to eventually compliance checking of designs. Some of the issues that need addressing in this endeavour include ambiguous language, inconsistent use of terms, contradicting requirements and handling multi-word expressions. The implementation of these tools is driven by NLP, ML and Semantic Web technologies. A semantic search engine was developed and validated against other popular and comparable engines with a corpus of 420 (out of about 800) documents used in the UK for compliance checking of building designs. In every search scenario, our search engine performed better on all objective criteria. Limitations of the approach are discussed which includes the challenges around licensing for all the documents in the corpus. Further work includes improving the performance of SPaR.txt (the tool created to identify multi-word expressions) as well as the information retrieval engine by increasing the dataset and providing the model with examples from more diverse formats of regulations. There is also a need to develop and align strategies to collect a comprehensive set of domain vocabularies to be combined in a Knowledge Graph.

1. Introduction

The well-publicised Hackitt Review [36] into the Grenfell Tower disaster points to some of the main reasons behind the ill-fated fire in the London blocks of flats in 2017. In order to avert similar disasters in the future, Dame Hackitt [36] clearly recommends that the building regulations and associated guidance, including the Approved Documents, need to be authored, applied and enforced in a fundamentally different way – also see Table 1:

* Corresponding author.
E-mail address: b.kumar@strath.ac.uk (B. Kumar).

https://doi.org/10.1016/j.aei.2024.102653
Received 21 February 2024; Received in revised form 21 May 2024; Accepted 14 June 2024
Available online 29 June 2024

Table 1
Summary of challenges in the current system, as identified in [36] and the broader literature.

Ambiguity, confusion and the ability to '*game the system*'.
The package of regulations and guidance (in the form of Approved Documents) is ambiguous and inconsistent;
"There can be differences of interpretation between different building control bodies and also between different individuals within the same building control body";
There is widespread confusion about what constitutes the regulations and what is guidance;
Many of the Approved Documents have not been comprehensively updated for some time (despite their pivotal role);
"The suite of guidance is very slow to adapt and update as new technologies and techniques become available" […]"this creates significant scope for gaming the system in a variety of ways".
Regulatory overload and unintended use.
"The cumulative impact of the Approved Documents changes an outcome-based system of regulation to one that is often inferred by users to be prescriptive";
In the UK, there are some 485 standards, 85 other government guidance, 176 industry guidance and 79 other government legislation documents, most of which are to be complied with whilst others are guidance documents
Those designing or constructing buildings are often focused on simply meeting the minimum requirements set out in Approved Document B (Fire Safety) rather than focusing on the performance- based requirements set out in the regulations;
Most building control organisations (county councils and approved inspectors) will typically have standard workflows to be used in the checking process. However, due to the large amount of manual labour involved these processes are slow and prone to errors.
Difficulty of the compliance checking process and life-cycles of assets.
The processes that drive compliance with building safety requirements are weak and complex with poor record keeping and change control;
There are around 75 performance-based requirements, relying on sophisticated judgements for their interpretation. This places increased reliance on the competence of those undertaking the design and construction of buildings and the skills and rigour of the regulators verifying the quality of the work that is done;
The relationship between performance regulation and the life cycle of a building should be explored in greater depth. Whereas many buildings undergo a change of use over their life span (sometimes several times), one should question how tightly coupled the performance requirements should be to the original intended use of the building (and who monitors and pays for later changes).

"[…] the current regulatory system for ensuring fire safety in high-rise and complex buildings is not fit for purpose."

Few studies investigate regulatory enforcement regimes themselves [92]. And there is currently no consensus on how to implement the required fundamental changes. As an example, user research indicates a desire for prescriptive regulations, at least for certain topics [69]. But the UK government, as noted by Dame Hackitt [36], wants to move towards an outcome-based approach where guidance would be developed and maintained by industry and validated by a "*Building Safety Regulator*". For some regulatory topics, the prescriptive approach makes little sense, e.g., sustainability. Prescriptive guidance on building performance with respect to climate impact, considering buildings are responsible for over 40% of carbon emissions, is difficult to maintain, especially in light of advancing technologies and changes in overall objectives. Further concerns with prescriptive regulations include that they may stifle creativity and innovation, and the guidance documents may be treated as '*de facto*' proof of compliance.

Despite a lack of consensus on how to overhaul the building regulations regime, it is clear that digitalisation will play a large role [70]. There has been a push towards digitalising the Architecture, Engineering and Construction (AEC) sector, with strong support from both policy makers and national governments.[1] As an example, in the UK digitalisation is a crucial component of the £600 billion construction sector deal.[2] This translates directly into the availability of research funding to rely on digital technologies with the aim of tackling challenges and transforming the industry to improve efficiency, sustainability, productivity and longevity.[3] UK and international examples of large-scale research initiatives include:

D-COM Network[4]: *"Drive forward the adoption of the digitization of regulations, requirements and compliance checking systems in the built environment"*
The ACCORD Project[5]: "Digital building permit and compliance verification"
DigiChecks[6]: "A new Digital Framework to manage permits and compliance checks in the construction industry"
CHEK[7]: "Change toolkit for digital building permit… an innovative toolkit supporting the digitalization of building permit issuing and automated compliance checks"

Overhauling the building regulations regime is not necessarily part and parcel of any plan to digitalise the building permit issuing process. Yet regulations are central to the overall permit issuing process, and digitalisation of the regulations is thought to improve consistency and structure of regulatory documents[52]. Therefore, we believe that strategies to digitalise the building permit process should take into consideration the need to overhaulthe regulations regime. In devising such a strategy, it is worth noting that large-scale initiatives, that include a broad selection of stakeholders, have an advantage over smaller studies.

Devising a strategy for digitalising the building permit process is not straightforward. On the one hand, a major issue is the extremely wide scope of such a strategy: the breadth of involved sectors, the range of processes and responsibilities, the diversity of topics covered, the number of regulatory and guidance documents, complexities in the structure of documents, as well as the variety of organisations responsible for authoring and maintaining the documents. For example, in 2020 the Building Regulations for England were supported by statutory guidance in 16 Approved documents, which referenced nearly 500 Standards, 85 other Government guidance documents and 176 industry guidance documents [62]. Since then, two further Approved Documents have been released. The myriad of potentially conflicting requirements make it impractical to come up with a detailed all-encompassing strategy. Yet, to some extent, an all-encompassing strategy is required to ensure that research towards solving sub-tasks can be aligned – this alignment is required both for the various studies on a specific sub-task, as well as between the bodies of research that exist for each sub-task. As an example, it is not possible to synthesise the current body of work on converting regulations to computer-processable rules – one reason being that there is no standard approach to formatting these rules [10,103].

Devising a strategy for digitalising the building permit process requires either a helicopter view of all sub-tasks and their requirements, or a set of solutions that are flexible enough to be adapted in further studies. In this paper, we investigate the latter, and describe the

[1] https://ec.europa.eu/docsroom/documents/45547/attachments/1/translations/en/renditions/native.
[2] https://www.gov.uk/government/publications/construction-sector-deal/construction-sector-deal.
[3] https://www.ukri.org/what-we-do/our-main-funds-and-areas-of-support/browse-our-areas-of-investment-and-support/transforming-construction/.
[4] https://www.dcom.org.uk/.
[5] https://accordproject.eu/.
[6] https://digichecks.eu/.
[7] https://chekdbp.eu/.

development of several resources and tools that can be integrated easily with adjacent and complimentary solutions. The scope of the research reported in this paper is limited to processing the texts found in the full corpus of building regulations and guidance documents, after proposing a platform-based approach that aims to support development of solutions for all the sub-processes in the compliance process. The processing of texts in the corpus of regulatory documents is central to development of the proposed platform, and hence our focus on it as a priority. We believe the other sub-processes (such as conversion of retrieved requirements to computer-processable rules) become more 'tractable' once all requirements have been retrieved and challenges of ambiguity and interpretation found in the corpus addressed. Documents are processed in their current form, but the tools are not restricted to PDF input. Their end-use may cover wide variety of project stages, project scales, as well as a broad range of potential users; such as guidance authors, designers and building inspectors.

Section 2 describes the envisaged platform strategy to digitalisation of building regulations process and the scope of this research. Section 3 introduces related work that motivates our general approach. Section 4 describes our approach to the development of meta- and domain-specific resources and findings to date. Section 5 discusses work supporting development of user-facing tools and section 6 presents conclusions and further research opportunities.

We show that a semi-automated approach to semantic enrichment of building regulations is possible, and can already help improve the ability to identify documents and sections that deal with a specific topic.

2. Scope and aims

2.1. A platform strategy for digitalisation of regulatory compliance

The overall building permit issuing process may be broken down into many sub-processes that, themselves, are complex and multi-disciplinary [12]. Not all of these sub-processes may be automated easily, nor should they in our opinion. With regards to overhauling the building regulations, one could consider the potential benefits of developing a dedicated regulatory document authoring and processing environment. Table 2 exemplifies further tools that may be developed in light of digitalising the authoring, use and checking of regulations.

Tools such as those listed in Table 2 may be developed in isolation. But in a platform approach, a range of tools can benefit from shared standards and resources, reducing development time and complexity [66]. A schematic of the proposed "Intelligent Regulatory Compliance" platform, i-ReC, is shown in Fig. 1. This visualizes how such a platform strategy distinguishes between (1) a layer of user-facing tools that rely on (2) a shared *meta*-layer of standards and resources. The aim of proposing a platform approach is the development of semantically rich, stakeholder-focused and unambiguous tools. The corpus of regulations includes the current and previous versions of building regulations, related building and health and safety legislation and the guidance set out in the Approved Documents and referenced second tier documents, such as standards and codes of practice. Identification of documents for inclusion within the corpus and managing new, withdrawn and revised documents will be facilitated by a set of corpus document manager tools. These tools will need to support processes such as maintaining an audit trail of changes, identifying and managing the impact of changes on other regulations within the corpus, identification of new concepts and updating the knowledge graph and providing change and impact information to user-facing tools and ongoing projects.

User-facing tools may support a range of disciplines including authors, designers and building inspectors, at all scales and project stages. The primary aim of such tools is to improve the usability of building regulations and building data [48] – which in turn promotes a safer and more sustainable built environment. To ensure that these user-facing tools can interact with each other, there is a need for shared standards and resources. An example would be to enable the sharing of information between systems that use slightly different terminology. For Building Information Models (BIM) such interoperability is currently provided through domain ontologies, such as the Building Topology Ontology (BOT) [77]. But these domain ontologies are not comprehensive enough and their classification schemes do not align with the terminology found in all the building regulations [49]. Both these issues limit the re-usability of such domain ontologies for an application like Automated Compliance Checking (ACC). Development of the proposed Knowledge Graph as a shared resource could address this as well as supporting other user-facing tools such as those shown in Table 2. Semantic enrichment of the regulations could facilitate improved information retrieval, both in the form of search, and through structured information requests created from user tools (e.g. BIM models) that retrieve only context-sensitive, relevant information from the entire corpus of regulations and guidance.

The central idea behind this model is to provide all stakeholders with an open access facility with an intelligent semantic search facility as well as a reasoning engine to identify inconsistent, conflicting and ambiguous terms. Such a platform will enable various stakeholders to retrieve relevant requirements from the corpus of regulatory documents whilst also being able to integrate their own tools to accomplish their objectives as required. The proposed approach is indeed an ambitious one but it promises to address many challenges outlined before in a unified way and work as a one-stop-shop for the various stakeholders ranging from the authors, designers, design checkers, constructors to end clients.

Table 2
Overview of the types of computational tools the authors envision in support of authoring, use and checking of building regulations.

Authoring	Use	Checking
• identifying related sections and regulations to compare and align	• user-interfaces that provide a comprehensive overview of relevant regulations, guidance or previously accepted solutions	• track and understand historical decisions through an audit trail
• Identifying relevant standards and other documents that could be invoked	• easing the access to closely related regulations	• identify regulations that require human intervention, e.g., ambiguous requirements
• suggesting appropriate templates to capture the regulation as computer-processable rules	• semantic enrichment of texts to improve understanding of structure and terminology	• improve the transparency of both manual and automated rule-checking (white-box validation)
• suggesting and linking terminology to ensure consistency • automatically labelling how easy it is to interpret parts of the regulatory text.	• linking regulations to relevant best practices	• aligning inspections with known failures or issues

2.2. Natural Language Processing approach

There is a need to investigate the sub-processes around digital permits in sufficient detail, and a need to integrate the tools that are being developed to support these sub-processes[12]. In this study, we decouple the tools from the specific sub-processes. Our aim is to explore the development of Natural Language Processing (NLP) tools and resources that support a wide range of user-facing tools. The reason is that building regulations are currently captured primarily in text, alongside tables and diagrams. As such, NLP tools can play a central role in supporting and achieving the necessary overhaul of building regulations and is central to the proposed platform approach.

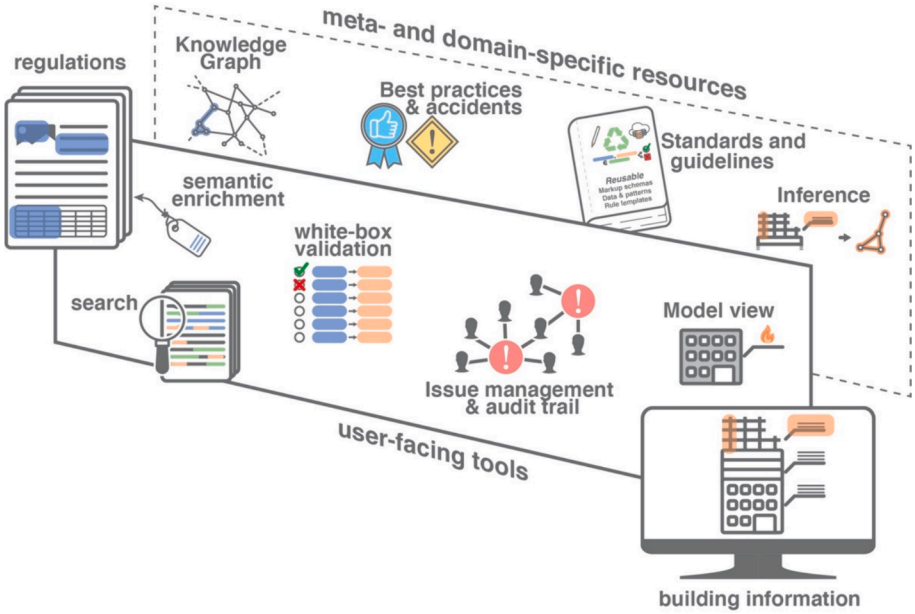


Fig. 1. Schematic overview of architecture of i-ReC, the envisioned platform strategy. Many user-facing tools would benefit from shared open standards and resources, e.g., datasets, shared vocabularies, guidelines, conversion algorithms and so on. In this paper we focus on NLP tools to process regulations, both from the perspective of meta- and domain-specific resources as well as the perspective of user-facing tools.

Developing a platform that encompasses meta- and domain-specific resources and user-facing tools, such as those listed in Table 2, can be relevant whenever text is encountered in the wider AEC domain. Examples include processing design briefs, change requests, accident reports, building product information sheets, contracts in the AEC domain, and so on. Our short-term goal is to improve the usability of building regulations, which we believe is fundamental to overhauling the regulations regime. The '*moonshot goal*' is to achieve full-fledged ACC – which we believe to be an unsolvable problem as explained in the subsequent section.

3. Related work

3.1. Automating compliance checking

Automating the compliance checking process has been researched for half a century [21,67,10,4]. Initial efforts towards ACC, most notably by Fenves [29] and his team, devised a regulatory provision representation scheme in the form of decision tables. The compliance process is facilitated by navigating through several of these linked decision tables, which encapsulated the logic enshrined in the regulatory documents. Subsequently, during the '80s and '90s, the effort moved towards applying Artificial Intelligence (AI) based tools and techniques. Rule-based compliance checking [33,78,51]acquired the centre- stage for a considerable period of time as it lent itself almost naturally to representing regulatory provisions, which were largely a collection of prescriptive rules at the time.

Also, in more recent years there is a plethora of research conducted in the realm of ACC and related sub-tasks [11,35,107,24], such as semantic annotation of regulations in support of converting text to rules [41,99,56]. With the emergence of new computational techniques and methods, novel approaches have been tried and tested. In recent developments, compliance checking has seen a move from rule-based approaches to Machine Learning (ML) based approaches [98,100,105,94,32], and Semantic web technologies have also been applied to compliance checking [13]. The premise of automating Compliance Checking (CC) includes the potential increase in efficiency and cost-effective labour, reductions in the number of errors and inconsistencies that are associated to manual checking, as well as improvements in the productivity and ease of customisation and innovation [26,21,68,4]. But much of the research towards ACC is only loosely connected to the overall permit issuing process, and does not take into account the practical integration with building permitting [12]. It may come as no surprise then that, despite considerable progress in the field of ACC, gaps remain in [67,10,4,70]: modelling building information, standardised approaches to modelling and aggregating rules, quality assurance and control of represented knowledge, the coverage of conceptualisations, changing the mindset of public officers, the assignment of legal responsibility for compliance, transparency of compliance assessments, standardised data and criteria for assessments on social, environmental and economic impact, and so on.

3.2. Natural Language processing applications in building regulations processing

ACC clearly remains an open research problem. One of the underlying challenges is the conversion of regulations to computer-processable rules [48]. The complexity of this conversion is compounded by the structure of regulations, as well as the way requirements are authored [103]. If regulations were authored in a computer-processable format, the role of NLP in ACC might be insignificant. But because building regulations are primarily presented as text, alongside tables and figures, NLP promises an essential part of proposed solution routes towards ACC. A few studies focus on developing a complete end-to-end ACC system that converts text to rules and applies the rules to building information models, examples include [98,104]. But the majority of solutions try to solve a smaller part of the ACC puzzle limiting the scope to a single document or indeed a small part of a document, with Information Extraction (IE) often being a central component [30].

3.2.1. Information extraction from regulatory documents

In a broader sense, IE revolves around extracting structured information from unstructured sources, such as text [82]. The type of extracted information may be limited to Named Entity Recognition

(NER) – note that strictly speaking domain terminology differs from Named Entities [46][8] – or could include the extraction of relations between informative words and phrases, attributes, constraints, and so on. IE over building regulations can be a goal by itself, e.g., a vocabulary of extracted domain terms may be used to support Information Retrieval (IR) [53,17]. Another aim may be semantic markup, e.g., denote the presence of specific domain terms, quantities, properties, and so on, in order to support down-stream labelling and parsing tasks [101]. Extracted information may be used to emphasise the salient parts and terms in a regulation, e.g., to determine the similarity between sections [52]. But grouping of related sections may also be based on external vocabularies, such as classifying regulations into a predefined set of classes, e.g., the classes found in a specific domain vocabulary [106]. Similarly, one could map extracted entities and other types of information to the classes in some domain vocabulary, e.g., in support of converting regulations to a set of rules that facilitates ACC [107]. Finally, IE may constitute the entire task of extracting a machine-understandable representation of the regulations described in text which is called semantic parsing.

3.2.2. Semantic parsing

Converting text to rules, or some other computer-executable logic form, is called semantic parsing in the field of NLP [65,5]. In this section, we will consider and exemplify the two paradigms that one might consider to be approaching semantic parsing[18],

- The traditional rationalist paradigm would be to prepare deterministic rules for converting text to code.
- The empiricist paradigm relies on statistical methods in the hope of learning reusable conversion patterns from examples.

A rule-based approach might consider the syntax of a text and the presence of certain keywords [45]. In a rule-based approach, a series of predefined transformations (conversion rules) are used so that text can be converted to some computer-processable format (compliance rules). But the manual programming of such conversion rules is a tedious task and, therefore, rule-based approaches are hard to scale. The identification and validation of conversion rules may be alleviated by relying on existing syntactic parsers and lexical resources, such as WordNet [108\, PropBank [43,109], VerbNet [83], FrameNet [80,7] and so on. A major obstacle is that parsers and lexical resources are of (extremely) limited use beyond the domain they were prepared for, and few-to-none were prepared with ACC in mind. This is particularly problematic because rule-based systems do not handle ill-formed or incomplete input well. The lack of domain resources for building regulations also impedes research towards speeding up the identification of potential syntactic patterns of interest through semi-supervised approaches, such as bootstrapping [14,1,95], or automated approaches based on simplifying assumptions, such as distant supervision [6,86].

A recent strong example of a rule-based approach to ACC is [79]. The authors rely on a *meta*-linguistic annotation framework called 4lang [44], which is language- and domain-independent. In earlier work they wrote a parser[9] for this framework that produces semantic graphs from which compliance rules are derived. The system's design is motivated by, and based on a deep knowledge of linguistic theory – which is decidedly uncommon in the domain of ACC. Despite the relative robustness of the approach, only a small and relatively simple set of regulations is processed. As a consequence, the authors claim that they can leave out-of-scope the handling of nested deontic operators and more complex deontic logic, such as exceptions [54]. In the domain of ACC it is not uncommon to see rule-based approaches that are restricted to very small datasets and specific problems [64].

The empiricist paradigm avoids *meta*-linguistic modelling, the presumption is that all necessary information for the analysis of text can be gleaned from text itself [42]. This paradigm includes Machine Learning (ML) approaches, which are more robust for ill-formed and incomplete inputs. But, usually the ML models are trained in a supervised fashion, which requires a substantial number of training examples before their performance is close to being acceptable. Due to the complexity of semantic parsing, it is difficult and costly to collect training data for supervised systems [16,40].

A recent example is [31]. This study uses a dataset of 718 samples that were carefully selected [32] from 10K rules represented in LegalRuleML [23]. These 10K rules were manually converted from a selection of regulations, the conversion effort took a team of 6 experts in relevant domains 6 months [22]. The ML model relies on a pre-trained general purpose Large Language Model – T5 [75] – and the authors explore improving results by easing the training process in several ways. While the model achieves relatively good results on the dataset, the authors note that the outputs are initial conversions that can be improved by experts. The model may not generalise well beyond regulations seen during training, e.g., documents that deal with different topics or use different wording. Notable findings include that results improve greatly (18.4%) through manual simplification of the legal clauses. Another significant (4.6%) improvement can be achieved by helping the model focus on extracting entities correctly – the authors suggest using external resources like dictionaries.

In conclusion, the scope of semantic parsing for ACC remains limited – regardless of both paradigm (rationalist, empiricist or combination) and state-of-the-art techniques used. But the primary obstacle is that building regulations, in their current PDF-based form, are simply too complicated to process. Sources of complexity include processing PDFs in the first place; handling Multi-Word Expressions (MWE), lists, figures, tables; references to other sections and documents and so on.

3.2.3. Is end-to-end ACC desirable?

Some requirements may simply not be amenable to be captured as a computer-processable rule. And a shift towards performance-based regulations is likely to increase the proportion of such rules. As an example, Malsane et al. [57] consider the following sentence from the Approved Documents (England):

> "The window or door should enable the person escaping to reach a place free from danger from fire."

Determining whether a place is '*free from danger of fire*' requires more than extracting information and re-formulating a sentence to computer-processable format. Rather, capturing what it means to be '*free from danger of fire*' requires at least some understanding of the mentioned concepts, context and relevant variables. The presence of such hidden ambiguities, assumptions, dependencies, and exceptions increases the complexity of implementing regulations as rules [88].

Processing of regulations becomes more complex when considering that sentences are inter-related through references, paragraphs, sections and documents, as well as tables and diagrams – also see section 4.1. It is because of this complexity that we believe that: fully converting all building regulations, in their current form, to computer-processable rules requires Natural Language Understanding (NLU). NLU is an AI-hard problem [97], meaning that to solve this problem is akin to solving the problem of general artificial intelligence. This roughly means that complete semantic parsing of building regulations is impossible. However, if the parsing and checking of simpler rules is automated, experts may focus their valuable time on more complex issues [88]. It can, thus, be helpful to develop semantic parsing solutions, and their effectiveness may be supported through NLP solutions that (1) identify regulations that may be processed and (2) break down slightly more

[8] Named Entities are, strictly speaking, real-world `*named*` objects, such as a person, location, organization etc. Domain terminology will need to encompass a broader set of labels, including e.g., `*ventilation strategy*` or a `*a place free from danger of fire*`.

[9] https://github.com/adaamko/wikt2def.

complex regulations so that they may be parsed, e.g.,[100].

A fundamental requirement of rules derived from regulations is that they remain human-interpretable – sometimes referred to as '*white-box*' within the context of CC [74]. The authors point out that fully automating the digital permit issuing process may not make sense at all. The reason is that the application and enforcement of regulations requires that responsibility can be appointed to someone – a computer cannot be held responsible for wrongly implementing a rule that was derived from a regulation.

4. Work towards meta- and domain-specific resources

This sections describes thoughts on and work towards meta- and domain-specific resources. The aim is to support digitalisation of the building regulations and user-facing tools that interface with building regulations, including tools for ACC. First, we outline what we believe are the main sources of complexity when processing regulations. Second, we focus on the complexity of compiling a domain lexicon – the set of defined classes and relations that, e.g., an ACC system relies on when composing rules. Third, we describe our work towards assembling such a lexicon and some of our findings.

4.1. Complexity of computer-processability of regulations

Most of the studies related to ACC often focus on single sentences that were manually selected from a small selection of chapters or sections within a narrow regulatory sub-domain. As an example, Zhang and El-Gohary [99] limit their scope to 3 chapters from the International Building Code and only extract requirements that include and demand measurable quantity or the presence of some building element. Reasons for reducing the scope may include the prevalence and complexity of certain types of requirements, as well as the research simply focusing only on a specific sub-domain [106]. In this section, we consider some notions of complexity from the perspective of NLP in relation to regulations. To exemplify what a simple quantitative regulation might look like, consider the following fictional example:

> Simple example: "The minimum distance of the door to the sink must be 120 cm."

A non-fictional example that is slightly more complex [102]:

> More complex example: "In areas where the average daily temperature in January is 25F or less or where there is a possibility of ice forming along the eaves causing a backup of water, an ice barrier that consists of at least two layers of underlayment cemented together or of a self-adhering polymer-modified bitumen sheet shall extend from the lowest edges of all roof surfaces to a point at least 24 in. inside the exterior wall line of the building."

Based on these two examples, Table 3 lists several properties of sentences that influence how easy (or indeed difficult) it is to convert them to a computer-processable format. Note that the two example sentences do not contain references to other bits of information that can be found in the corpus of building regulations. In a sense they are self-contained, apart from the common-sense knowledge that may be required to interpret and process them. Many sentences in regulations are not self-contained and, instead, refer to tables, figures, lists of sentences, other sections or even entire documents. And, as noted in section 3.2.3 performance regulations, which are open to interpretation, are hard to capture as computer-processable rules. Besides, building regulations are constantly amended, so any rules derived from text will require frequent updates and checks.

The ease of processing a sentence may be gleaned from features that might be quantified, such as: the sentence length, the number of prepositions and domain-specific terms, the number of co-references within the sentence as well as references towards other sections, tables and figures. Other features that are harder to quantify include the number of ambiguous statements, e.g., see [103]forinsights into the different ways that ambiguity pervades building regulations. Examples of studies that investigate the ability to convert regulations to rules include:

- Uhm et al. [91] investigate 27 Request for Proposal documents and find that many of the requirements are hard to translate to computer-processable rules. The type of project, e.g., healthcare facility or courthouse, seems directly related to variation in the number of processable requirements – between 2 % to 55 %. They also investigate the types of nouns and verbs that occur, and mappings to domain vocabularies and types of functions required to validate compliance.
- Macit İlal and Günaydın [56] manually select 297 building-related clauses from a municipal regulatory document. In these clauses they identify 258 rules of which, following manual classification, 58 % are self-contained and can be converted to logical statements.
- Soliman-Junior et al. [89] manually identify 820 clauses in a Brazilian regulatory document on healthcare facilities. In these clauses they identify 1284 requirements, of which 54 % is classified as qualitative and 34 % quantitative. 63 % of the requirements are thought to be re-presentable as logical statements, but they also find that many of the concepts that occur in the requirements cannot be mapped to a domain lexicon.
- Soliman-Junior et al. [90] manually identify more than 3800 requirements from 5 UK healthcare design regulations and guidance documents, comprising Health Building Notes, Health Technical Memoranda and Building Regulations.. Following a variety of manual classification approaches they find:
  - 63% are qualitative and 35% are quantitative.
  - ~50% are subjective, necessitating human interpretation. The ambiguities also make it difficult to map these requirements to logical statements.
  - 47% of the requirements can be mapped to logical statements, and 99% of these exhibit low complexity (class 1 and 2 following [88].
  - More than 10% of manually selected requirements contain references to diagrams, tables, sections and documents. The number of references is thought to rely on the overall structure of regulations, and the level of detail that a document deals with – general design guidance tends to contain more references.
  - For one of the five documents, 28% of the requirements can be translated by a commercial text-to-rule conversion system. This equates to only 53% of the requirements that were thought to be easily converted to a logical statement.

A better understanding of the types and potential contents of regulations can help clarify and organise regulations [89]. Table 4 presents a high-level classification scheme for the complexity of regulations, loosely inspired by the above-mentioned studies and Solihin and Eastman [88]. The classification distinguishes between: the complexity of validating the requirements, the complexity of processing the various structures in which regulations are presented, and the complexity of capturing the concepts and classes that occur in regulations. Based on such a classification scheme one can gain insights into the overall complexity of regulations. Furthermore, one may consider training a classifier to aid the identification of regulations that might need editing.

If one randomly picks a section in one of the UK building regulations, it is easy to see that only a fraction of the texts are easily processable by a computer. To confirm this, we randomly select 10 sections of ~100 words from 420 British regulatory documents. Among these documents are codes of practice and guidance documents. Two domain experts manually categorise these 10 texts following Table 4. We find that:

- It is uncommon to find self-contained simple sentences that express single word terms and a requirement that is easy to validate. Even when a section is relatively simple to process considering one

Table 3
Example properties that affect the ease of converting a sentence to a computer-processable rule.

length	Generally the complexity of processing a sentence increases linearly with the length of the sentence.
structure	The complexity of processing a sentence is compounded by the presence of prepositions and prepositional phrases.
ambiguities	Complexity increases terminology and information is ambiguous, e.g., "*areas with an average daily temperature in month X* " may be different each year. Some classes of objects that should comply to a regulation are merely hinted at and may be defined by a number of variables and specific interactions.
uncommon terms	Processing uncommon terminology, such as domain terms, may require adaptation of rule-based approaches. With regards to ML-based approaches, terms that are not (often) seen during the training of general domain Large Language Models the embeddings of domain-specific terms can be expected to be poor.
MWEs	Multi-Word Expressions (MWE) are terms that consist of multiple words that, together, express a single unit of information, also see section 4.2. There are currently no reliable strategies to represent long MWEs accurately – usually the average of constituent embeddings is used to represent the entire string. As a result long MWEs of, e.g., more than 5 words, receive a weak representation that is relatively similar to the representation of any string of comparable length.

Table 4
Classifications scheme for the complexity of processing building regulations. We divide the complexity into 3 categories; the *lexicon* is the explicitly defined vocabulary of classes that are available when composing computer-processable rules, the *data validation* refers to the amount of validation required to prove that the building or product data aligns with the requirements of the regulation, the *text complexity* refers to the structure and arity of interrelated requirements.

Level	Lexicon	Data validation	Text complexity
1	All terminology consists of single words that can be mapped directly to known classes, e.g., 'door' is mapped to 'IFCdoor'.	The regulation applies to data that is explicitly available in BIMs, e.g., 'wall height' is likely to be captured in a BIM.	Simple regulations, e.g., a self- contained sentence with precisely defined terms and few to no prepositional phrases.
2	Some terminology consists of multiple words, and some terms require that new classes are added to the domain lexicon. Determining the correct semantics of words may require disambiguation, e.g., the 'head' of a door.	Need for simple arithmetic operations to derive the required data from BIMs, e.g., the straight line distance be- tween two points.	A single sentence that is hard to process, e.g., a long sentence that contains prepositions and prepositional phrases. The sentence may express conditional statements and exceptions.
3	Some terminology consists of multiple words. When adding the terms to the domain lex- icon, an explicit definition of some terms requires a combi- nation of classes, e.g., 'door in series' may be defined using a combination of classes and conditions – 'door', 'swing', 'space', 'distance', and so on.	Need for complex computations to derive the required data from BIM, potentially in combination with external knowledge. Examples include the computation of abstract concepts, such as 'line of sight', or implicit relationships, such as 'close to'.	Structurally complex regulations that contain, e.g., lists or multiple closely related sentences. Besides the need for co-reference resolution, a hierarchy of statements across sentences may need to be con– structed.
4	Adding some of the classes to the lexicon is complicated, e.g., determining a suitable label for events and other classes that involve many variables and interactions (time, location, temperature, and so on). An example is an area 'where there is a possibility of ice forming along the eaves causing a backup of water'. Another complication occurs when new technologies can require the re-definition of existing classes.	Performance based requirements, as well as some of the unintentional ambiguous regulations, necessitate manual checking. Manual CC may be supported through knowledge-based tools and various simple and complex computations over the data that is present in a BIM	Regulations spread over multiple sections or documents necessitate defeasible logic, where conclusions can be defeated on the basis of sub- sequent information [73]. The references introduce additional parameters, combinations and interactions that are hard-to- manage – bringing processing closer to reading comprehension (NLU).

category (at best we find level 2), complexities arise when considering the other two categories.

- When we consider the sections as a whole, rather than cherry-picking requirements that are easy to process, 9 out of 10 sections fall into level 4 for at least one of the categories. The remaining section falls into level 3 on each of the categories.
- We encounter multiple sentences that are longer than 50 words. A normal length for a sentence may be 15 to 20 words. 4 of our sections contain enumerations or lists of sentences. 6 of our sections refer to other sections or documents.

4.2. Towards a lexicon to map domain terminology

As well as dealing with the complexities of processing the sentences within regulations, it is fundamental to the platform-based approach to develop a lexicon of domain terminology. For example, a critical factor for ACC is the ability to determine, for every requirement, which parts of a BIM (building model) are being checked [90]. This requires that the vocabulary of terms used in building regulations can be automatically mapped to BIM objects, attributes, and so on. Beyond ACC, the ability to map BIM objects to domain terms can support a variety of user-facing tools, such as asset specification and management [3], Alani et al. [2].

A major obstacle is that the range of labels found in BIMs is currently too limited to represent all the terminology found in regulations [48]. This is because neither the BIM systems' native data structures nor the IFC interoperable exchange standard map directly and entirely on to the terms and objects found in the corpus of regulatory documents.

The building regulations are expressed using a large number and variety of domain terms [49]. Among the standards and codes of practice documents are a number of non-comprehensive vocabulary documents. Many regulatory documents also contain a section with terms and definitions, as well as an index of terms. It is relatively straightforward to extract these explicit lists of terms and definitions from most documents, but licensing restrictions limit sharing and re-using them.

We find over 8K unique defined terms in just 9 vocabulary documents with a total of 640 pages, e.g., a '*drop apron*'[10] is defined as a '*flashing (01) fixed vertically at the eaves (01) NOTE Found mainly on flat roofs (01)*'. On the other hand, in the 1,274 pages of the open-access Approved Documents (England) we only find ~ 300 unique defined terms – and some terms have different definitions across approved documents, such as '*wet room*':

- 'WC or bathroom compartment with tanking and drainage laid to fall to a connected gulley capable of draining the floor area when used as a shower.' (Approved Document M – Access to and use of buildings, Volume 1, 2015 edition.)

[10] Term 06 32,221 in BS 6100–6 2008.

- 'A room used for domestic activities (such as cooking, clothes washing and bathing) that produce significant amounts of airborne moisture, e.g. a kitchen, utility room or bathroom. For the purposes of Part F of the Building Regulations, sanitary accommodation is also regarded as a wet room.' (Approved Document F – Ventilation, Volume 1, 2021 edition.)

If the intended use-case for a lexicon is computational reasoning over the classes, this requires explicit and formal definitions of these classes. Comprising a single, all-encompassing ontology necessitates the resolution of any terminological and conceptual incompatibilities [87], such as the two definitions of '*wet room*' found in the Approved Documents (England). The need for case-by-case resolution can be reduced by providing an upper-level ontology, which prescribes the canonical entities to which sub-classes must be aligned. In the AEC domain the Building Topology Ontology (BOT) provides such an upper-level ontology. The purpose of BOT is to support BIM Maturity Level 3, which envisions seamless information exchange that is interoperable, distributed, web-based and interdisciplinary [77]. But in many cases a mapping between regulatory text and objects in a BIM does not require computational reasoning.

4.2.1. An informal lexicon

Informal approaches to capturing domain terms can be as simple as a finite list of domain terms, also known as a controlled vocabulary. The semantics of such a term-list may be enriched with definitions, e.g., in a glossary, as well as by adding some relations like synonymy, e.g., in a thesaurus. A taxonomy adds a class hierarchy, through informal or formal '*is-a*' and sometimes '*instance-of*' relations. In some cases, one might argue that a taxonomy is an ontology, but usually an ontology captures additional semantics, e.g., value restrictions, disjunctions, cardinality constraints, part-of relations, and inverse relationships. The formal semantic architecture of an ontology enables computational reasoning over the defined data structure and knowledge captured within [28], which allows checking for logical contradictions and implicit sub-and super-class relations, as well as classifying and retrieving instances[72].

An ontology-based approach to a comprehensive lexicon is complicated by a variety of characteristics that apply to the AEC domain; large size and/or low stability of a corpus, large range of domains and terminology covered, ill-defined terminology, some key users lack expertise in one or more domains covered, the need for curators with authoritative judgement [84]. Therefore, we argue in [48] that formally and comprehensively defining all AEC domain terms, properties and rules is a Herculean task that is never-ending as regulations are constantly amended. On the other hand, informal labels on BIM elements may suffice for a mapping to building regulations when an exact match of the label is present. And in many cases, simply adding synonymy relations may resolve terminological mismatches.

4.2.2. Linked data and semantics

Considering potential downstream applications, and their requirements with regards to a lexicon, it makes sense to capture the terminology as Linked Data [48]. Linking terminology to existing resources eases the reuse of knowledge captured in, e.g., an informal taxonomy like Uniclass [34] and ontologies like BOT. From the perspective of labelling BIM elements, these links may support choosing relevant labels, e.g., a user-facing tool may suggest an overview of prevalent terms found in the building regulations. From the perspective of validation, one can expect that some terminology benefits from more formal semantics like '*is-a*' and exact matches. This can help when a requirement applies to all sub-classes of some term. Similarly, a frame-based approach [80,7] may be required to capture the multitude of semantic relations expressed by more complex terms – the level 3 and 4 classes in the lexicon category of Table 4. Examples include events and classes that inherently rely on the interactions between variables, such as '*area where the average temp is lower than X* ' mentioned in section 4.1. We conclude that such semantics should be captured following a Linked Data approach, e.g., relying on SHACL[11] or SHEX[12] to capture compound classes that require relationships between building elements, simple calculations, and/or conditional statements [48,71].

4.2.3. Named entity Recognition

Rather than collecting a comprehensive controlled vocabulary from scratch, one might consider extending a large existing classification system like Uniclass [34]. Uniclass is already used to label some types of building information [3], so for these instances of labels the identification of text occurrences equates to mapping between text and labelled data. However, from the 15K classes found in Uniclass only 598 (4%) terms occur verbatim in the Approved Documents (England) [49]. The rest of the terms found in the Approved Documents (England) can be divided into three groups:

- A first group of terms is expressed in a slightly different surface form from their corresponding Uniclass label. A classifier may be trained to map these texts to their respective labels, but such a dataset requires examples. And collecting examples for all 15K classes – either manually or semi-automated – is comparable in terms of effort as linking each of the classes to corresponding text-based occurrences.
- A second group of terms are semantically similar to one or more labels found in Uniclass, but the mapping requires a bit of shoehorning. An example is the class '*hot finished hollow section member*' (Requirement 4.3.2 on unprotected members in BS 5950–8 (1990).) for which the closest Uniclass equivalent is '*Carbon steel hot-finished hollow sections*' (Pr_20_76_52_16).
- A third group of terms simply does not have an equivalent label in Uniclass, such as '*party wall*' [48].

It is not known how many labels need to be added to Uniclass to cover the terms in the second and third groups. Assuming that they make up a considerable number of terms with varying complexities, we follow the rationale of [25] and argue that existing vocabularies are not suited for semi-supervised NER approaches over the building regulations.

4.2.4. Discovering domain terms

It is possible to extract concepts in an unsupervised fashion, e.g., based on syntax and collocation[19,27][25]. Beyond phrase chunking, a highly relevant task to building a domain lexicon of technical terms is discovering Multi-Word Expressions (MWEs), i.e terms that consist of multiple words that, together, express a single unit of information [46] [20]. The identification of domain terminology is both broader in scope than MWE discovery – single words could be domain terms – and narrower, as non-technical MWEs are irrelevant [8,93]. Handling MWEs is a key issue for NLP systems [85,81,76] and has been a source of decreased performance in ACC studies. Examples that explicitly indicate the need for better handling of MWEs include automatically breaking down complex regulations [100], as well as automating semantic enrichment of building regulations [101]. To stress the importance of handling MWEs, consider that:

- ~80 % of the defined terms in the Approved Documents (England) consist of multiple words.
- ~94 % of the 15K Uniclass labels consist of multiple words.

4.3. Domain lexicon development: Methods and findings

In Kruiper et al. [49] we present an automated approach to automatically identify (1) which parts of a sentence may correspond to

[11] https://www.w3.org/TR/shacl/.
[12] https://github.com/shexSpec/shex/wiki/ShEx.

Table 5
Statistics for the Scottish Building Regulations corpus. The number of defined terms, word-level tokens and sentences found in the domestic and non-domestic regulations.

	Domestic	Non-Domestic	Total
Terms defined in definitions section	128	127	128
Defined terms in text after lemmatisation and lower-casing	233	247	292
Number of terms linking to definitions section	4,687	5,368	10,055
Number of tokens	131,666	151,499	283,165
Vocabulary	8,282	8,925	9,837
Number of sentences	6,313	7,293	13,606
Mean sentence (word-level token) length, excluding punctuation	20.86	20.77	20.81
Standard deviation	11.96	12.32	12.16

terms, and (2) which of these candidate terms belong to the AEC domain. The approach relies on our earlier work where we developed a shallow parser for the Scottish building regulations (see Table 5), SPaR.txt [46]. SPaR.txt is trained to determine which words in a sentence belong to the same unit of meaning, and is evaluated on identifying *multi-word* entities or *concepts*. An advantage over strictly unsupervised approaches, such as phrase chunking, is that SPaR.txt can handle discontiguous MWEs – such as *multi-word concepts* in the previous sentence or the example in Fig. 2.

We distinguish domain-relevance by comparing how often terms occur in the Approved Documents (England), and how often they occur in a background corpus [49,60]. The output is an automatically generated Knowledge Graph (KG) of candidate domain terms, where we automatically link the extracted terms to the concepts found in existing vocabularies like Uniclass. Fig. 3 shows an extract from the KG, visualizing terms relating to 'ventilation', created in the GraphDB software. Colours of nodes: (red) spans, such as '*vent*' and a subclass '*air vent*', (purple) concepts, such as the Uniclass term '*Trickle vents*' (Pr_30_59_94_90), (blue) primary source nodes, such as the SPaR.txt paper '*http://dx.https://doi.org/10.18653/v1/2021.nllp-1.14*'.

Both our SPaR.txt and the Text-to-KG tools fall in the category of meta- and domain-specific resources. The code, data and instructions for these working prototypes are freely available online,[13] so that the wider community may use them and expand upon them. The purpose of these tools is to support the study and development of user-facing tools, in order to better support users of building regulations [48].

4.3.1. Supporting manual composition of a lexicon

We initially explore the manual collection of terms for a KG. The manual approach provides us with a baseline, workflow and insight in the requirements of the KG. Our aim is to develop an informal taxonomy for three small sub-domains, where the class hierarchy supports organisation of terms rather than formally defining sub-class relations. Specifically, we rely on the Simple Knowledge Organisation System (SKOS) vocabulary [61] to capture hierarchy with skos:broader and skos:narrower. Domain experts developed each sub-domain taxonomy through evaluation of terminology within relevant industry standards, guidance, dictionaries, classification systems and product information, to extract terms and definitions. We find that the manual work is tedious and slow. Per hour, annotators add ~ 6 concepts and ~ 4 links to external resources to the KG. Issues include that annotators aren't sure whether terminology actually occurs in regulations, which means the KG terms may not provide a mapping to text. They also note the difficulty of gauging whether the collected terms comprehensively capture a sub-domain. And that determining the relevance of terms is easier when definitions are present, and when the source of terms is known to be reliable [49].

We conduct a small qualitative comparison, where manual KG curation is supported by our automatically generated KG [49]. Through visualisation of linked term candidates and searching in the graph, the speed of adding terms to the KG was increased 15-fold – a strong indicator that a user-facing tool for KG curation can greatly speed up this type of work by domain experts. Annotators find that it is helpful that the source and provenance of terms can be tracked in the KG, and that it is easy to identify related terms. Especially when definitions are present, even if such definitions are derived from a less reliable source like WikiData. Importantly, the approach to generate the KG is scalable for the most part. Only certain node–node metrics are harder to compute with an increasing number of nodes in the KG.

4.3.2. Identifying terms and relations

With regards to our approach to MWE discovery, we expect that general domain embeddings may perform equally or even better than embeddings derived from domain-specific texts. The reason is that SPaR.txt is a sequence tagger that relies more on latent syntactic properties. However, we find that general domain embeddings perform poorly on semantically oriented tasks, such as clustering and computing semantic similarity. Some terms may not occur often in general domain text, e.g., '*rybat*' or '*grout*', which reduces the ability to differentiate their representations from morphologically similar terms. However, we also find that many MWEs in the AEC domain are comprised of words that, individually, are commonly found in general domain texts. Examples include '*mortar snot*' and '*cloaked verge tile*'. An issue is that embeddings of sequences of tokens, such as any MWE, are usually composed of a weighted sum over constituent tokens [63]. This means that many of the very long domain terms, such as '*target primary energy rate*', receive a relatively weak representation that complicates matters, e.g., clustering or computing semantic similarities. Moreover, the compositional meaning of a term like '*green roof*' has little to do with the colour '*green*'. We find that having access to domain definitions provides additional information that can greatly improve representations,[14] improving the ability to suggest candidate relations for a KG [49].

Beyond suggesting relations based on similarity, it may be possible to extract relations between concepts from text, e.g., through (Semi-)Open IE [9,50]. However, during initial tests with Semantic Role Labelling (SRL) we find that the results are relatively poor due to the complexity of many sentences and noise stemming from PDF- extraction. Challenges in this area include the approach to align SRL arguments and KG concepts, the filtering of relations of interest, as well as validating that relations capture the expected semantics. As an example, a phrase like '*a dwelling that is part of a mixed-use building*' does not imply that a '*dwelling*' is always part of a '*mixed-use building*' – yet, when focusing only on the semantic roles for the verb '*is*' in this phrase, that is the implication.

All this points to the need for linked data that can facilitate the re-use of a large variety of existing resources, thus, enabling new ways of analysing building regulations and support the digitalsation of regulations processing in other ways.

The foregoing discussions establishes the need for linked data, which should enable the re-use of a large variety of existing resources, and facilitate new ways of analysing building regulations. This will also support the digitisation of regulations in other ways including the focus on supporting the domain experts and this should align with the requirements of FAIR (findability, accessibility, interoperability and reusability) data.

[13] SPaR.txt: https://github.com/rubenkruiper/SPaR.txt Text-to-KG: https://github.com/rubenkruiper/irec.

[14] For relevant code and linking domain terms, also see https://github.com/rubenkruiper/LDAC_BSDD_hackathon

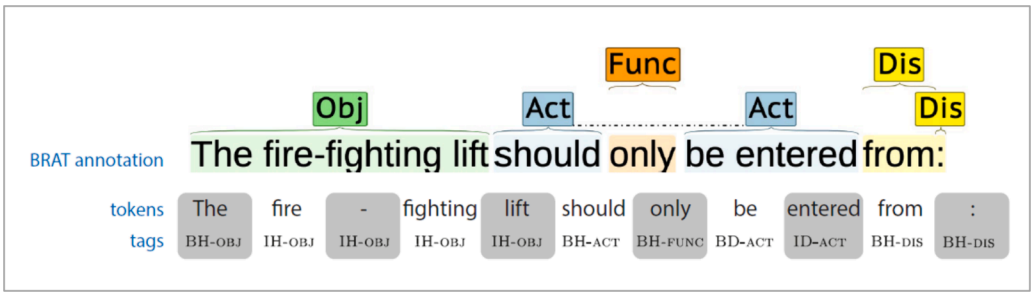


Fig. 2. Example of an annotated sentence. The determiner at the start of the OBJECT span is taken to be part of the span. A discontiguous ACTION span is interjected by a FUNCTIONAL span that modifies the Verb-Phrase. During training the sentence is tokenized and the aim is to predict the correct tags for each token.

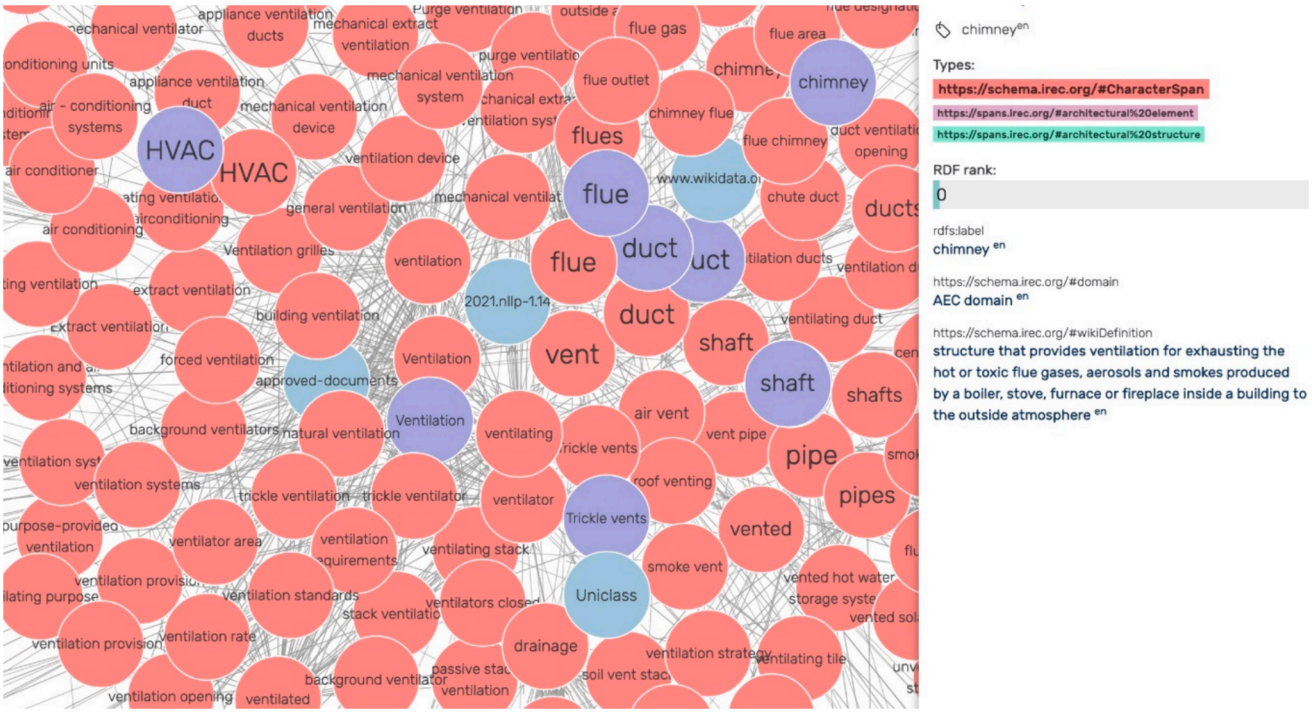


Fig. 3. KG extract visualizing terms relating to 'ventilation'.

4.4. Knowledge formalisation

Our work is fundamentally about processing engineering knowledge in the form of regulatory requirements. Our approach is driven by the failure of earlier approaches to represent this kind of knowledge formally as rules and other formalisms [29,33,78,51]. Those knowledge representation formalisms were found to be quite restrictive due to the nature of the knowledge being represented and processed, particularly since the adoption of a performance-based approach to authoring the regulatory requirements. In terms of crafting a rule-based processing of regulatory requirements, there is a huge amount of manual effort required to convert the content of the regulatory documents into rules due to the complexity of the semantics encapsulated because of ambiguities, use of synonyms and sometimes conflicting requirements among other things. This research presents an alternative approach for regulatory knowledge representation by utilising established computational methods (i.e. NLP/ML and Semantic Web) to extract and process the regulatory knowledge directly from regulatory documents. We have found this to be a more effective approach in the various evaluations of our work. In any case, NLP is associated with human intelligence and the basic theory of NLP is also highly related to the various knowledge-related theories. Ever since the earliest days of Artificial Intelligence [59], knowledge representation has played a significant role in the development history of NLP, mainly focusing on exploring symbolic knowledge representations and using symbolic systems to enable machines to understand and reason languages [38]. In addition to NLP, knowledge graphs have been used in this research to formally capture and represent regulatory knowledge to enable symbolic reasoning and processing of regulatory requirements. It is well established that knowledge graphs arose out of research in knowledge representation and reasoning, among other things [37].

5. Work towards user-facing tools

NLP plays a fundamental role in improving the usability of building regulations [48]. As described in section 3.2.2, some parts of the regulatory documents simply cannot be converted to a computer- processable formats. Therefore, one can reasonably expect that there will always be parts of the regulations captured as text, diagrams and tables.

5.1. Information retrieval

In this section we briefly describe our work on an IR system to improve search for building regulations [47]. Despite the importance of

access to relevant regulations, support for retrieval is inadequate [15,58,55]. One issue is that IR researchers in the AEC domain generally do not make their code or data publicly available. This makes it impossible to reuse their approach or compare results. Therefore, a crucial component of a general strategy towards ACC is the adaptation of the FAIR principles to ensure Findability, Accessibility, Interoperability and Reusability [110].

In our study, through two rounds of interviews we investigated how domain practitioners search for relevant regulations and what types of queries they use [47]. Practitioners note that they work with manually devised checklists and rely on the Approved Documents to navigate the large body of documents. In line with literature, we found that construction professionals have difficulties finding relevant documents [15]. One issue is that the retrieval solutions they employ simply do not provide enough insights into the relevance of results. Often the information need is both navigational, e.g., searching for a specific document, as well as informational, e.g., searching for the information content regardless of source [39]. For these reasons, we developed a passage-retrieval system that divides documents into 100-word passages that are indexed and retrieved separately.

The automatically generated KG (described earlier) makes it possible to quickly identify groups of related candidate terms, which enables query and document expansion.

As part of the study, we investigated the use of the KG to expand user queries, with the aim of improving the search results when similar terminology is used. Based on the interviews we developed a dataset of 42 queries and corresponding narratives that describe the desired information need. We found that the queries formulated by participants are often significantly longer than the average ~ 2 words of web-queries[15] [96]. With an average query length of ~ 6 words, the information need is often captured relatively well and we find that query expansion does not necessarily improve retrieval results. On the other hand, we found that document expansion can improve the recall significantly [47]. As such we recommend semantic enrichment of documents for effective semantic search.

Due to the way documents are indexed in the British Standards Online (BSOL) portal, this system often does not return results. It actually is unable to retrieve any documents for our 42 queries. Our prototype system is able to retrieve a relevant top-3 result for 35 of our 42 queries. Our user interface ranks document-level results based on the number of retrieved relevant passages in a document which showed that a user retrieved text passages with greatly improved ability to determine the relevance of retrieved results.

5.2. General remarks for user-facing tools

A balance has to be found between the needs and requirements of both authors and users of regulations. Checking the compliance with regulations sometimes requires information that isn't always captured in BIM [88].

This raises the question as to whether to place the burden of additional work on (1) those who create the 3D BIMs, or (2) those who develop processing of building data, in order to derive further information necessary for checking. In some cases, the additional information may be derived from information that is present in a BIM.

6. Conclusion

A platform-based approach for digitalising regulatory requirements processing spanning authoring, designing and compliance checking was presented. The proposed strategy allows collecting an overview of crucial requirements for isolated applications, that affect the development of related tools and resources. The proposed approach comprises of several tools integrated together as required with the processor of corpus of regulatory documents underlying the whole platform accessed by all the various tools. The paper also provided brief descriptions of the proof-of-concept implementation of some of the tools and associated approaches using NLP, ML and semantic web technologies that could form parts of the platform approach. These are SPaR.txt which identifies MWEs in the documents, which is utilised in an information extraction semantic search engine validated over a corpus of 420 used in the UK for regulatory compliance of designs. The automatic generation of knowledge graphs used by the search engines was also described as well as the query and document expansion techniques used by it.

Some of the areas that could be developed further include:

- The performance of SPaR.txt could be improved by increasing the dataset and providing the model with examples from more diverse formats of regulations. With the help of a curated set of concept candidates, one would want to focus on entity linking, that is the identification and disambiguation of concepts in text, and mapping them to a class in the KG automatically. Other tasks of interest are coreference resolution and relation extraction, including relations that occur across sentence boundaries. Considering the many references to other regulations, sections and documents, it is important to extract and model document structure in a meaningful way as well.
- There is a need to develop and align strategies to collect a comprehensive set of domain vocabularies, which we imagine to be combined in a KG.
  - o Such work would be project-based and may include terminology captured at different levels of formality. And so it would be important to have protocols in place to make sure information from different projects can be aligned easily. One example approach that may be of interest is bioschemas, (https://bioschemas.org/.) which aims to ease the markup of texts by defining (1) types and properties, and (2) shapes over these types and properties to enable validating their correct use.
  - o There is also the opportunity and requirement to develop tools that support the integration of new terminology into the KG, e.g., compare which (equivalent, related or similar) terms already are present and which are missing, as well as compare how terms are defined. One example of work in this area is to build some tools to ease the alignment of a project-vocabulary's terminology with the existing in vocabularies in bSDD. (https://github.com/rubenkruiper/LDAC_BSDD_hackathon)
- Future work towards Information Retrieval should aim to increase the size of the dataset. While our data suggests that semantic markup of regulations greatly benefits search, other approaches to document expansion should be developed and tested.
- In general, more time should be spent on cleaning the processed regulatory texts to improve results.

7. Limitations of the research

Many regulations are only accessible as PDF files, and correctly extracting the text from PDF documents is not straightforward. The diversity of formatting found throughout regulations further complicates this issue. In a selection of ~ 400 British Standards we find single column, two-column and three-column formats, and a large variety of positioning and formatting approaches for text, figures and tables. On top of this, licensing restrictions on most regulations prohibit sharing of data that either includes or is derived from their contents. For some tasks, such as ACC, licensing restriction also impede processing open-access guidance documents due to the many references to licensed documents.

[15] https://www.statista.com/statistics/269740/number-of-search-terms-in-internet-research-in-the-us/ [Accessed April 2023].

CRediT authorship contribution statement

Ruben Kruiper: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis. **Bimal Kumar:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization. **Richard Watson:** Writing – review & editing, Supervision, Investigation, Funding acquisition, Conceptualization. **Farhad Sadeghineko:** Validation, Investigation. **Alasdair Gray:** Supervision, Software, Investigation, Formal analysis. **Ioannis Konstas:** Software, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research is part of the intelligent Regulatory Compliance (i-ReC) project, a collaboration between Northumbria University and Heriot-Watt University. We are grateful to the Building Research Establishment (BRE), the Construction Innovation Hub (CIH), as well as Northumbria University for funding this research. The primary funders were CIH through BRE as their lead resaerch partner. The grant was originally made to Northumbria University and the grant ID was 120725.

References

[1] E. Agichtein, L. Gravano, J. Pavel, V. Sokolova, A. Voskoboynik, Snowball: a prototype system for extracting relations from large text collections, SIGMOD Rec. 30 (2) (2001) 612.
[2] Y. Alani, N. Dawood, J. Patacas, S. Rodriguez, H. Dawood, A semantic com- mon model for product data in the water industry, Journal of Information Technology in Construction 26 (2001) 566–590.
[3] Y. Alani, N. Dawood, S. Rodriguez and Dawood, H.. Whole Life Cycle Construction Information Flow using Semantic Web Technologies: A Case for Infrastructure Projects. In *Proc. 37th CIB W78 Information Technology for Construction Conference (CIB W78)*, pages 141–155, 2020.
[4] R. Amor, J. Dimyadi, The promise of automated compliance checking, Developments in the Built Environment 5 (2021) 100039. Elsevier.
[5] Y. Artzi, L. Zettlemoyer, Weakly Supervised Learning of Semantic Parsers for Map- ping Instructions to Actions, Transactions of the Association for Computational Linguistics 1 (2013) 49–62.
[6] N. Bach, S. Badaskar. A review of relation extraction. *Literature review for Language and Statistics II*, https://www.cs.cmu.edu/~nbach/papers/A-survey-on-Relation-Extraction.pdf, 2007.
[7] C.F. Baker, FrameNet: A Knowledge Base for Natural Language Processing, In Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore, Number 1968, pages1–5, 2014.
[8] T. Baldwin and S. N. Kim. Multiword Expressions. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing*, chapter 12, pages 267–292. Chapman and Hall, second edition, 2010.
[9] M. Banko, M. Cafarella, S. Soderland, M. J. Broadhead and Etzioni, O. Open information extraction from the web. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2670–2676, 2007.
[10] T.H. Beach, J.-L. Hippolyte, Y. Rezgui, Towards the adoption of automated regulatory compliance checking in the built environment, Journal of Automation in Construction, 118 (2020) 103285. Elsevier.
[11] T. Bloch, R. Sacks, Clustering information types for semantic enrichment of building information models to support automated code compliance checking, ASCE Journal of Computing in Civil Engineering (2020), https://doi.org/10.1061/(ASCE)CP.1943-5487.0000922.
[12] T. Bloch, J. Fauth, The unbalanced research on digitalization and automation of the building permitting process, Journal of Advanced Engineering Informatics 58 (2023) 102188. Elsevier.
[13] K.R. Bouzidi, B. Fies, C. Faron-Zucker, A. Zarli, N.L. Thanh, Semantic Web Approach to Ease Regulation Compliance Checking in Construction Industry, Future Internet. 4 (3) (2012) 830–851, https://doi.org/10.3390/fi4030830, 2012.
[14] S. Brin, Extracting Patterns and Relations from the World Wide Web, The World Wide Web and, Databases 53 (9) (1999) 172–183.
[15] T. Cerovsek, in: Advancing regulation retrieval with profiling, controlled vocabularies and networked services, IEEE, 2009, pp. 257–264.
[16] D.L. Chen, Fast online lexicon learning for grounded language acquisition 1 (2012) 430–439.
[17] C.P. Cheng, G.T. Lau, K.H. Law, J. Pan, A. Jones, Regulation retrieval using industry specific taxonomies, Artificial Intelligence and Law 16 (3), pages 277–303, 2008.
[18] K. Church, A Pendulum Swung too Far, Linguistic Issues in Language Technology 2 (4), pages 1–26, 2007.
[19] M. Collins, Y. Singer, Unsupervised models for named entity classification, in: *1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora*, 1999.
[20] M. Constant, G. Eryiğit, J. Monti, L. Van Der Plas, C. Ramisch, M. Rosner, A. Todirascu, Multiword expression processing: A survey, Comput. Linguist. 43 (4), pages 837–892, 2017.
[21] J. Dimyadi and R. Amor, Automated Building Code Compliance Checking. Where is it at? In *Proceedings of the CIB World Building Congress 2013 and Architectural Management & Integrated Design and Delivery Solutions (AMIDDS)*, number 380, pages 172–185, 2013.
[22] J. Dimyadi, S. Fernando, K. Davies, R. Amor, Computerising the New Zealand Building Code for Automated Compliance Audit. 6th New Zealand Built Environment Research Symposium (NZBERS2020) 6 (2020) 39–46.
[23] J. Dimyadi, G. Governatori and R. Amor, Evaluating legaldocml and legalruleml as a standard for sharing normative information in the aec/fm domain. In *Proceedings of the Joint Conference on Computing in Construction (JC3)*, volume 1, pages 637–644. Heriot-Watt University, Edinburgh, UK. Heraklion, Greece, 2017.
[24] O. Doukari, D. Greenwood, K. Rogage, M. Kassem, Object-Centred Automated Compliance Checking: a Novel, Bottom-Up Approach, Journal of Information Technology in Construction 27, pages (2022) 335–362.
[25] D. Downey, M. Broadhead, O. Etzioni, Locating complex named entities in web text, IJCAI International Joint Conference on Artificial Intelligence, pages 2733–2739, 2007.
[26] C. Eastman, J.-M. Lee, Y.-S. Jeong, J.-K. Lee, Automatic rule-based checking of building designs, Journal of Automation in Construction 18 (2009) 1011–1033. Elsevier.
[27] O. Etzioni, M. Cafarella, D. Downey, A.M. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates, Unsupervised named-entity extraction from the Web: An experimental study, Artif. Intell. 165 (1) (2005) 91–134.
[28] D. Fensel, Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce, 2nd edition., Springer, 2003.
[29] S.J. Fenves, Tabular decision logic for structural design, ASCE Journal of Structural Division 92 (6) (1966) 473–490.
[30] S. Fuchs, Natural Language Processing for Building Code Interpretation: Systematic Literature Review Report, Technical Report May, University of Auckland. (2021), https://doi.org/10.13140/RG.2.2.29107.55845.
[31] S. Fuchs, J. Dimyadi, M. Witbrock, and R. Amor, Improving the Semantic Parsing of Building Regulations through Intermediate Representations. In Proceedings of 30th annual meeting of *EG-ICE*, pages 1-10, London, 2023.
[32] S. Fuchs, M. Witbrock, J. Dimyadi, R. Amor, Neural Semantic Parsing of Building Regulations for Compliance Checking, IOP Conference Series: Earth and Environmental Science 1101 (9) (2022).
[33] J. Garrett,, A Knowedge-based Standards Processor for Structural Component Design, PhD Thesis, Carnegie-Mellon University, 1987.
[34] J. Gelder, The principles of a classification system for BIM: Uniclass 2015 1 (2015) 287–297.
[35] D. Greenwood, S. Lockley, S. Malsane and J. Matthews, Automated compliance checking using building information models. In: *The Construction, Building and Real Estate Research Conference of the Royal Institution of Chartered Surveyors*. RICS, London, 2010.
[36] J. Hackitt, Building a Safer Future - Independent Review of Building Regulations and Fire Safety: Final Report (issue December). 2018 ID CCS1117446840.
[37] C. Gutierrez, J.F. Sequeda, Knowledge Graphs: Tracking the historical events that lead to the interweaving of data and knowledge, Communications of ACM 64 (3) (2021) 96–104, https://doi.org/10.1145/3418294, 2021.
[38] X. Han, W. Chen, Z. Liu, Y. Lin, M. Sun, Knowledge Representation Learning and Knowledge-Guided NLP, in: Z. Liu, Y. Lin, M. Sun (Eds.), Representation Learning for Natural Language Processing, Springer, Singapore, 2023, https://doi.org/10.1007/978-981-99-1600-9_9, 2023.
[39] B. Hedin, S. Tomlinson, J.R. Baron, D.W. Oard, Overview of the TREC 2009 legal track, NIST Spec, Publ (2009) 1–9.
[40] J. Herzig, J. Berant, Neural semantic parsing over multiple knowledge-bases, 2, Association for Computational Linguistics (ACL, 2017, pp. 623–628.
[41] E. Hjelseth and N. Nisbet, Capturing Normative Constraints By Use of the Semantic Mark-Up Rase. In Proceedings of *CIB W78-W102 2011: international conference, Sophia Antipolis, 25 October 2011: CIB W78, Sophia Antipolis, pp.26–28, 2011*.
[42] D. Jones, Non-hybrid Example-based Machine Translation Architectures. In *Proceedings of TMI-92*, pages 163–171, 1992.
[43] P. Kingsbury and M. Palmer, From TreeBank to PropBank. *Proceedings of the Interna- tional Conference on Language Resources and Evaluation (LREC)*, pages 1989–1993, 2002.
[44] A. Kornai, J. Acs, M. Makrai, D. M. Nemeskey, K. Pajkossy, and G. Recski, Competence in lexical semantics. In *Proceedings of the fourth joint conference on lexical and computational semantics*, pages 165–175, 2015.

[45] R. Kruiper, Computer-aided biomimetics: semi-open relation extraction from scientific biological texts, Heriot-Watt University, 2020. PhD thesis.

[46] Krupier, R., Konstas, I., Gray, A., Sadeghineko, F., Watson, R., & Kumar, B. (2021). SPAR.txt, a cheap Shallow Parsing approach for Regulatory texts. In N. Aletras, I. Androutsopoulos, L. Barrett, C. Goanta, & D. Preotiuc-Pietro (Eds.), *Proceedings of the Natural Legal Language Processing Workshop 2021* (pp. 129-143). Association for Computational Linguistics (ACL). https://aclanthology.org/2021.nllp-1.14/.

[47] Krupier, R., Konstas, I., Gray, A. J. G., Sadeghineko, F., Watson, R., & Kumar, B. (2023a). Document and query expansion for information retrieval on building regulations. In *Proceedings of the 30th EG-ICE: International Conference on Intelligent Computing in Engineering* (pp. 1-12). University College London. https://www.ucl.ac.uk/bartlett/construction/research/virtual-research-centres/institute-digital-innovation-built-environment/30th-eg-ice-1.

[48] Krupier, R., Konstas, I., Gray, A. J. G., Sadeghineko, F., Watson, R., & Kumar, B. (2023b). Don't shoehorn, but link compliance checking data. In *LDAC 2023: Linked Data in Architecture and Construction 2023* (CEUR Workshop Proceedings). CEUR. https://linkedbuildingdata.net/ldac2023/abstracts.html.

[49] Krupier, R., Konstas, I., Gray, A., Sadeghineko, F., Watson, R., & Kumar, B. (2023c). Taking stock: a Linked Data inventory of Compliance Checking terms derived from building regulations. In *LDAC 2023: Linked Data in Architecture and Construction 2023* (CEUR Workshop Proceedings). CEUR. https://linkedbuildingdata.net/ldac2023/abstracts.html.

[50] R. Kruiper, J.F.V. Vincent, J. Chen-Burger, M.P.Y. Desmulliez, I. Konstas, In Layman's Terms: Semi-Open Relation Extraction from Scientific Texts, In *arXiv Preprint* (2020) arXiv:2005.07751.

[51] B. Kumar, Knowledge Processing for Structural Design, PhD Thesis, Edinburgh University, 1989.

[52] G. T. Lau, K. H. Law and Kumar, B. A regulatory information infrastructure with application to accessibility codes. In Coleman, R., McDonald, A., and Hamlyn, H., editors, *Proceedings of Include 2003*, London, 2003.

[53] G.T. Lau, K.H. Law, G. Wiederhold. Legal information retrieval and application to E-rulemaking, ACM Press, New York, New York, USA, 2005, pp. 146–154.

[54] T. Libal, A meta-level annotation language for legal texts. In *Logic and Argumentation: Third International Conference, CLAR 2020, Hangzhou, China, April 6–9, 2020, Proceedings 3*, pages 131–150. Springer, 2020.

[55] H.T. Lin, N.W. Chi, S.H. Hsieh, A concept-based information retrieval approach for engineering domain-specific technical documents, Adv. Eng. Inf. 26 (2) (2012) 349–360, 2012.

[56] S. Macit İlal, H.M. Günaydın, Computer representation of building codes for automated compliance checking, Autom. Constr. 82(May 2016):43–58, (2017).

[57] S. Malsane, J. Matthews, S. Lockley, P.E. Love, D. Greenwood, Development of an object model for automated compliance checking, Autom. Constr. 49(PA): pages 51–58, (2015).

[58] L. McGibbney, B. Kumar. A knowledge-directed information retrieval and management framework for energy performance building regulations, American Society of Civil Engineers (ASCE, 2011.

[59] J. McCarthy, Programs with common sense. In Proceedings of the Teddington Conference on the Mechanization of Thought Processes, pages 75–91, London,1959.

[60] A. Meyers, Y. He, Z. Glass, J. Ortega, S. Liao, A. Grieve-Smith, R. Grishman, O. Babko- Malaya, The termolator: Terminology recognition based on chunking, statistical and search-based scores, Frontiers in Research Metrics and Analytics 3 (2018). , pages 1–14, 2018.

[61] A. Miles, S. Bechhofer. SKOS Simple Knowledge Organization System Reference, 2009.

[62] Ministry of Housing, Communities & Local Government (MHCLG), Final Report of the Expert Group on Structure of Guidance to the Building Regulations, Available at: https://www.gov.uk/government/publications/final-report-of-the-expert-group-on-structure-of-guidance-to-the-building-regulations, 2020.

[63] J. Mitchell, M. Lapata, Composition in Distributional Models of Semantics, Cognit. Sci. 34 (8) (2010) 1388–1429, 2010.

[64] S. Moon, G. Lee, S. Chi, H. Oh, Automated Construction Specification Review with Named Entity Recognition Using Natural Language Processing, Journal of Construction Engineering and Management, 147 (1) (2021) 04020147.

[65] R.J. Mooney, in: Learning for Semantic Parsing, Springer, 2007, pp. 311–324.

[66] M. Muffatto, Introducing a platform strategy in product development, Int. J. Prod. Econ. 60 (1999) pages 145–153, 1999.

[67] N.O. Nawari, Automating Code Compliance Checking, MDPI - Buildings 9 (4) (2019) 86, 2019.

[68] R.A. Niemeijer, B. De Vries, J. Beetz, Freedom through constraints: User-oriented architectural design, Journal of Advanced Engineerig Informatics 28 (1) (2014) pages 28–36, 2014.

[69] NBS, NBS research finds users value Approved Documents documents 2017 Retrieved from https://www.thenbs.com/knowledge/nbs-research-finds-users-value-approved-, 2017.

[70] F. Noardo, D. Guler, J. Fauth, G. Malacarne, S.M. Ventura, M. Azenha, P.-O. Olsson, L. Senger, Unveiling the actual progress of digital building permit: Getting awareness through a critical state of the art review, Build. Environ. 213 (2022) 108854.

[71] E. Nuyts, J. Werbrouck, R. Verstraeten, L. Deprez, Validation of Building Models against Legislation Using SHACL LDAC2023), 3633 (2023) 164–175, 2023.

[72] D. Oberle, How ontologies benefit enterprise applications, Semantic Web 5 (6) (2014) pages 473–491, 2014.

[73] M. Pertierra, S. Lawsky, E. Hemberg, U.M. O'Reilly, Towards formalizing statute law as default logic through automatic semantic parsing. In, CEUR Workshop Proceedings 2143, (2017).

[74] C. Preidel, A. Borrmann, in: BIM-based code compliance checking, Building Information Modeling Technology Foundations and Industry Practice, Springer International Publishing, 2018, pp. 367–381.

[75] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (1), pages 5485–5551, 2020.

[76] C. Ramisch, C. R. Cordeiro, A. Savary, V. Vincze, V. B. Mititelu, A. Bhatia, M. Buljan, M. Candito, P. Gantar, V. Giouli, T. Güngör, A. Hawwari, U. Iñurrieta, J. Kovalevskaite, S. Krek, T. Lichte, C. Liebeskind, J. Monti, C. P. Escartín, B. QasemiZadeh, R. Ramisch, N. Schneider, I. Stoyanova, A. Vaidya and Walsh, A. Edition 1.1 of the Parseme shared task on automatic identification of verbal multiword expressions. In *LAW-MWE-CxG 2018 - Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions, Proceedings of the Workshop*, pages 222–240, 2018.

[77] M.H. Rasmussen, P. Pauwels, C.A. Hviid, J. Karlshøj, in: Proposing a Central AEC Ontology That Allows for Domain Specific Extensions, Edinburgh. Heriot-Watt University, 2017, pp. 237–244.

[78] W. Rasdorf, Generic Design Standards Processing in an Expert System Environment, ASCE Journal of Computing in Civil Engineering 2 (1) (1988) 1988, https://doi.org/10.1061/(ASCE)0887-3801(1988)2:1(68).

[79] Recski, G., Lellmann, B., Kovacs, A., and Hanbury, A. Explainable rule extraction via semantic graphs. In *ASAIL/LegalAIIA@ ICAIL*, pages 24–35, 2021.

[80] J. Ruppenhofer, M. Ellsworth, M. Petruck. *FrameNet II: Extended theory and practice*, 2010. https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf.

[81] I.A. Sag, T. Baldwin, F. Bond, A. Copestake, D. Flickinger, Multiword expressions: A pain in the neck for NLP, In Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2276 (2002) pages 1–15.

[82] S. Sarawagi, Information Extraction, Foundations and Trends® in Databases 1 (3) (2007) 261–377.

[83] K.K. Schuler, VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon, University of Pennsylvania, 2005. PhD thesis.

[84] C. Shirky Ontology is Overrated: Categories, Links, and Tags. Clay Shirky's Writings About the Internet, 2005.

[85] J.M. Siskind, A computational study of cross-situational techniques for learning word- to-meaning mappings Cognition 1996 61(1–2 SPEC, ISS ():) (1996) 39–91.

[86] A. Smirnova, P. Cudré-Mauroux, Relation Extraction Using Distant Supervision, ACM Comput. Surv. 51 (5) (2018) 1–35.

[87] B. Smith, D.M. Mark, Geographical categories: an ontological investigation, Int. J. Geogr. Inf. Sci. 15 (7) (2001) 591–612.

[88] W. Solihin, C. Eastman, Classification of rules for automated BIM rule checking development, Autom. Constr. 53 (2015) 69–82.

[89] J. Soliman-Junior, C.T. Formoso, P. Tzortzopoulos, A semantic-based framework for automated rule checking in healthcare construction projects, Can. J. Civ. Eng. 47 (2) (2020) 202–214.

[90] J. Soliman-Junior, P. Tzortzopoulos, J.P. Baldauf, B. Pedo, M. Kagioglou, C. T. Formoso, J. Humphreys, Automated compliance checking in healthcare building design, Autom. Constr. 129 (2021) 103822.

[91] M. Uhm, G. Lee, Y. Park, S. Kim, J. Jung, J.-K. Lee, Requirements for computational rule checking of requests for proposals (rfps) for building designs in south korea, Adv. Eng. Inf. 29 (3) (2015) 602–615.

[92] J. Van der Heijden, J. De Jong, Towards a better understanding of building regulation, Environ. Plann. B. Plann. Des. 36 (6) (2009) 1038–1052.

[93] A. Villavicencio, M. Idiart, Discovering Multiword Expressions, Nat. Lang. Eng. 25 (6) (2019) 715–733, https://doi.org/10.1017/S1351324919000494. Cambridge University Press.

[94] C. Wu, P. Wu, J. Wang, R. Jiang, M. Chen, X. Wang, Developing a hybrid approach to extract constraints related information for constraint management, Autom. Constr. 124 (January) (2021) 103563.

[95] W. Wu, H. Li, H. Wang, K.Q. Zhu, Semantic Bootstrapping: A Theoretical Perspective, IEEE Trans. Knowl. Data Eng. 29 (2) (2017) 446–457.

[96] J. Xu, W.B. Croft, Query expansion using local and global document analysis, In *SIGIR Forum (ACM Special Interest Group on Information Retrieval)* (1996) 4–11.

[97] R.V. Yampolskiy, Turing test as a defining feature of ai-completeness, In the Footsteps of Alan Turing, Artificial Intelligence, Evolutionary Computing and Metaheuristics, 2013, pp. 3–17.

[98] J. Zhang, N.M. El-Gohary, Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking, Autom. Constr. 73 (2016) 45–57.

[99] J. Zhang, N.M. El-Gohary, Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking, J. Comput. Civ. Eng. 30 (2) (2016) 04015014.

[100] Zhang, R. and El-Gohary, N. A machine learning-based method for building code requirement hierarchy extraction. *Proceedings, Annual Conference - Canadian Society for Civil Engineering*, 2019-June:1–10.

[101] R. Zhang, N. El-Gohary, A Machine-Learning Approach for Semantically-Enriched Building-Code Sentence Generation for Automatic Semantic Analysis, In *Construction Research Congress* (2020) 1261–1270.

[102] R. Zhang, N.M. El-Gohary, in: A Clustering Approach for Analyzing the Computability of Building Code Requirements, Construction Research Congress 2018, American Society of Civil Engineers, Reston, VA, 2018, pp. 86–95.

[103] Zhang, Z., Ma, L., and Broyd, T. Rule capture of automated compliance checking of building requirements: a review. *Proceedings of the Institution of Civil Engineers-Smart Infrastructure and Construction*, pages 1–14, 2023.
[104] Z. Zheng, Y.C. Zhou, X.Z. Lu, J.R. Lin, Knowledge-informed semantic alignment and rule interpretation for automated compliance checking, Autom. Constr. 142 (August) (2022) 104524.
[105] B. Zhong, X. Xing, H. Luo, Q. Zhou, H. Li, T. Rose, W. Fang, Deep learning-based extraction of construction procedural constraints from construction regulations, Adv. Eng. Inf. 43 (December 2019) (2020) 101003.
[106] P. Zhou, N. El-Gohary, Ontology-based multilabel text classification of construction regulatory documents, J. Comput. Civ. Eng. 30 (4) (2016) 04015058.
[107] P. Zhou, N. El-Gohary, in: Text and Information Analytics for Fully Automated Energy Code Checking, Sustainable Civil Infrastructures, Springer Science and Business Media B.V, 2019, pp. 196–208.
[108] G.A. Miller, WordNet: a lexical database for English, Communications of the ACM 38 (11) (1995) 39–41.
[109] M. Palmer, D. Gildea, N. Xue, Semantic role labelling, Synthesis Lectures on Human Language Technologies, 3, Springer, 2010.
[110] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, daSilva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. Van Der Lei, E. Van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, Comment: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (2016) 1–9, 2016.