

Taking stock: a Linked Data inventory of Compliance Checking terms derived from Building Regulations

Ruben Kruiper^{1,2,*}, Ioannis Konstas¹, Alasdair J.G. Gray¹, Farhad Sadeghineko², Richard Watson² and Bimal Kumar²

¹*School of Mathematics and Computer Sciences Heriot-Watt University, Edinburgh, United Kingdom*

²*Department of Architecture and Built Environment Northumbria University, Newcastle, United Kingdom*

Abstract

Compliance Checking (CC) would be a lot easier if we could automatically map between (1) terms that occur in building regulations and (2) elements of buildings and building products. However, the terminology used in the regulations is vastly different from the terminology found in Building Information Models (BIM). We are therefore forced to somehow shoehorn the vocabulary of regulatory terms into a set of classes that may well be several orders of magnitude smaller. This paper aims to reduce the gap between terms found verbatim in the regulations, and the classes that exist in Linked Data Vocabularies in Architecture and Construction. We explore the automated extraction of domain terminology from building regulations, and interlink the resulting terms with existing controlled vocabularies like Uniclass. The resulting Knowledge Graph (KG) can be used to suggest relevant and related domain terminology, which improves collecting the inventory of Linked Data terms required for CC.

Keywords

Automated Compliance Checking, Knowledge Graph, Natural Language Processing, Linked Data

1. Introduction

Most building work requires approval, hence regulations are frequently accessed by professionals in the construction industry – from architects to building inspectors and contractors [1]. On top of this, regulatory compliance is of critical importance to existing building, as corroborated by incidents like the disastrous Grenfell fire [2]. Despite a plethora of motivations [3, 4, 5], and a long history of research [6, 7], a solution to Automated Compliance Checking (ACC) remains at large – also see our other paper submitted to this workshop [8]. The crux of most ACC motivations is to improve the usability of the building regulations – in terms of effectiveness, efficiency and ease-of-use. A related strand of ACC research specifically focuses on supporting human experts during a compliance audit or even during design [9, 10, 11]. We believe that providing support during Compliance Checking (CC) is a viable approach, in contrast to ACC.

In this paper we explore the development of a domain lexicon for CC, captured in a Linked Data Knowledge Graph (KG). Such a lexicon could greatly benefit research on intelligent

Proceedings LDAC2023 – 11th Linked Data in Architecture and Construction, June 15–16, 2023, Matera, Italy


*Corresponding author.

✉ r.kruiper@northumbria.ac.uk (R. Kruiper)

ORCID 0000-0002-2288-3743 (R. Kruiper); 0000-0002-6720-4425 (I. Konstas); 0000-0002-5711-4872 (A. J.G. Gray); 0000-0002-9359-6493 (F. Sadeghineko); 0000-0002-3868-8088 (R. Watson); 0000-0002-2539-4902 (B. Kumar)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Regulatory Compliance (i-ReC) tools, e.g., by easing the mapping between the regulatory texts and CC rules. However, manually constructing and maintaining even small parts of this lexicon is a tedious task, see Appendix A. Our aim is to speed up the identification of a comprehensive set of relevant concepts, related terms and surface forms – the form in which a concept occurs in text. Section 2 describes related work on deriving a lexicon for CC from text. Section 3 describes our aim, the data we use, and our approach to suggesting related terms to a human annotator. Section 4 provides some insight in our preliminary results. Despite our relatively small set of input regulations, we are able to quickly provide annotators with a relatively comprehensive overview of related terminology and surface forms. We believe our exploratory approach is of interest to the wider community, and share our data and code so that others may extend and improve the work as they see fit¹.

2. From domain vocabulary to conceptualisation

In this paper, we use ‘*building regulations*’ to refer to any regulations captured in building standards, codes of practices, guidance documents and so on. We use the terms ‘*lexicon*’ and ‘*conceptualisation*’ interchangeably to refer to the set of classes and properties required to compose compliance checking rules. Deriving a conceptualisation from text is part of any approach to ontology [12] or taxonomy [13] learning from text. Expected steps include term extraction, concept identification, and internal and external concept linking.

Term Extraction: This step revolves around identifying which words, or groups of words, denote a term – a single unit of information. Related Natural Language Processing tasks include Information Extraction (IE) [14], Semantic Role Labelling (SRL) [15], and domain term discovery [16]. In the ACC domain relevant studies include tagging of regulations with semantic markup, either manually, e.g., RASE [17], or automatic, e.g., Semantic Information Elements [18]. One challenge identified by these studies is handling terms that consist of multiple words, so-called Multi-Word Expressions (MWE) [19]. Proper handling of MWEs is a key issue in NLP [20, 21, 22], and is especially relevant for IE in technical domains [23].

Concept Identification: This step involves turning the most salient terms into the classes and properties of a conceptualisation. Related NLP tasks include entity resolution [24, 25] and canonicalisation [26]. Ideally, concepts are provided with a domain-specific definition and a set of surface forms [12]. Beyond existing ontologies and classification schemes, such as the Building Topology Ontology (BOT) [27] and UNICLASS [28], useful resources include domain dictionaries. While construction domain dictionaries exist, even as part of the British Standards, their terms and definitions are often proprietary – limiting their use in generating a shared conceptualisation for CC. To the best of our knowledge there exists no work on automated concept identification in the CC domain.

Internal Concept Linking: This step refers to the identification of relations between the concepts in a controlled vocabulary, e.g., a conceptualisation. NLP research exists on automatically learning hierarchical relations between terms, e.g., [29]. Synonym identification often relies on similarity measures, e.g., [30, 31, 32], while hyper/hyponymy (super/subclass)

¹We encourage readers to replicate our results and extend our approach, by sharing our code and data at: <https://github.com/rubenkruiper/irec>

detection may also rely on syntactic patterns, e.g., [33]. In the CC domain WordNet [34] has been used to link concepts [35] and work exists on a feature-based relation classifier [36].

External Concept Linking: This step refers to identifying relations between the concepts in a controlled vocabulary, and concepts found in other resources like BOT and UNICLASS. A relation of interest is the indication that two classes are identical, e.g., ‘*skos:exactMatch*’ and ‘*owl:equivalentClass*’. Such identity relations make it possible for independently constructed datasets to use each others’ information [37]. To the best of our knowledge there exists no work on automating external concept linking or alignment for CC.

3. Method

3.1. Aim

Appendix A describes our initial work towards manually developing a conceptualisation for the CC domain. Three domain experts each identify a group of terms related to a subtopic, e.g., ‘*thermal insulation*’. They identify links between terms found in existing vocabularies, and extend the terminology where they believe this is required. The three annotators note various issues, including:

- not being sure if terms added to the KG actually occur in the regulations
- not knowing when the collected terms comprehensively describe a small subdomain
- the tediousness of identifying new terms and relations, especially when definitions are missing and sources may not be reliable

Over a combined 6 days the annotators identify merely 302 vocabulary terms and link them to 214 external resources. Our **aim** is to speed up such manual efforts to identify relevant CC concepts, related terms and surface forms. It is important to distinguish (1) a vocabulary of words that occur verbatim in the regulations from a (2) a controlled vocabulary. The former may simply refer to the set of all the words that occur in the regulations, in our case this set also includes MWEs. The latter is a set of predefined and preferred terms that may be used in a domain taxonomy or other knowledge organisation system – the conceptualisation.

Table 1

Overview of the number of sentences in the foreground corpus (MAD) and the background corpus, as well as the number of SPAR.TXT outputs before and after cleaning. Note that about 61% of the retained spans are a MWE.

	MAD	EU regulations
sentences	20,598	11,106
SPAR.TXT outputs		
Unprocessed objects	123,359	72,375
– Unique	43,937	10,408
Cleaned objects	72,625	60,842
– Unique	5,584	2,948
Combined total unique spans		7,940
– MWEs		4,855 (61.15%)

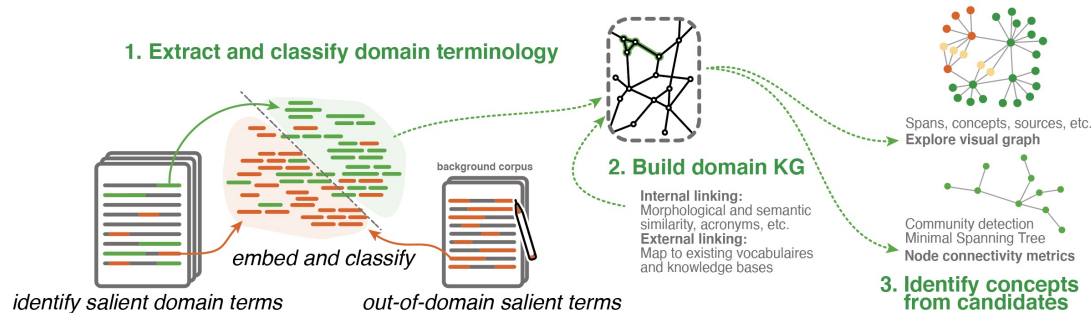


Figure 1: Overview of our strategy to support the manual identification of salient domain concepts, directly from the regulations, steps 1 to 3 correspond to Sections 3.3, 3.4 and 3.5.

3.2. Data and approach

In order to make our code and data available we rely on the UK Merged Approved Documents [38], which we will refer to as MAD. This is a set of open access British Building regulations and guidance that make up a total of 1,247 pages. To help identify which terms are specific to the building domain, we rely on a background corpus of 5 openly available EU regulations on the design of medical devices [39, 40, 41, 42, 43]. As suggested by [44] we selected the background corpus based on the similar text genre and similar time periods of publication. Table 1 provides some insight in the sizes of both input text corpora.

Figure 1 illustrates our approach. First, we extract two sets of salient noun-based spans of words – one from MAD and one from a background corpus of EU regulations on the design of medical devices. Section 3.3 explains how we crudely classify the word spans as construction domain or out-of-domain, taking into account the span’s source corpus and frequency-based characteristics. Section 3.4 explains how the domain terms are used to generate a KG, with nodes automatically linked to UNICLASS [28] and Wikidata [45] – a large general domain knowledge base. Section 3.5 describes various node-similarity measures and properties that we compute between spans in the KG. After adding these additional relations to the KG, we identify closely related terms through querying the graph, as well as by relying on network metrics like clustering coefficient, centrality, and degree. The resulting set of relations, and overall relatedness of terms, makes it possible to present annotators with an overview of terms that are likely to belong to the same topic.

3.3. Term extraction and domain classification

In previous work, we developed a Shallow Parser for Regulatory texts (SPAR.TXT) [46]. SPAR.TXT is originally trained to discover and identify terms – either single words or MWEs expressing a single unit of information – in the Scottish Building regulations [47]. We run SPAR.TXT over our corpora and limit the identification of OBJECT spans that express noun-based phrases, e.g., real-world objects or otherwise distinguishable concepts. Examples include ‘*the fire-fighting lift*’ and ‘*storage building*’, but also ‘*design process*’ and ‘*theory*’. We post-process the SPAR.TXT results relatively strictly, to improve the quality of terms presented to human annotators. Table 1

provides an overview of the numbers of candidate terms extracted from both sets of texts. Inspired by [16] we distinguish domain from out-of-domain concepts based on the frequency that a term was found in the foreground corpus and background corpus. We adopt a modified Term Frequency-Inverse Document Frequency (TF-IDF) metric:

$$TF-IDF(t) = \log\left(1 + \frac{f_{c_t}}{f_{c_t} + b_{c_t}}\right) * \log(avgIDF_t) \quad (1)$$

with f_{c_t} the number of times term t occurs in the foreground corpus, b_{c_t} the background corpus count, and $avgIDF$ the averaged IDF weight over the subword tokens of term t .

We compute term embeddings using a basic case-sensitive pre-trained BERT language model [48], which we multiply with the average IDF-weight of a term’s tokens and normalise as suggested by [32]. We compute the Nearest Neighbours (NN) graph to identify the 500 most similar embeddings for each term, relying on cosine similarity. Figure 2 plots the number of NNs that only appear in the foreground corpus against the modified TF-IDF value for terms. Note that our TF-IDF value defaults to 0 if a term only occurs in the background corpus. Terms are labelled as out-of-domain when both (1) their TF-IDF value is below 0.6, and (2) less than 200 out of the 500 NNs can be found only in the foreground corpus. This labelling strategy splits the total 7,940 terms into 4,958 domain and 2,982 general/out-of-domain terms. Note that these numbers depend highly on the size of the input data, cleaning, and the domain-classification values that we picked.

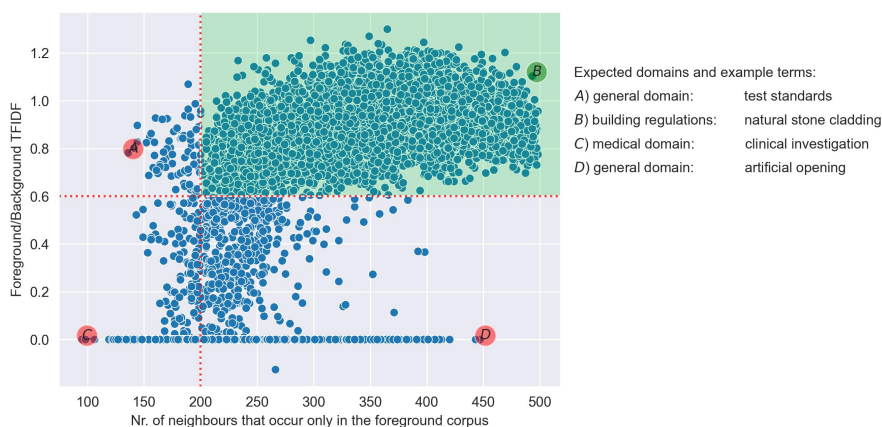


Figure 2: Plot of frequency-based statistics (TF-IDF) against semantic similarity of terms (number of NNs). Terms with a high TF-IDF value occur more often (or only) in the UK Merged Approved documents.

3.4. Building the KG

We build our KG as a directed edge-labelled RDF graph. We distinguish between (1) the 295 defined terms that were manually extracted from MAD and (2) spans of text. The former are added as instances of ‘*skos:Concept*’, and provided with their preferred label, alternative

labels and definitions as found in MAD. The latter are currently added as *‘irec:CharacterSpan’*² instances; these include terms that were automatically extracted from text, terms found in the index term lists of MAD and terms found in our manually assembled KG. If a span was identified by SPAR.TXT, this is indicated using *‘prov:wasAttributedTo’*.

Defined concepts and spans are kept in separate namespaces, the corresponding nodes are linked through *‘skos:exactMatch’*. We add a UNICLASS namespace to store UNICLASS terms that occur verbatim in MAD, again linking them to a span node that has the same *‘rdfs:label’* – we create this span if it did not exist in the KG yet. We also add a WIKIDATA namespace, to which we add nodes for those concepts and spans that can be found in WIKIDATA. The source of nodes is annotated using *‘prov:hasPrimarySource’*.

Where available, we add WIKIDATA definitions and classification labels to the span nodes in our graph. WIKIDATA contains much noise, and many out-of-domain interpretations – e.g., the span *‘wall’* is amongst others identified as a *‘Unix utility’*. We manually annotate the relevance of 1,220 distinct WIKIDATA classes that are linked to the defined and index terms from MAD, and find that 45.7% of these are irrelevant. We only keep definitions and corresponding class labels, if the label is within our set of classes that were annotated as relevant. We currently add definitions using *‘irec:wikiDefinition’* and add class labels as separate spans that are linked with *‘rdf:type’*. We also provide an indication of whether a term may be specific to the CC domain or not, based on the domain classification described in Section 3.3.

For the defined terms found in MAD, 25% of the 295 preferred and alternative labels occur in WIKIDATA. And while 29% of our spans can be found in WIKIDATA, this number drops to 13% after filtering out irrelevant WIKIDATA classes. This indicates that WIKIDATA (besides being very noisy), as expected, does not provide the coverage of building domain terms that would be needed for CC. Finally, we run SPAR.TXT on each of the WIKIDATA definitions. The aim is to identify whether defined spans are related, based on overlapping terms in their definitions – inspired by [49]. From the definitions we identify an additional 2,741 spans, which are added to the graph and related to the respective defined spans with *‘irec:definitionRelation’*.

²We define the CharacterSpan class and various properties in our own *‘irec’* namespace, some of these may be substituted with more commonly used resources to improve interoperability.

Table 2

Overview of the final number of concepts in the KG. Counts are based on a node’s UID (Unique Identifier) or preferred label. Most of the concepts that lack a definition are derived from UNICLASS.

Grouped by	node UID	preferred label	Nr. of concepts with identical label		
Concepts	2,143	1,624	MAD	UNICLASS	WIKIDATA
– MAD	295	295	0	0	90
– UNICLASS	598	580	0	18	93
– WIKIDATA	1.250	825	90	93	425
			Nr. of defined concepts by source		
			MAD	Uniclass	WikiData
– Defined	1,540	1,071	295	27	1,218
– Undefined	603	585	0	571	32

3.5. Identifying domain concepts

The KG contains 2.1K concepts with 1.6K unique preferred labels. Table 2 provides some insight in the number of concepts, their labels and their definitions. Seven labels have both a concept node with definition and a concept node without a definition; an example is *floating floor*. Notably, we do not include any definitions from UNICLASS into our KG.

There exist several indicators for spans in the KG to be a good candidate for a CC conceptualisation. In the first place, one could rely on the presence of an exact match to a concept from MAD, UNICLASS or WIKIDATA. Besides a span being represented by one or more concepts, one could consider looking at closely related spans. To this end, we compute several span-span features, such as whether a span occurs in the definition of another span – inspired by [49]. This means we run SPAR.TXT over all definitions and identify a set of 2.7K additional spans. Other computed features include:

- Morphological similarity, e.g., based on word overlap and edit distance we determine that *structural element* is morphologically similar to *element of structure*.
- Semantic similarity, we embed spans and the 5 NNs for each span.
- Domain classification, we rely on our crude domain specificity – see Section 3.3 – classification to assign a label to spans that were not seen in our corpora, e.g., *waterproofing membrane* occurs in the definition for *green roof* and is classified as CC domain.
- We identify potential acronyms in MAD, e.g., *LPA* stands for *local planning authority*.
- Antonym-based features, based on the 3.3K antonyms found in WordNet [34].

Nodes in our KG that are highly connected, are likely to include inflections, alternative labels and sub/super classes. We explore using the Louvain method for community detection [50] to

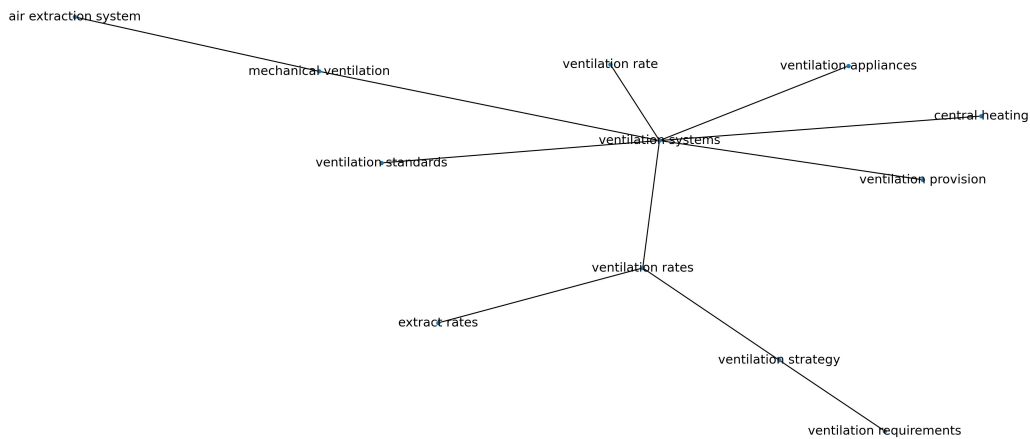


Figure 3: Overview of terms related to *mechanical ventilation*. An ego network was computed with a given span label (radius 5), which was recursively divided into smaller communities (max size 20). The visualisation is created by computing the Minimum Spanning Tree for the community that includes the user defined span label.

identify highly connected groups of nodes. To this end we convert the KG to a weighted graph with span-labels as nodes, where edge weights are manually set based on the relation types. This makes it possible to compute and visualise spans that are highly related in the KG, see Figure 3 for an example.

4. Initial annotator feedback

Together with the annotators we used a visual interface of the graph, as well as the graph metrics computed before, to:

- Explore a new subtopic to define for an CC conceptualisation.
- Compare the terms in an existing subtopic against those captured in a graph, to see if there are additional terms that we might want to add.

First, the new subtopic we aim to map is terminology revolving around ‘ventilation’. Figure 4 visualises a part of the terms in the KG revolving around ventilation. From the KG and graphs like the one shown in Figure 3, annotators quickly identify a hierarchy of high-level terms, along with some subclasses and definitions. As an example, the relatively high level term ‘flue’ has subclasses ‘vent’, ‘duct’, and ‘chimney’. However, our annotators were asked to rely on the same spreadsheet approach as used in the manual exploration of building a KG– see Appendix A. They find that working in a spreadsheet severely limits the types of relations they would like to add to the KG. As an example, while a ‘cable duct’ may be a type of ‘duct’, but it falls outside of the scope of ventilation-related terms. Similarly, ventilation terminology is interrelated with

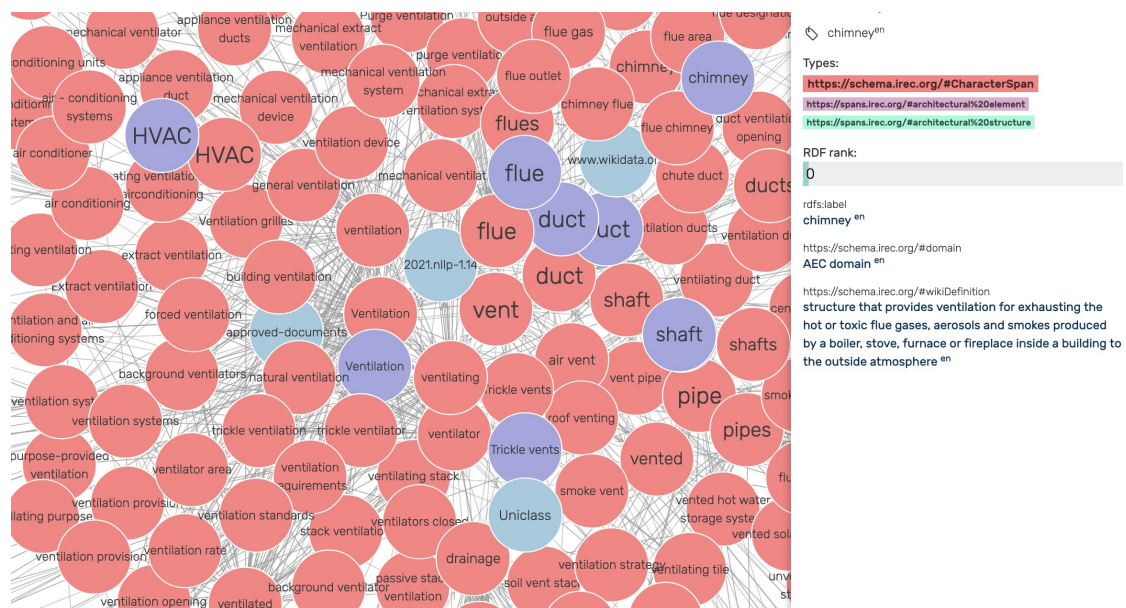


Figure 4: Visualisation of terms revolving around ‘ventilation’, created in GraphDB. Colours of nodes: (red) spans, such as ‘vent’ and a subclass ‘air vent’, (purple) concepts, such as the UNICLASS term ‘Trickle vents’ (Pr_30_59_94_90), (blue) primary source nodes, such as the SPAR.TXT paper ‘2021.nllp-1.14’.

drainage terminology. At this stage, it may be beneficial to work with vocabulary editors, such as VocBench 3 [51].

We compare the manually assembled terms revolving around ‘*thermal insulation*’ against those found in our KG. Our manual approach required two days of manual annotation which resulted in a hierarchy of 30 concepts, with a total of 40 alternative labels, 19 definitions and 13 links to concepts in external resources. Using our KG approach, the annotators were able to identify 15 additional terms that hadn’t been considered before in ten minutes. These included missing classes at various depths in the hierarchy, as well as alternative labels. For the new and existing concepts additional definitions from WIKIDATA were found, although these are often too generic – e.g., ‘*thermal insulation*’ (Q918306) is defined as ‘*insulation against heat transfer*’.

Overall, our annotators appreciate the graphical overviews and having terms, definitions and source information gathered together. They find that the related terms are grouped together quite well, which makes it easier to get a comprehensive overview of alternative and related terms for a concept. The access to various surface forms (inflections) is more useful from a computing perspective. However, while adding new terminology following our spreadsheet approach – see Appendix A – the annotators stress the need for a better editing environment. The consensus is that, ideally, the graphical overview of the KG allows adding new concepts and relations directly.

5. Conclusion and limitations

Solving ACC touches on a variety of open research problems, such as semantic parsing, canonicalisation and ontology matching. We present work on one of these sub-tasks, namely identifying the lexicon of terms that may be used to compose CC rules. Candidate terms are for a large part directly extracted from building regulations. By adding the terms to a KG, it is possible to capture a large variety of relations between the candidates, as well as link them to external resources. External resources that we relied on are an example of an existing vocabulary within the building domain, UNICLASS, and an example of a large general domain knowledge base, WIKIDATA. We show that (1) regulatory texts can provide a basis for a CC conceptualisation, and (2) using a KG to capture terminology can serve as a promising support tool for developing such a conceptualisation.

Our approach is exploratory and should be seen as a proof-of-concept. Most of the computing steps may be tweaked, or replaced. As an example, the way we currently compute some features, such as morphological similarity between spans $\mathcal{O}(n^2)$, limits the scalability of our approach. Further research would be required to determine, e.g., the value of computed features, the relevance of WIKIDATA classes, the domain classification. As noted, our KG schema contains several idiosyncratic RDF classes and properties that may be replaced with more widely used equivalents. Furthermore, due to licensing restrictions our KG only covers the MAD. This limits the scope and usability of the KG, as well as its use for suggesting terms that should be included in a CC lexicon.

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. Our thanks also go to Ian Babelon and Huyam Abudib for their help with annotation.

References

- [1] J. McKechnie, S. Shaaban, S. Lockiey, Computer Assisted Processing of Large Unstructured Document Sets: A Case Study in the Construction Industry, in: Proceedings of the ACM Symposium on Document Engineering, 2001, pp. 11–17.
- [2] C. Cook, How legal drafting may be central to fire safety debate - BBC News, 2017. URL: <https://www.bbc.co.uk/news/uk-41049510><http://www.bbc.co.uk/news/uk-41049510>.
- [3] F. Meijer, H. Visscher, L. Sheridan, Building regulations in Europe Part I: A comparison of the systems of building control in eight European countries, Delft University Press Science, Delft, 2014.
- [4] R. A. Niemeijer, B. De Vries, J. Beetz, Freedom through constraints: User-oriented architectural design, *Advanced Engineering Informatics* 28 (2014) 28–36. doi:10.1016/j.aei.2013.11.003.
- [5] C. Preidel, A. Borrmann, BIM-based code compliance checking, in: *Building Information Modeling: Technology Foundations and Industry Practice*, Springer International Publishing, 2018, pp. 367–381. URL: https://doi.org/10.1007/978-3-319-92862-3_22. doi:10.1007/978-3-319-92862-3{_}22.
- [6] J. Dimyadi, 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)*, 380, 2013, pp. 172–185.
- [7] N. O. Nawari, Automating Code Compliance Checking, *MDPI - Buildings* 9 (2019) 86. URL: <https://www.mdpi.com/2075-5309/9/4/86>.
- [8] R. Kruiper, I. Konstas, A. J. Gray, F. Sadeghineko, R. Watson, B. Kumar, Don't Shoehorn, but Link Compliance Checking Data (2023).
- [9] J. Dimyadi, C. Clifton, M. Spearpoint, R. Amor, Computerizing Regulatory Knowledge for Building Engineering Design, *Journal of Computing in Civil Engineering* 30 (2016). doi:10.1061/(asce)cp.1943-5487.0000572.
- [10] J. Dimyadi, R. Amor, Automating Conventional Compliance Audit Processes, in: *14th IFIP International Conference on Product Lifecycle Management (PLM)*, 2017, pp. 324–334.
- [11] J. Dimyadi, R. Amor, BIM-based compliance audit requirements for building consent processing., in: J. Karlshøj, R. Scherer (Eds.), *eWork and eBusiness in Architecture, Engineering and Construction*, September, CRC Press, 2018, pp. 465–471. URL: <https://www.taylorfrancis.com/books/9780429013652>. doi:10.1201/9780429506215.
- [12] P. Buitelaar, P. Cimiano, B. Magnini, *Ontology Learning from Text : An Overview*, Learning (2004) 1–10.

- [13] F. Ameri, B. Kulvatunyou, N. Ivezic, K. Kaikhah, Ontological conceptualization based on the SKOS, *Journal of Computing and Information Science in Engineering* 14 (2014). doi:10.1115/1.4027582.
- [14] S. Sarawagi, Information Extraction, *Foundations and Trends® in Databases* 1 (2007) 261–377. URL: <http://pages.cs.wisc.edu/~anhai/courses/784-fall13/ieSurvey.pdf>. doi:10.1561/15000000003.
- [15] M. Palmer, D. Gildea, N. Xue, Semantic role labeling, volume 3, 2010. doi:10.2200/S00239ED1V01Y200912HLT006.
- [16] 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* 3:19 (2018) 1–14. doi:10.3389/FRMA.2018.00019/FULL.
- [17] E. Hjelseth, N. Nisbet, Capturing Normative Constraints By Use of the Semantic Mark-Up Rase, in: *Proceedings of CIB*, March, 2011, pp. 26–28. URL: <http://itc.scix.net/data/works/att/w78-2011-Paper-45.pdf>.
- [18] J. Zhang, N. M. El-Gohary, Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking, *Journal of Computing in Civil Engineering* 30 (2016) 04015014. doi:10.1061/(asce)cp.1943-5487.0000346.
- [19] 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*, pp. 1261–1270.
- [20] J. M. Siskind, A computational study of cross-situational techniques for learning word-to-meaning mappings, *Cognition* 61 (1996) 39–91. doi:10.1016/s0010-0277(96)00728-7.
- [21] 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)*, volume 2276, 2002, pp. 1–15. doi:10.1007/3-540-45715-1_1.
- [22] C. Ramisch, S. R. Cordeiro, A. Savary, V. Vincze, V. B. Mititelu, A. Bhatia, M. Buljan, M. Candido, 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, A. Walsh, 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, 2018*, pp. 222–240. URL: <https://gitlab.com/parseme/sharedtask-guidelines/issues>.
- [23] T. Baldwin, S. N. Kim, Multiword Expressions, in: N. Indurkha, F. J. Damerau (Eds.), *Handbook of Natural Language Processing*, second ed., Chapman and Hall, 2010, pp. 267–292.
- [24] L. Getoor, A. Machanavajjhala, Entity resolution, *Proceedings of the VLDB Endowment* 5 (2012) 2018–2019. URL: <https://dl.acm.org/doi/10.14778/2367502.2367564>. doi:10.14778/2367502.2367564.
- [25] V. Christophides, V. Eftymiou, T. Palpanas, G. Papadakis, K. Stefanidis, An Overview of End-to-End Entity Resolution for Big Data, *ACM Computing Surveys* 53 (2021). doi:10.1145/3418896.
- [26] L. Galárraga, G. Heitz, K. Murphy, F. M. Suchanek, Canonicalizing Open Knowledge Bases,

- in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14, 2014, pp. 1679–1688. doi:10.1145/2661829.2662073.
- [27] M. H. Rasmussen, P. Pauwels, C. A. Hviid, J. Karlshøj, Proposing a Central AEC Ontology That Allows for Domain Specific Extensions, in: Lean and Computing in Construction Congress - Volume 1: Proceedings of the Joint Conference on Computing in Construction, July, Heriot-Watt University, Edinburgh, 2017, pp. 237–244. URL: http://itc.scix.net/cgi-bin/works/Show?_id=lc3-2017-153. doi:10.24928/JC3-2017/0153.
- [28] J. Gelder, The principles of a classification system for BIM: Uniclass 2015, Proceedings of the 49th International Conference of the Architectural Science Association 1 (2015) 287–297. URL: https://anzasca.net/wp-content/uploads/2015/12/028_Gelder_ASA2015.pdf.
- [29] R. Fu, J. Guo, B. Qin, W. Che, H. Wang, T. Liu, Learning Semantic Hierarchies via Word Embeddings, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Stroudsburg, PA, USA, 2014, pp. 1199–1209. URL: <http://ir.hit.edu.cn/~car/papers/acl14embedding.pdf><http://aclweb.org/anthology/P14-1113>. doi:10.3115/v1/P14-1113.
- [30] T. K. Landauer, S. T. Dumais, R. Anderson, D. Carroll, P. Foltz, G. Pumas, W. Kintsch, L. Menn, L. Streeter, A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge, *Psychological Review* 1 (1997) 211–240.
- [31] P. D. Turney, Distributional Semantics Beyond Words: Supervised Learning of Analogy and Paraphrase (2013). URL: <https://arxiv.org/pdf/1310.5042.pdf>.
- [32] W. Timkey, M. van Schijndel, All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2021, pp. 4527–4546. URL: <https://aclanthology.org/2021.emnlp-main.372>. doi:10.18653/v1/2021.emnlp-main.372.
- [33] V. Shwartz, Y. Goldberg, I. Dagan, Improving Hypernymy Detection with an Integrated Path-based and Distributional Method (2016). URL: <https://arxiv.org/pdf/1603.06076.pdf><http://arxiv.org/abs/1603.06076>.
- [34] G. a. Miller, WordNet: a lexical database for English, *Communications of the ACM* 38 (1995) 39–41. doi:10.1145/219717.219748.
- [35] J. Zhang, N. El-Gohary, a Semantic Similarity-Based Method for Semi-Automated Ifc Extension, Proceedings of 5th International /11th Construction Specialty Conference (2015). doi:<https://open.library.ubc.ca/cIRcle/collections/52660/items/1.0076395>.
- [36] J. Zhang, N. El-Gohary, An Automated Relationship Classification to Support Semi-Automated IFC Extension, in: Construction Research Congress 2016, American Society of Civil Engineers, Reston, VA, 2016, pp. 2039–2049. URL: <http://ascelibrary.org/doi/10.1061/9780784479827.203>. doi:10.1061/9780784479827.203.
- [37] J. Raad, B. Wouter, F. van Harmelen, N. Pernelle, S. Fatiha, Detecting erroneous identity links on the web using network metrics, in: The Semantic Web – ISWC 2018, volume 11136 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2018, pp. 217–232. URL: http://dx.doi.org/10.1007/978-3-030-00671-6_13<http://link.springer.com/10.1007/978-3-030-00671-6>. doi:10.1007/978-3-030-00671-6.

- [38] HM Government, The Building Regulations 2010: The merged approved documents, Technical Report June, 2022. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082748/Merged_Approved_Documents_Jun2022_.pdf.
- [39] C. o. t. E. U. European Parliament, Council Directive 90/385/EEC - EN, 1990. URL: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31990L0385:en:HTML>.
- [40] C. o. t. E. U. European Parliament, Council Directive 93/42/EEC- EN, 1993. URL: <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31993L0042:EN:HTML>.
- [41] C. o. t. E. U. European Parliament, EU Directive 98/79/EC - EN, 1998. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31998L0079&from=EN>.
- [42] C. o. t. E. U. European Parliament, Regulation (EU) 2017/745 - EN, 2017. URL: <https://eur-lex.europa.eu/eli/reg/2017/745/oj>.
- [43] C. o. t. E. U. European Parliament, Regulation (EU) 2017/746 - EN, 2017.
- [44] Gwang-Yoon Goh, Choosing a Reference Corpus for Keyword Calculation, *Linguistic Research* 28 (2011) 239–256. doi:10.17250/khisli.28.1.201104.013.
- [45] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, *Communications of the ACM* 57 (2014) 78–85. doi:10.1145/2629489.
- [46] R. Kruiper, I. Konstas, A. J. Gray, F. Sadeghineko, R. Watson, B. Kumar, SPaR.txt, a Cheap Shallow Parsing Approach for Regulatory Texts (2021) 129–143. doi:10.18653/v1/2021.nllp-1.14.
- [47] Scottish Government, Building Standards Technical handbook 2020: Domestic, 2020. URL: <https://www.gov.scot/publications/building-standards-technical-handbook-2020-domestic/>.
- [48] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: arXiv preprint arXiv:1810.04805, 2018. URL: <https://github.com/tensorflow/tensor2tensorhttp://arxiv.org/abs/1810.04805>.
- [49] B. Kim, H. Choi, H. Yu, Y. Ko, Query Reformulation for Descriptive Queries of Jargon Words Using a Knowledge Graph based on a Dictionary, in: *International Conference on Information and Knowledge Management, Proceedings, Association for Computing Machinery*, 2021, pp. 854–862. doi:10.1145/3459637.3482382.
- [50] V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment* 2008 (2008) 1–12. doi:10.1088/1742-5468/2008/10/P10008.
- [51] A. Stellato, M. Fiorelli, A. Turbati, T. Lorenzetti, W. Van Gemert, D. Dechandon, C. Laaboudi-Spoiden, A. Gerencsér, A. Waniart, E. Costetchi, J. Keizer, VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons, *Semantic Web* 11 (2020) 855–881. doi:10.3233/SW-200370.
- [52] R. V. Guha, D. Brickley, S. MacBeth, Schema.org: Evolution of Structured Data on the Web, *Queue* 13 (2015). doi:10.1145/2857274.2857276.
- [53] A. Miles, S. Bechhofer, SKOS Simple Knowledge Organization System Reference, 2009. URL: <https://www.w3.org/TR/skos-reference/http://www.w3.org/TR/skos-reference/>.
- [54] C. P. Cheng, G. T. Lau, K. H. Law, J. Pan, A. Jones, Regulation retrieval using industry specific taxonomies, *Artificial Intelligence and Law* 16 (2008) 277–303. doi:10.1007/s10506-008-9065-5.

A. Manually developing a KG for ACC

We start by exploring the manual extension and linking of controlled vocabularies in the building domain. While this approach is not scalable, the development can provide important lessons for approaching the creation of a domain vocabulary [52]. We model our vocabulary using the Simple Knowledge Organisation System (SKOS), a common data model for sharing and linking knowledge resources like thesauri, taxonomies and classification schemes [53].

An example of a controlled vocabulary in the building domain is UNICLASS [28]. However, only 598 (4%) of the 15K UNICLASS terms occur verbatim in the 1.274 pages of the UK Merged Approved documents (MAD). This means that there is a severe mismatch between the wording used in the regulations and the wording used in UNICLASS. On the one hand, UNICLASS has a far wider coverage. On the other hand, considering the wide coverage of UNICLASS one would expect that many of its terms can be found in a relatively generic set of building regulations like the Approved Documents. In conclusion, UNICLASS can be expected to require substantial extension or reformatting if it is to provide a shared conceptualisation that aligns with the regulation texts.

A.1. Approach

Our first step is to develop a workflow for the collection and annotation of terminology that works for our annotators. We then record new entities and relations through existing vocabulary editors, such as VocBench 3 [51]. However, after several weeks and a multitude of meetings, we settled on using a spreadsheet to capture three small subdomains of interest. See Table 3 for an example of the template we used. The aim is to ensure consistency in capturing concepts, alternative labels, definitions, exact matches and other SKOS-based triples.

A.2. Findings

Developing a workflow and adding terms to the KG manually takes a tremendous amount of time. With the workflow in place it is still challenging to identify synonyms and hyponyms of terms. Our three annotators indicated to have spent at least two full days each on composing their part of the graph. As a result of these six days of work, only 302 vocabulary terms were identified with a total of 130 alternative labels. In 214 cases, a term was linked to the unique identifier of a matching term in an external resource; UNICLASS (49%), NRM3 (4%), and a selection of British Standards vocabularies (49%). Note that terms and definitions from the British Standards vocabularies cannot be shared due to licensing restrictions. Furthermore, the six days of work does not include the development of a workflow for annotation or developing scripts to integrate their separate annotations.

Initially, annotators identified relevant terms through search engines and keyword search in classification systems and standards. However, they found it difficult to identify appropriate sources of domain terminology. The quality of the sources is not always easy to judge, and a lack of definitions complicates the identification of relations between terms. Another hurdle is determining whether the terms identified are actually used in the regulations. Classification systems, and UNICLASS in particular, are found to contain classes that are inconsistent with every

day language use in industry. To exemplify this, many leaf nodes in UNICLASS are amalgamations of properties, such as *'fibre cement profiled sheet self-supporting cladding systems'*. This finding corroborates the statement that classification schemes represent the needs of the issuing agencies, rather than the needs of users [54].

In the end annotators agreed on a standard set of sources to use, accepting the risk of missing specialist sources for particular subject areas – which may include the more obscure terms that would be useful to capture in the KG. Sources such as the British Standards vocabularies are noted to be particularly valuable. In many cases definitions makes it easier to identify synonymous or closely related terms. As such, definitions reduce the need for domain knowledge and make the annotation process less error prone. Nevertheless, establishing which relationship should exist between terms in the KG can be difficult.

It is hard to determine when a comprehensive overview of terminology has been achieved. Annotators found themselves disagreeing on which terms to include and how they relate. They thought *'skos:related'* is too vague, and tended to interpret *'skos:broader'* as a *'rdfs:subClassOf'* relation. The latter would imply that both terms are classes in the KG schema, just as *'skos:Concept'* is a class.

Table 3

Part of the annotation template used for the manual curation of a KG, with several hierarchical examples.

Code	Concept / preferred label	Alternative label(s)	Related concepts	Definition	Application	Source	Exact Matches
1.1.2	thermal insulation products					BS EN ISO 9229-2020;	
1.1.2.2	in-situ thermal insulation product			Thermal insulation product produced or taking its final form at the site of application and that achieves its properties after installation.		BS EN ISO 9229-2020;	
1.1.2.2.7	polyisocyanurate foam insulation	PIR foam insulation; PIR; Polyisocyanurate insulation;	1.1.2.2.8;	Polyisocyanurate foam thermal insulation product, which is foamed in-situ insulation.	Between rafters; may also be applied to (the underside of) slates and tiles to stabilise where nail fatigue is an issue;	UNICLASS; Pr_25_31_28_65;	