Semantic Web Lab
HERIOT WATT UNIVERSITY
Northumbria University NEWCASTLE

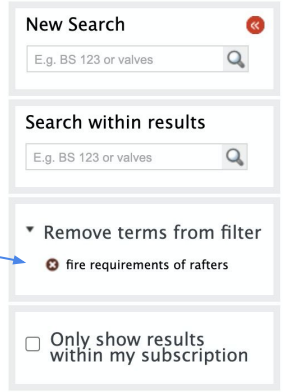
Ruben Kruiper
Ioannis Konstas
Alasdair Gray
Farhad Sadeghineko
Richard Watson
Bimal Kumar

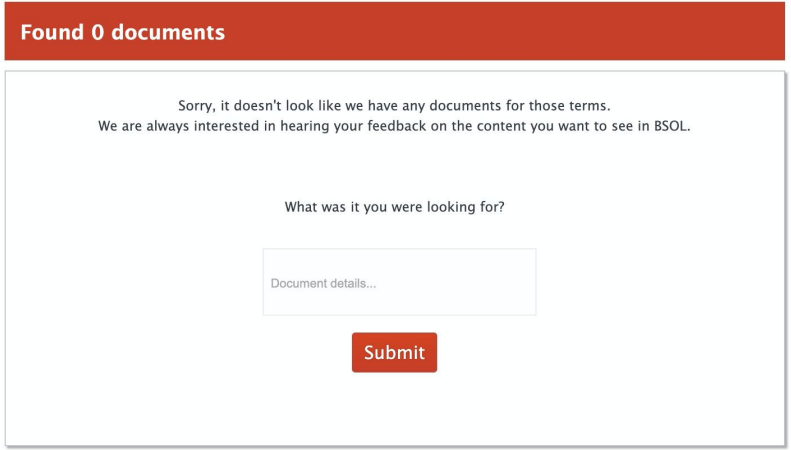
Document and Query Expansion for Information Retrieval on Building Regulations

Project: Intelligent Regulatory Compliance (I-ReC)
Problem: Searching for relevant regulations

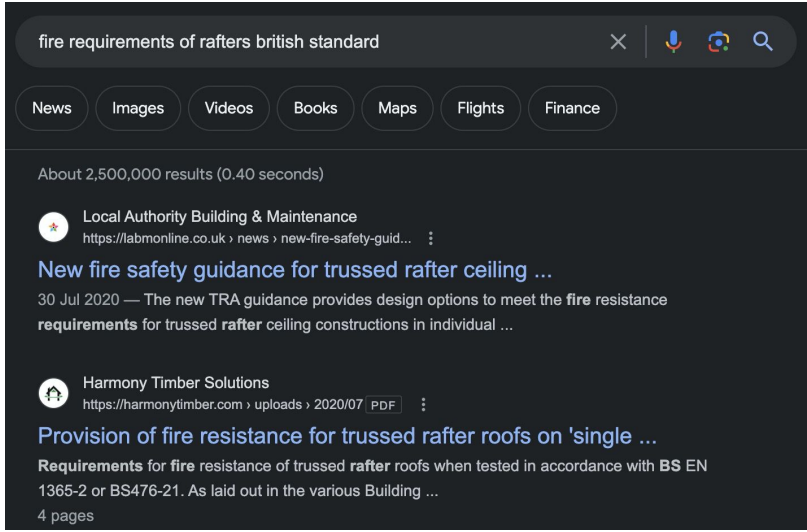
CONSTRUCTION INNOVATION HUB bre

Simple query?
"fire requirements of rafters"



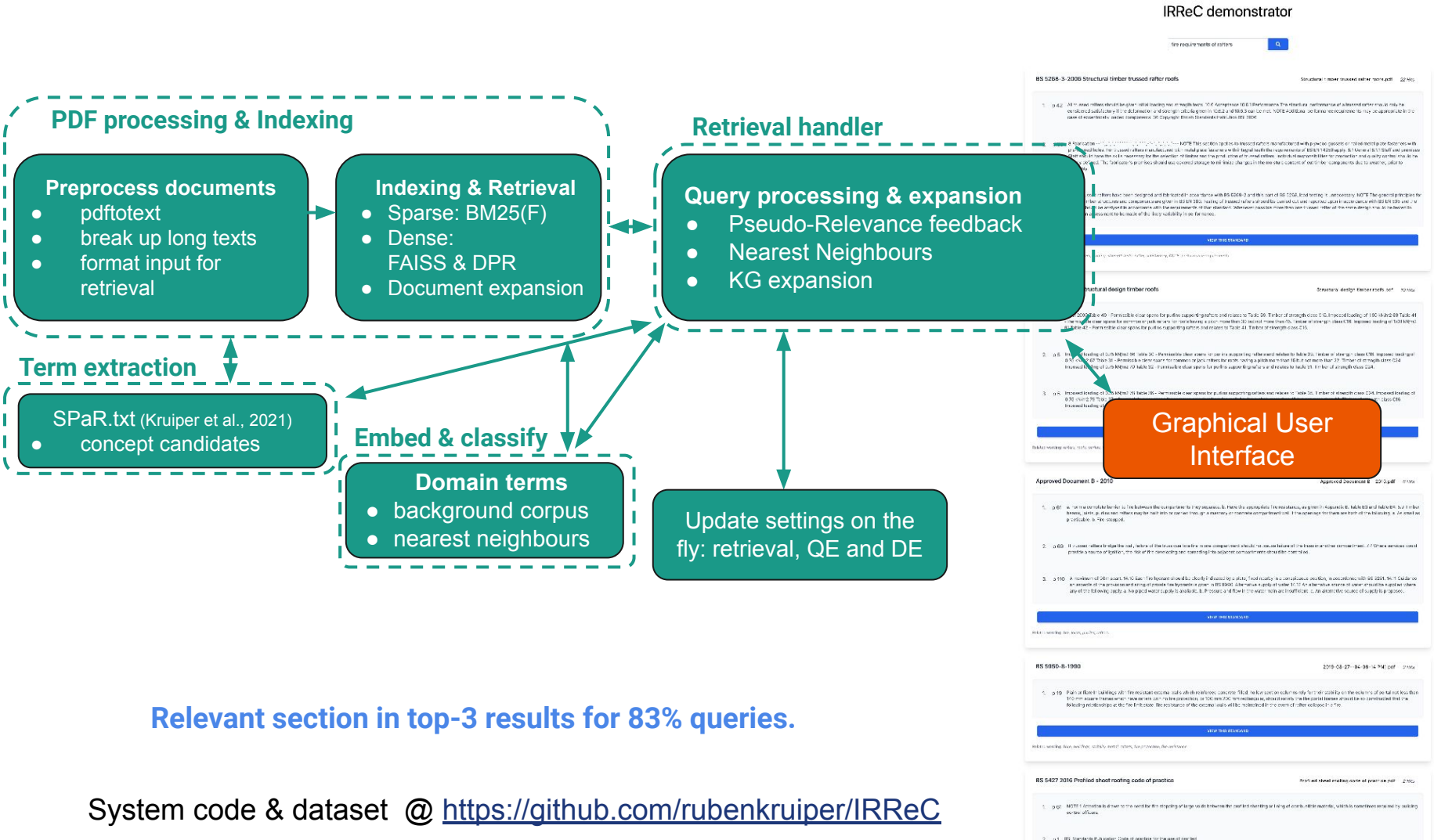


Ask the Trussed Rafter Association?



Check which standards have been consulted for existing products?

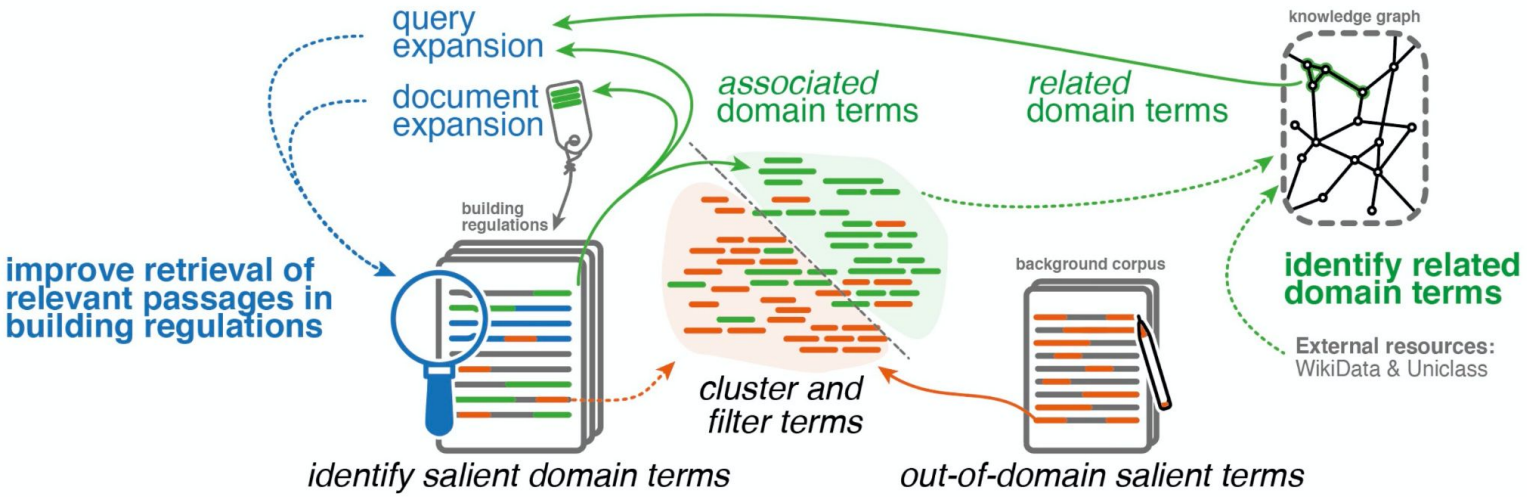
Also see (Cerovsek, 2009)



Relevant section in top-3 results for 83% queries.

System code & dataset @ https://github.com/rubenkruiper/IRReC

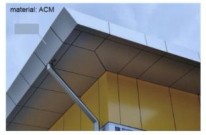
- Problem
- Passages
- Search
  - Dense IR
  - Expansion
- Findings



- Problem
- Passages
- Search
  - Dense IR
  - Expansion
- Findings

	Queries	Expanded queries
Avg. length (words)	5.86	20.19
Std. deviation	2.10	4.37
Shortest	3	12
Longest	13	34

Collecting queries to evaluate



bimacademy

Scope: active/Passive roof sub-assembly
Questions 1st interview round
1. Could you list 5 terms you would use to name the labelled element in this figure?
2. W.r.t. compliance, what requirements would you consider for this element?
3. Are there any specific structural requirements to consider, and how would you search for relevant regulations?
4. Are there any fire safety requirements to consider, …
etc.

8 x 30 minute interviews
12 questions
±1K queries
→
6 x 60 minute interviews
find relevant standards
reformulate queries
→
855 unique queries
random selection of 100
replaced 21 vague or very
similar queries

During annotation the **information need** was still **unclear** for many queries
→
2 domain experts
reformulate 20 queries
each and provide narrative
→
42 query & narrative pairs, punctuation or sentence context for NER

Example query & narrative pair
Initial query: roof covering drainage
Narrative: What are the requirements for pitched roof drainage in terms of adequately carrying rainwater to a suitable outlet, like gutters.
New query: pitched roof, drainage requirements

First: how do practitioners actually search for relevant regulations?

- Often combine multiple facets to their information need, e.g., *'fire requirements'* of *'external cladding'* on *'flats'*.
- Queries significantly longer than the average ±2 words of web-queries.
- Many (±50%) of the extracted domain terms are MWEs.

Problem

Passages

Search

- Dense IR
- Expansion

Findings

BSOL relies on ISO's International Classification for Standards (ICS).

Extend ICS?

- **Which topics/classes?**
- **Which classes to assign?**
  Regulations contain many domain terms (Lin et al., 2012)
  - complicates classification of documents
  - complicates how relevant an entire document is to a user-query

- Problem
- Passages
- Search
  - Dense IR
  - Expansion
- Findings

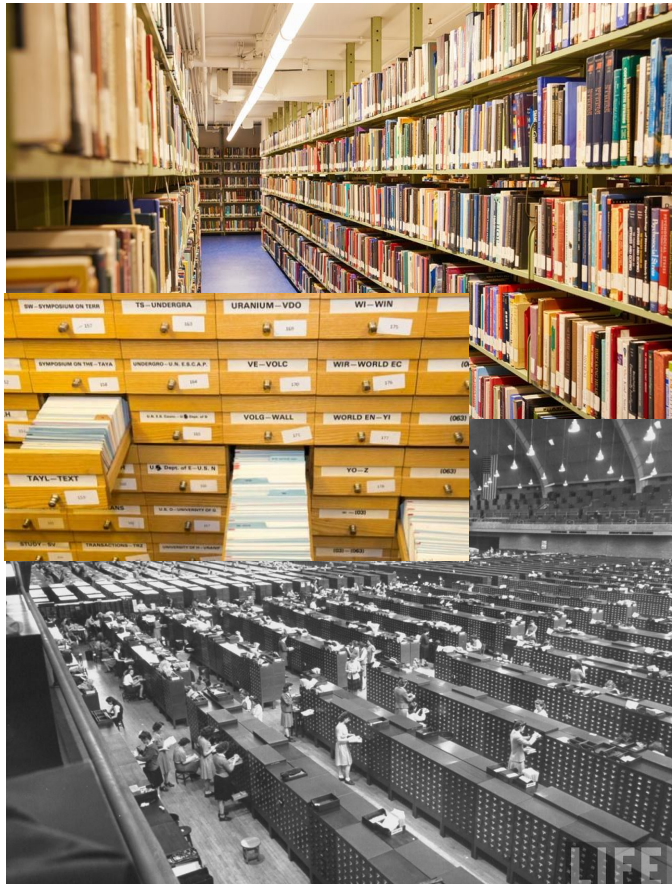
BSOL relies on ISO's International Classification for Standards (ICS).

~~**Extend ICS?**~~

- **Which topics/classes?**
- **Which classes to assign?**
  Regulations contain many domain terms (Lin et al., 2012)
  - complicates classification of documents
  - complicates how relevant an entire document is to a user-query

Classification is easier if you divide documents into shorter passages

- More fine-grained classification (Sannier & Baudry, 2012)
- Labelling domain terms in passages (Lin et al., 2012)



Compute similarity between query and document

e.g. BM25

Inverse Document Frequency
~ how often the word occurs in all documents

sum over query words

word count for document

modify influence a single query word can have

$$score(D,Q) = IDF(w_i) \sum_{i \in Q} \frac{f_{w_i,D} \cdot (k_1 + 1)}{f_{w_i,D} + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avglen(D)})}$$

$$Q = \{w_1, \ldots, w_n\}$$

modify importance of term in relation to relative length of document

Compute similarity between query and document

e.g. BM25

Inverse Document Frequency
~ how often the word occurs in all documents

sum over query words

word count for document

modify influence a single query word can have

$$score(D,Q) = \sum_{i \in Q} IDF(q_i) \cdot \frac{f_{w_i,D} \cdot (k_1 + 1)}{f_{w_i,D} + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avglen(D)})}$$

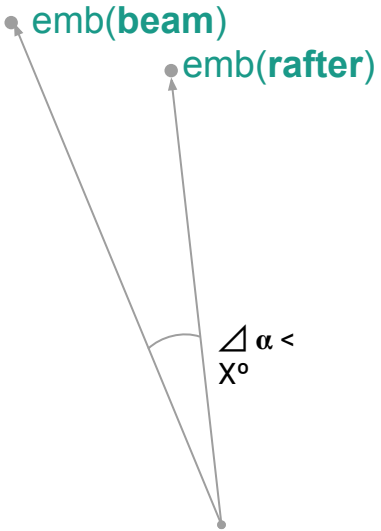
$$f_{w_i,D} = 0 \rightarrow score(Q,D) = 0$$

$$Q = \{w_1, \ldots, w_n\}$$

modify importance of term in relation to relative length of document

- Problem
- Passages
- Search
  - Dense IR
  - Expansion
- Findings

as a single structural element because of the moment-resisting connections used, especially at the column/rafter joints. Thus, in cases where the external wall of the building cannot be wholly unprotected, the rafter members of the frame, as well as the column members, may need to be fire protected. The design method for this is set out in SCI Publication P313.



Alternatively, check that upon the notional removal of each supporting column and each beam supporting one or more columns, or any nominal length of load-bearing wall (one at a time in each storey of the building), the building remains stable and that the area of floor at any

Similar embeddings means words occur in similar contexts.

Not necessarily same semantics.

For the purpose of structural analysis, trussed rafters may be represented by beam elements set along system lines and connected together at nodes (see Figure 1). For the greatest accuracy, system lines

Whether the bracing system is either a purpose designed bracing system at rafter or ceiling level, a prefabricated wind girder or ring beam, the bracing system chosen should be capable of resisting the imposed loads (wind and/or instability) and of limiting the horizontal deflection due to these loads to 10 mm

document expansion

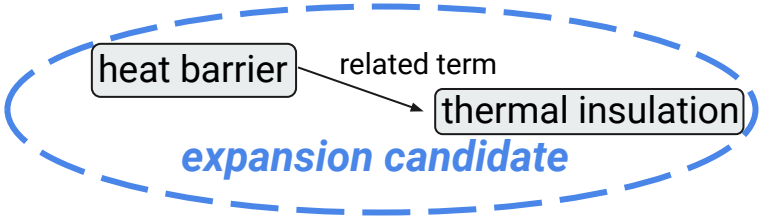
identify **salient** and add **associated** terms

query expansion

identify **salient** and add **associated** terms

$$score(D, Q) = score(D^+, Q^+)$$

user query



Query:
heat barrier → Expanded query:
heat barrier, thermal insulation

- Problem
- Passages
- Search
  - Dense IR
  - Expansion
- Findings

420 British Standards in PDF format → 288K 100-word passages

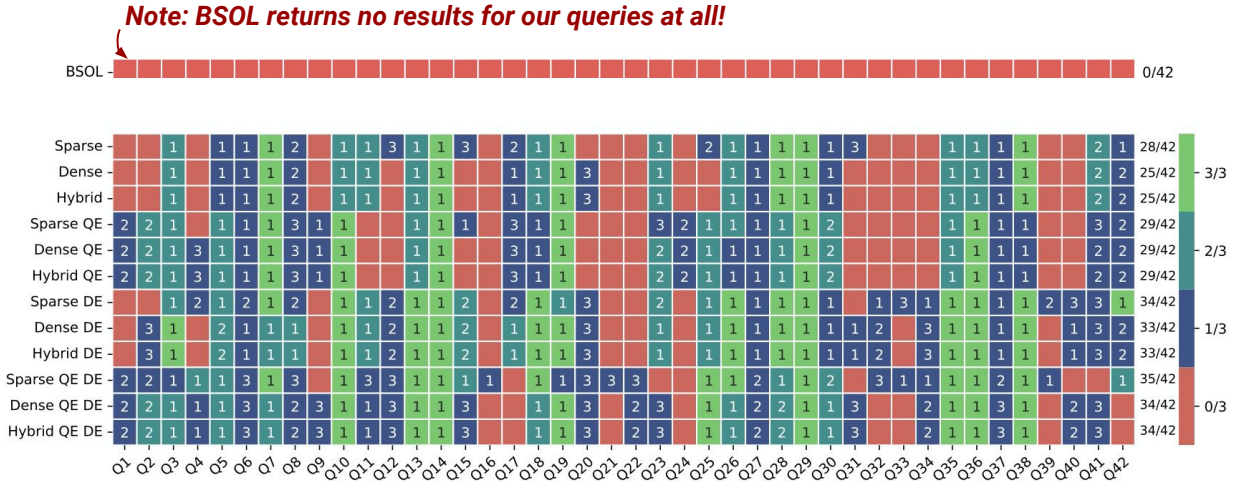
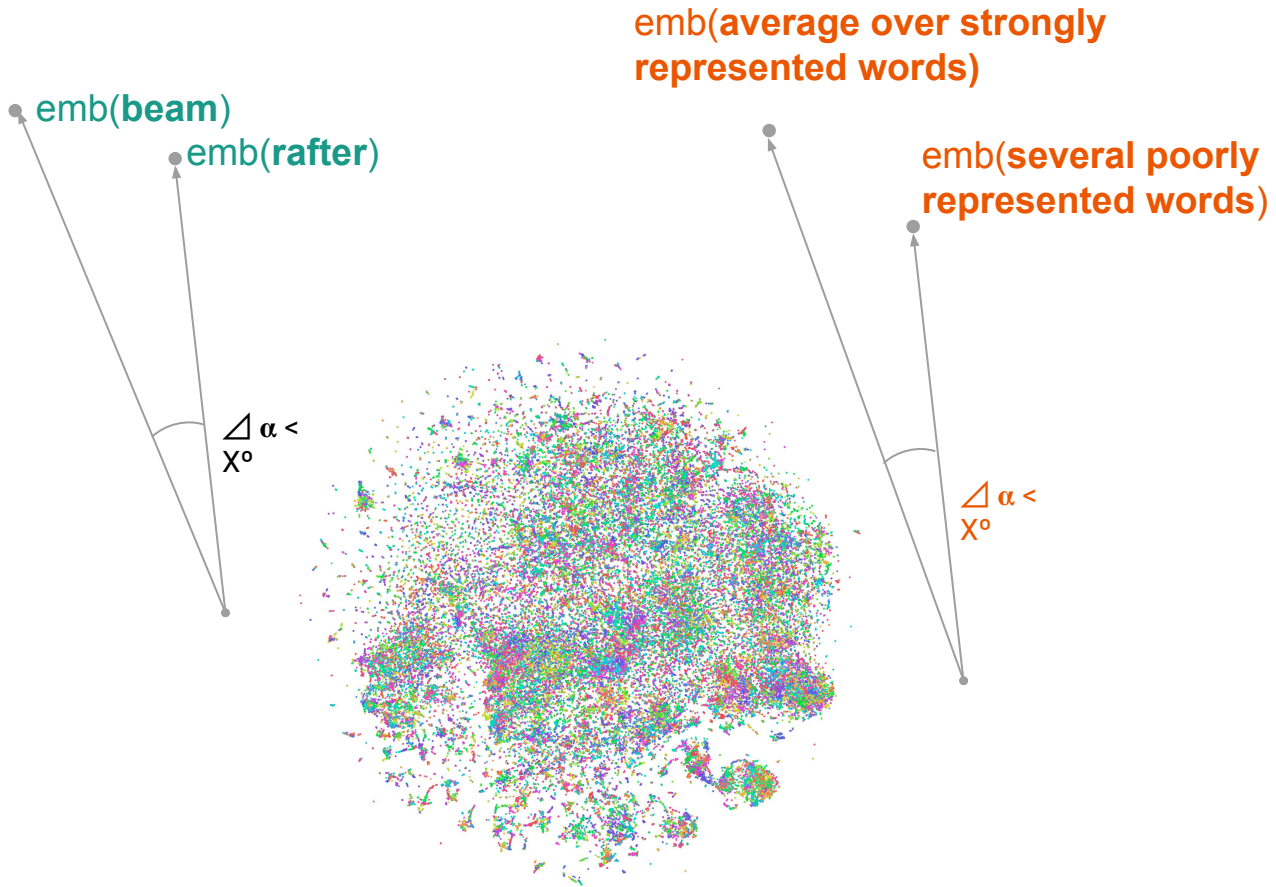


Figure 3: Overview of annotation results per query for each of our IR settings. Colours reflect the number of positive results, out of a possible total of 3 for each query. The numbers inside the cells indicate the best rank for a positive result. On the right-hand side an indication is given of how many queries were answered out of the total 42 queries.

- Problem
- Passages
- Search
  - Dense IR
  - Expansion
- Findings



- Problem
- Passages
- Search
  - Dense IR
  - Expansion
- Findings

document expansion

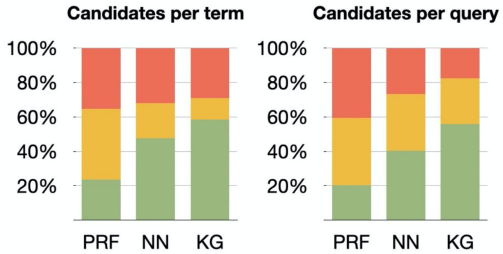
identify **salient** and add **associated** terms

query expansion

identify **salient** and add **associated** terms

$$score(D, Q) = score(D^+, Q^+)$$

user query



PRF: Pseudo Relevance Feedback
NN: Nearest Neighbours
KG: Knowledge Graph (Kruiper et al., 2023)

Query drift
Decreased accuracy w.r.t. representing information need

Unnecessary fluff
Probably doesn't help much, but not necessarily wrong
- *strict: wrong*
- *lenient: fine*

Good query candidates
Help focus on important aspects of the query, e.g., add synonyms and closely related terms that might be sought after

- Problem
- Passages
- Search
  - Dense IR
  - Expansion
- Findings

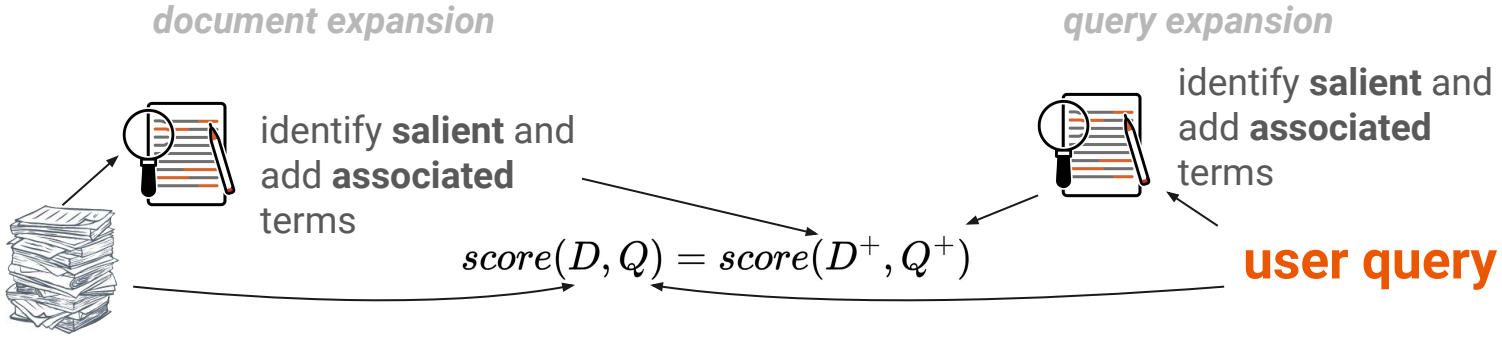


Table 3: Overview of results for each of our IR settings, all measures are based on the top 3 results for each query. Bold-faced values indicate best performance. MRR stands for Mean Reciprocal Rank, MAP for Mean Average Precision, the F1 score is a harmonic mean between precision and accuracy.

	sparse	sQ	sD	sQD	dense	dQ	dD	dQD	hybrid	hQ	hD	hQD
avg.F1	47.62	48.10	**59.76**	59.05	44.05	48.10	59.29	56.90	44.05	48.10	59.29	56.90
MRR	56.35	50.00	**66.27**	65.48	49.21	51.98	61.51	63.89	49.21	51.98	61.51	63.89
MAP	54.96	50.00	**65.28**	64.48	47.82	52.78	61.71	63.10	47.82	52.78	61.71	63.10
Total/129	50	50	**65**	63	47	50	**65**	60	47	50	**65**	60

avg. F1 does not care about order of results, simply a measure of precision and recall on the top *k* results
MRR looks only at the single highest-ranked relevant item
MAP takes into account all relevant results in the top *k*

Problem

Passages

Search

- Dense IR
- Expansion

Findings

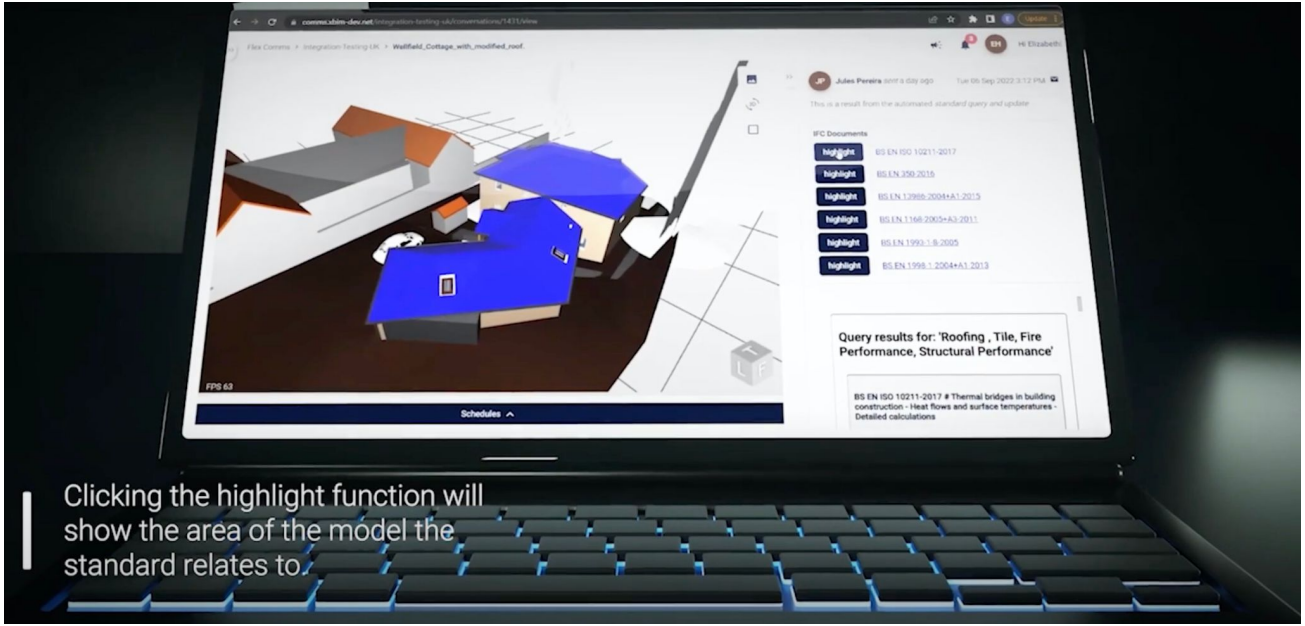
Narrative: *What standard provides information about the installation requirements for waterproofing at abutments?*

Query	Document	Retrieved passage
waterproofing of abutments	BS 5534-2014+A2 (2018) Slating and tiling for pitched roofs and vertical cladding — Code of practice	6.5.6 Abutments COMMENTARY ON 6.5.6 When fixing abutments for single-lap clay and concrete interlocking tiles: a) slate or tile fillets bedded in mortar are not recommended unless special details are provided to avoid the mortar cracking at the abutment face as a result of roof settlement, etc. ; b) cement fillets are not recommended. 6.5.6.1 Top edges On top edges, the top course tiles should allow the cover flashing to overlap by not less than 150 mm. 6.5.6.2 Side abutments Side abutments should be detailed with either sheet metal or proprietary flashings.

Narrative: *What are the requirements for installing metal stud structures for non-domestic buildings, including internal partition walls?*

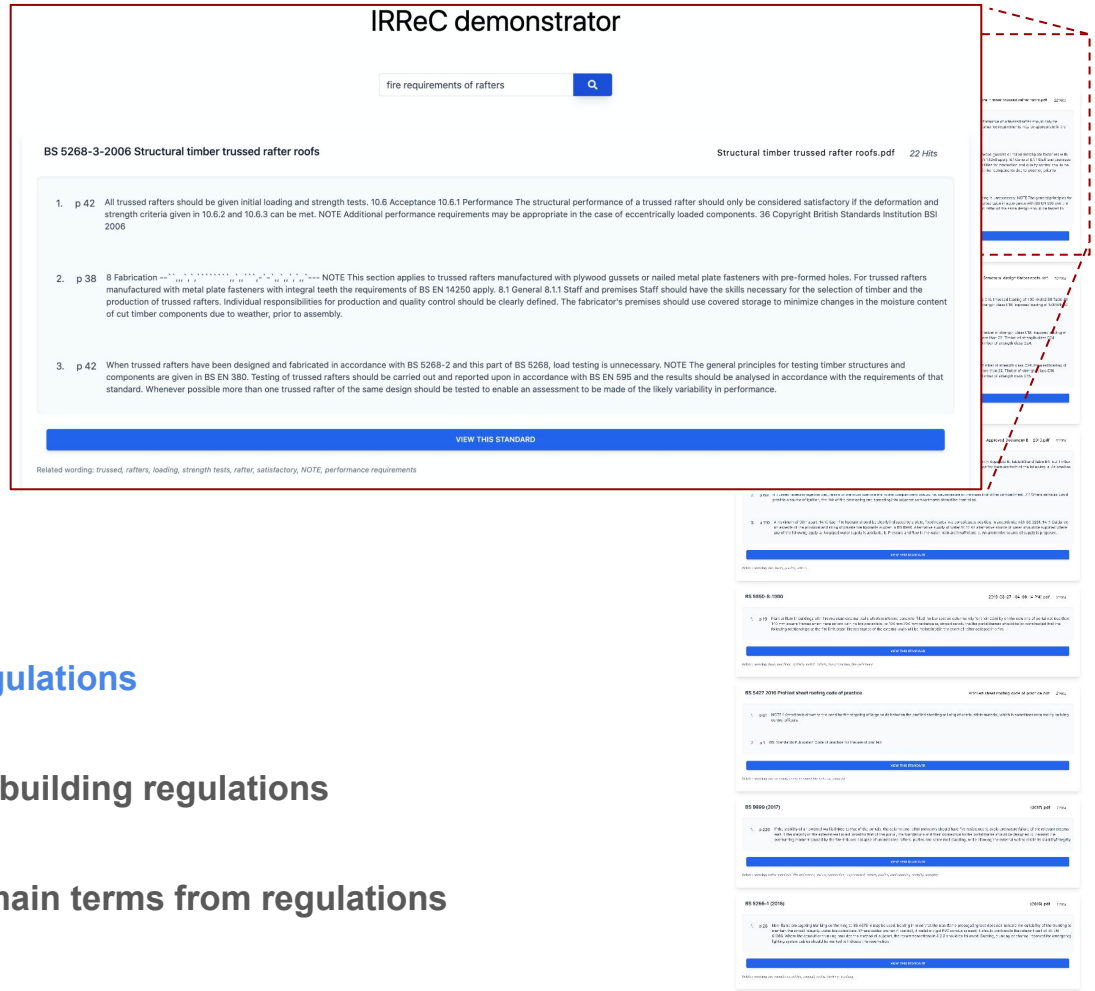
Query	Document	Retrieved passage
requirements for installing metal studs	BS EN ISO 13287 (2019) Personal protective equipment. Footwear. Test method for slip resistance	INTERNATIONAL STANDARD Personal protective equipment -- Footwear -- Test method for slip resistance 1 Scope This document specifies a method of test for the slip resistance of PPE footwear. It is not applicable to special purpose **footwear** containing spikes, **metal studs** or similar. Footwear claiming `slip resistance' would be deemed an item of personal protective equipment. NOTE For product development purposes, sole units, outsoles or other soling components such as top pieces may be tested.

- Problem
- Passages
- Search
  - Dense IR
  - Expansion
- Findings



Thank you!

Find me for a demo



- **Information Retrieval over building regulations**
  https://github.com/rubenkruiper/IRReC
- **Identifying Multi-Word Expressions in building regulations**
  https://github.com/rubenkruiper/SPaR.txt
- **Generating a Knowledge Graph of domain terms from regulations**
  https://github.com/rubenkruiper/irec