Semantic Web Lab

HERIOT WATT UNIVERSITY

Ruben Kruiper
Ioannis Konstas
Alasdair Gray

Northumbria University NEWCASTLE

Farhad Sadeghineko
Richard Watson
Bimal Kumar

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

CONSTRUCTION INNOVATION HUB

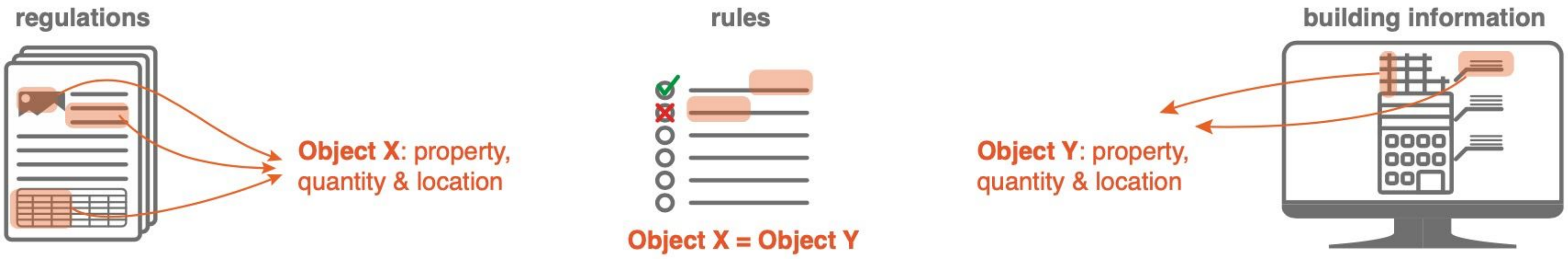
bre

Overview

- Aim
    - Collect a controlled vocabulary for Compliance Checking (CC)
- Approach
    - Extract terminology from building regulations
    - Classify domain of extracted terminology
    - Identify relevant terms from additional resources
    - Build a span-based graph
    - Explore using the span-based graph

Aim: Controlled vocabulary for Compliance Checking (CC)

Aim: shared conceptualisation for Compliance Checking

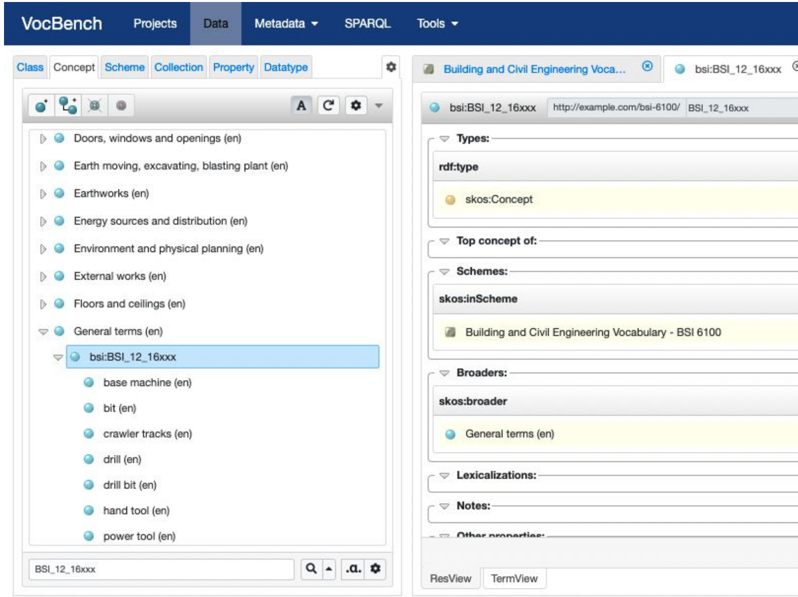

regulations
Object X: property, quantity & location
rules
Object X = Object Y
building information
Object Y: property, quantity & location

- Lexicon ~ the complete set of meaningful units in our ACC vocabulary
  - Which terms would be required for formulating Compliance Checking rules?
- Potential sources of terms:
  - **Building Regulations** (standards, codes of practice, eurocodes, guidance, etc.)
  - BIM models (vendor specific terms, interoperability ontologies, etc.)
  - Domain thesauri and vocabularies (**Uniclass**, NRM3, BsDD, etc.)
  - Building product datasheets (vendor specific names, product characteristics, etc.)
  - more?

“hot finished hollow section member”	owl:equivalentClass	*(Pr_20_76_52_16) Carbon steel hot-finished hollow sections*
	rdfs:type	*(TE_10_10_50) Structural members*
	skos:broader	*TATA Celsius® hot finished hollow sections*
	prov:location	*BS 5950-8 1990, BS EN 1090-3 2019,* and so on…
	etc.	

Aim: shared conceptualisation for Compliance Checking

- First, explore manual approach to build KG
  - What terms and relations can we expect?
  - Editing the data
    - *VocBench*, *Tematres*, *MS Excel*
  - Format
    - *Simple Knowledge Organisation System (SKOS)*
  - Reuse existing sources
    - *Uniclass, NRM3, BSI vocabularies, IFCowl, …*
- Workflow:
  1. Manual annotation in MS Excel
  2. OntoRefine to convert CSV to SKOS
  3. Further editing in VocBench



Doors, windows and openings (en)
Earth moving, excavating, blasting plant (en)
Earthworks (en)
Energy sources and distribution (en)
Environment and physical planning (en)
External works (en)
Floors and ceilings (en)
General terms (en)
bsi:BSI_12_16xxx
base machine (en)
bit (en)
crawler tracks (en)
drill (en)
drill bit (en)
hand tool (en)
power tool (en)

BSI_12_16xxx

bsi:BSI_12_16xxx http://example.com/bsi-6100/ BSI_12_16xxx

Types:
rdf:type
skos:Concept

Top concept of:

Schemes:
skos:inScheme
Building and Civil Engineering Vocabulary - BSI 6100

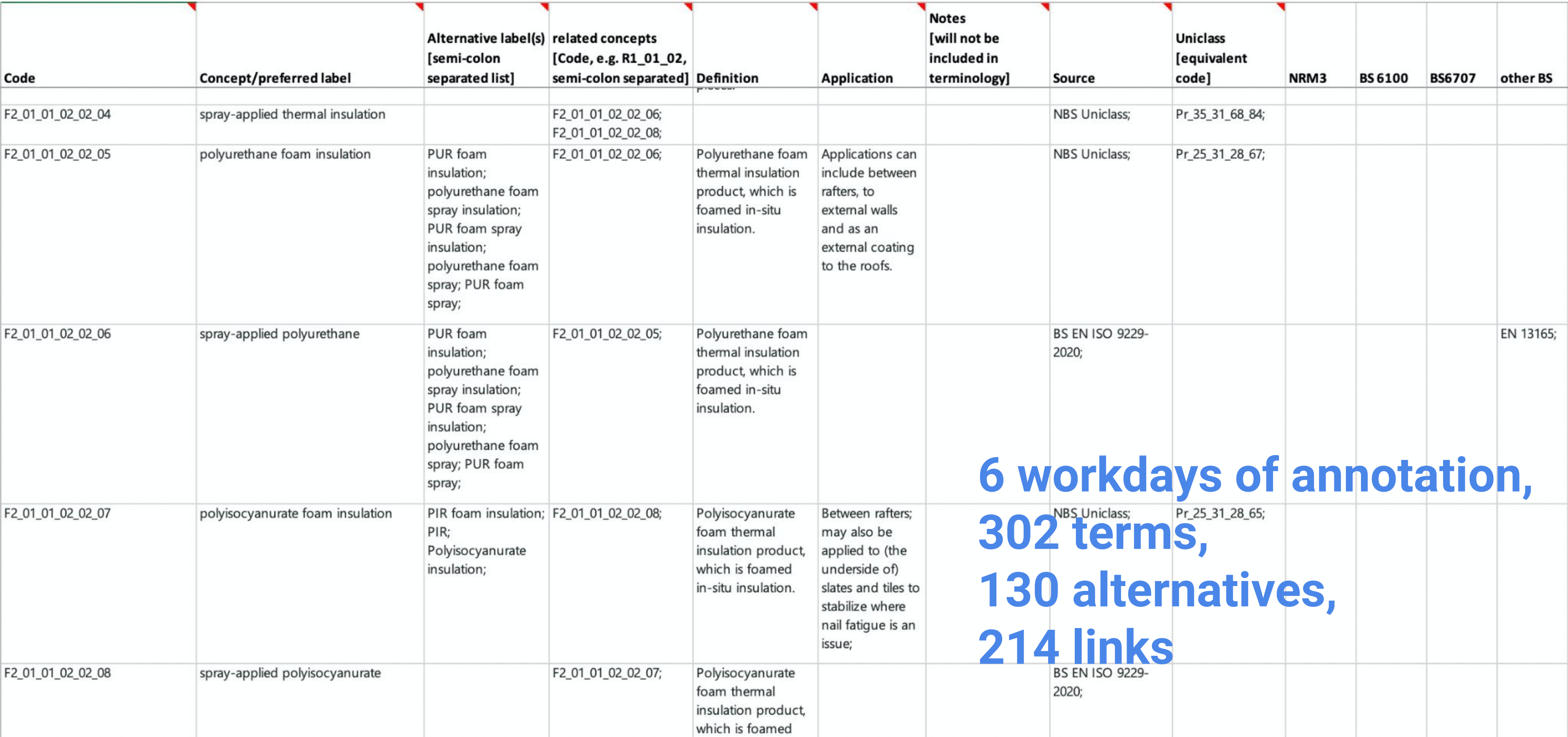
Broaders:
skos:broader
General terms (en)

Lexicalizations:

Notes:

Other properties:

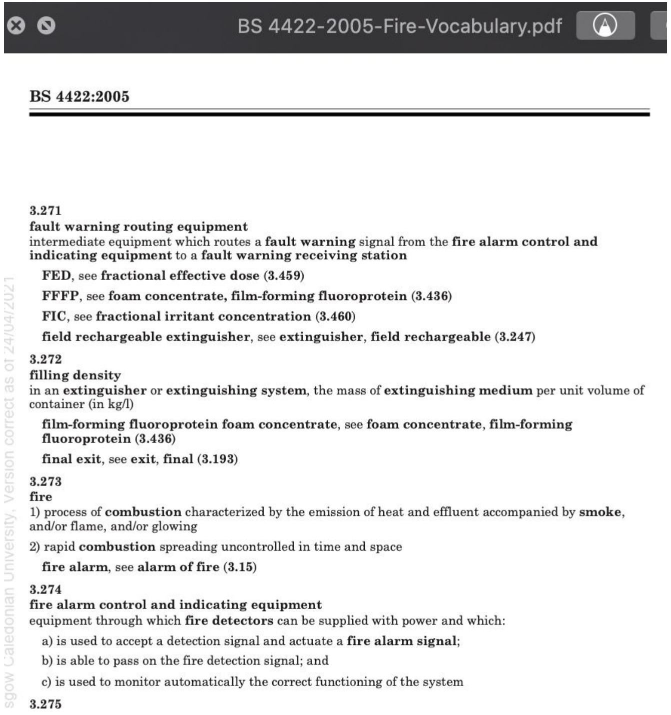
Aim: shared conceptualisation for Compliance Checking



Code	Concept/preferred label	Alternative label(s) [semi-colon separated list]	related concepts [Code, e.g. R1_01_02, semi-colon separated]	Definition	Application	Notes [will not be included in terminology]	Source	Uniclass [equivalent code]	NRM3	BS 6100	BS6707	other BS
				pieces.								
F2_01_01_02_02_04	spray-applied thermal insulation		F2_01_01_02_02_06; F2_01_01_02_02_08;				NBS Uniclass;	Pr_35_31_68_84;				
F2_01_01_02_02_05	polyurethane foam insulation	PUR foam insulation; polyurethane foam spray insulation; PUR foam spray insulation; polyurethane foam spray; PUR foam spray;	F2_01_01_02_02_06;	Polyurethane foam thermal insulation product, which is foamed in-situ insulation.	Applications can include between rafters, to external walls and as an external coating to the roofs.		NBS Uniclass;	Pr_25_31_28_67;				
F2_01_01_02_02_06	spray-applied polyurethane	PUR foam insulation; polyurethane foam spray insulation; PUR foam spray insulation; polyurethane foam spray; PUR foam spray;	F2_01_01_02_02_05;	Polyurethane foam thermal insulation product, which is foamed in-situ insulation.			BS EN ISO 9229-2020;					EN 13165;
F2_01_01_02_02_07	polyisocyanurate foam insulation	PIR foam insulation; PIR; Polyisocyanurate insulation;	F2_01_01_02_02_08;	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 stabilize where nail fatigue is an issue;		NBS Uniclass;	Pr_25_31_28_65;				
F2_01_01_02_02_08	spray-applied polyisocyanurate		F2_01_01_02_02_07;	Polyisocyanurate foam thermal insulation product, which is foamed			BS EN ISO 9229-2020;					

6 workdays of annotation,
302 terms,
130 alternatives,
214 links

Aim: shared conceptualisation for Compliance Checking



BS 4422:2005

3.271
fault warning routing equipment
intermediate equipment which routes a **fault warning** signal from the **fire alarm control and indicating equipment** to a **fault warning receiving station**

FED, see **fractional effective dose** (3.459)

FFFP, see **foam concentrate, film-forming fluoroprotein** (3.436)

FIC, see **fractional irritant concentration** (3.460)

field rechargeable extinguisher, see **extinguisher, field rechargeable** (3.247)

3.272
filling density
in an **extinguisher** or **extinguishing system**, the mass of **extinguishing medium** per unit volume of container (in kg/l)

film-forming fluoroprotein foam concentrate, see **foam concentrate, film-forming fluoroprotein** (3.436)

final exit, see **exit, final** (3.193)

3.273
fire
1) process of **combustion** characterized by the emission of heat and effluent accompanied by **smoke**, and/or flame, and/or glowing

2) rapid **combustion** spreading uncontrolled in time and space

fire alarm, see **alarm of fire** (3.15)

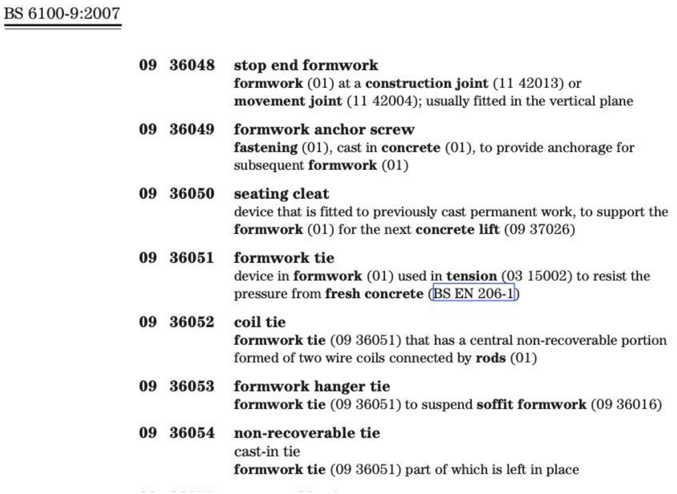
3.274
fire alarm control and indicating equipment
equipment through which **fire detectors** can be supplied with power and which:

a) is used to accept a detection signal and actuate a **fire alarm signal**;

b) is able to pass on the fire detection signal; and

c) is used to monitor automatically the correct functioning of the system

3.275



BS 6100-9:2007

09 36048 **stop end formwork**
formwork (01) at a **construction joint** (11 42013) or **movement joint** (11 42004); usually fitted in the vertical plane

09 36049 **formwork anchor screw**
fastening (01), cast in **concrete** (01), to provide anchorage for subsequent **formwork** (01)

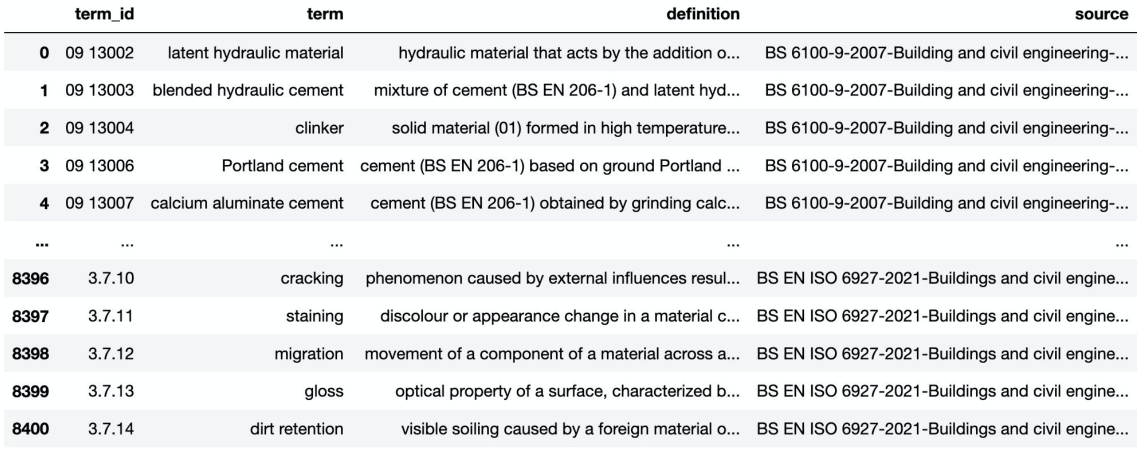
09 36050 **seating cleat**
device that is fitted to previously cast permanent work, to support the **formwork** (01) for the next **concrete lift** (09 37026)

09 36051 **formwork tie**
device in **formwork** (01) used in **tension** (03 15002) to resist the pressure from **fresh concrete** (BS EN 206-1)

09 36052 **coil tie**
formwork tie (09 36051) that has a central non-recoverable portion formed of two wire coils connected by **rods** (01)

09 36053 **formwork hanger tie**
formwork tie (09 36051) to suspend **soffit formwork** (09 36016)

09 36054 **non-recoverable tie**
cast-in tie
formwork tie (09 36051) part of which is left in place



	term_id	term	definition	source
0	09 13002	latent hydraulic material	hydraulic material that acts by the addition o...	BS 6100-9-2007-Building and civil engineering-...
1	09 13003	blended hydraulic cement	mixture of cement (BS EN 206-1) and latent hyd...	BS 6100-9-2007-Building and civil engineering-...
2	09 13004	clinker	solid material (01) formed in high temperature...	BS 6100-9-2007-Building and civil engineering-...
3	09 13006	Portland cement	cement (BS EN 206-1) based on ground Portland ...	BS 6100-9-2007-Building and civil engineering-...
4	09 13007	calcium aluminate cement	cement (BS EN 206-1) obtained by grinding calc...	BS 6100-9-2007-Building and civil engineering-...
...	...	...	...	...
8396	3.7.10	cracking	phenomenon caused by external influences resul...	BS EN ISO 6927-2021-Buildings and civil engine...
8397	3.7.11	staining	discolour or appearance change in a material c...	BS EN ISO 6927-2021-Buildings and civil engine...
8398	3.7.12	migration	movement of a component of a material across a...	BS EN ISO 6927-2021-Buildings and civil engine...
8399	3.7.13	gloss	optical property of a surface, characterized b...	BS EN ISO 6927-2021-Buildings and civil engine...
8400	3.7.14	dirt retention	visible soiling caused by a foreign material o...	BS EN ISO 6927-2021-Buildings and civil engine...

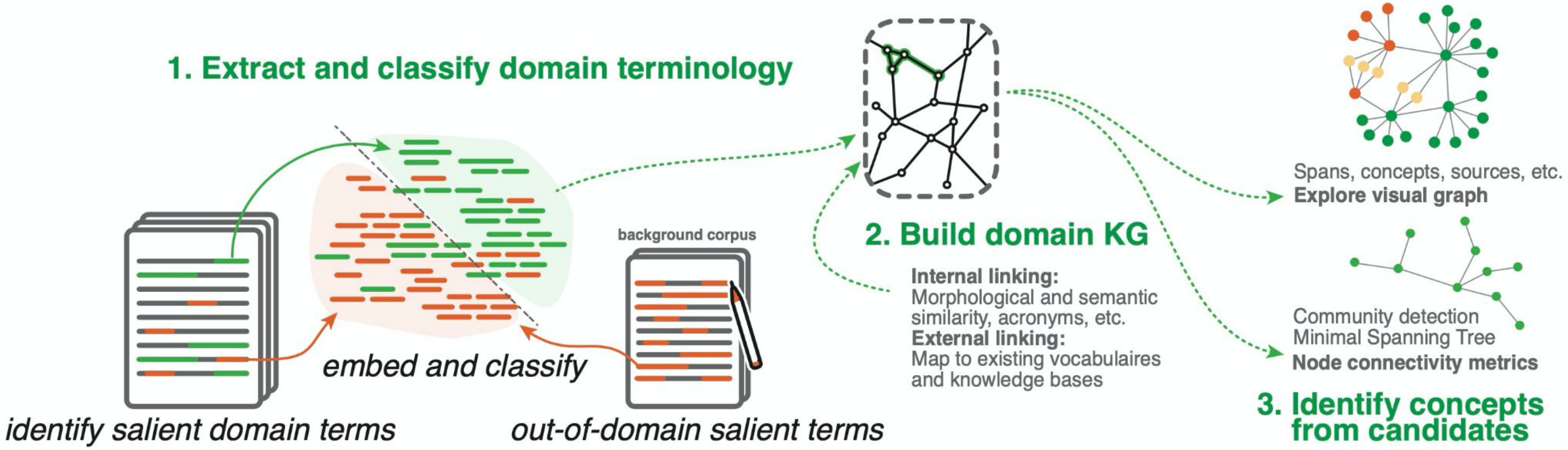
Scraping 16 vocabularies already results in 8K terms and definitions, (mostly civil engineering pdfs and restricted licensing)

Approach to automate the collection of terms

Approach

Generally, the first steps for deriving an ontology or taxonomy from text are:

1. Term extraction from text
2. Identify which terms are candidates for concepts
3. Internal and external concept linking

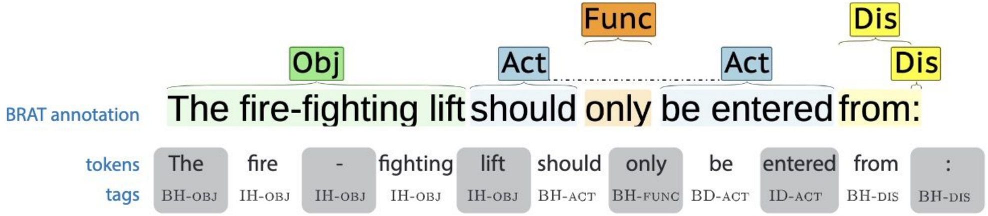


1. Extract and classify domain terminology
background corpus
embed and classify
identify salient domain terms
out-of-domain salient terms
2. Build domain KG
Internal linking:
Morphological and semantic similarity, acronyms, etc.
External linking:
Map to existing vocabulaires and knowledge bases
Spans, concepts, sources, etc.
Explore visual graph
Community detection
Minimal Spanning Tree
Node connectivity metrics
3. Identify concepts from candidates

1. Term extraction from building regulations

UK Merged Approved Documents

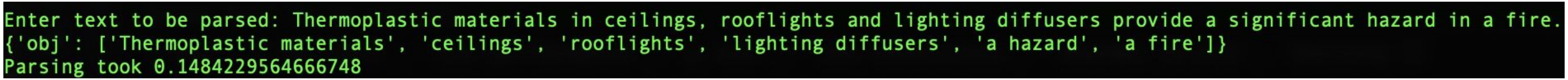
1. Convert PDF documents to a text-based format
2. Split texts into sentences
3. Run SPaR.txt to identify candidate terms for the KG
   - Aim is to capture Object spans
   - Including Multi-Word Expressions (MWE)



tokens	The	fire	-	fighting	lift	should	only	be	entered	from	:
tags	BH-OBJ	IH-OBJ	IH-OBJ	IH-OBJ	IH-OBJ	BH-ACT	BH-FUNC	BD-ACT	ID-ACT	BH-DIS	BH-DIS

Figure 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, see the tagging scheme described in Section 4.1. The identifier for this sentence in the dataset is '*d_2.14.4_i3_s_0*'.

- OBJECT spans indicate either real-world objects or distinguishable concepts. They include proper nouns, compounds, multi-word terms, and multi-word Named Entities, such as '*the Target Emissions Rating*', '*offensive fire-fighting*' and '*BS 8000-15: 1990*'. We include determiners as part of the OBJECT span during annotation, see Figure 2.

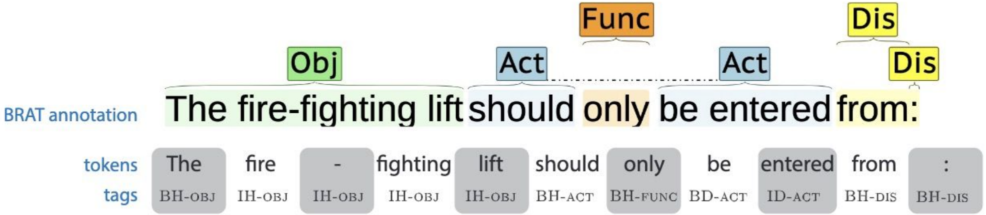

```
Enter text to be parsed: Thermoplastic materials in ceilings, rooflights and lighting diffusers provide a significant hazard in a fire.
{'obj': ['Thermoplastic materials', 'ceilings', 'rooflights', 'lighting diffusers', 'a hazard', 'a fire']}
Parsing took 0.1484229564666748
```

Data and code @ https://github.com/rubenkruiper/SPaR.txt

1. Term extraction from building regulations

UK Merged Approved Documents

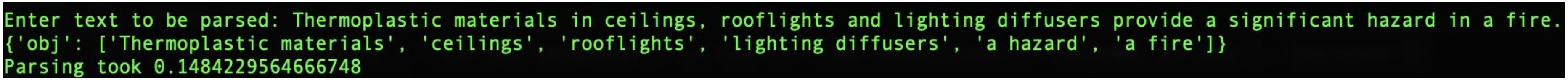
1. Convert PDF documents to a text-based format → includes noise
2. Split texts into sentences → more noise
3. Run SPaR.txt to identify candidate terms for the KG → lots of noise
   - Aim is to capture Object spans
   - Including Multi-Word Expressions (MWE)



tokens	The	fire	-	fighting	lift	should	only	be	entered	from	:
tags	BH-OBJ	IH-OBJ	IH-OBJ	IH-OBJ	IH-OBJ	BH-ACT	BH-FUNC	BD-ACT	ID-ACT	BH-DIS	BH-DIS

Figure 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, see the tagging scheme described in Section 4.1. The identifier for this sentence in the dataset is '*d_2.14.4_i3_s_0*'.

- OBJECT spans indicate either real-world objects or distinguishable concepts. They include proper nouns, compounds, multi-word terms, and multi-word Named Entities, such as '*the Target Emissions Rating*', '*offensive fire-fighting*' and '*BS 8000-15: 1990*'. We include determiners as part of the OBJECT span during annotation, see Figure 2.



```
Enter text to be parsed: Thermoplastic materials in ceilings, rooflights and lighting diffusers provide a significant hazard in a fire.
{'obj': ['Thermoplastic materials', 'ceilings', 'rooflights', 'lighting diffusers', 'a hazard', 'a fire']}
Parsing took 0.1484229564666748
```

Data and code @ https://github.com/rubenkruiper/SPaR.txt

ection 2B: Sizes of certai floors and roofs for dwe sk from house longhorn b

ing of members

Guidance on the sizing of certain members ors and roofs is given in 'Span tables for timber members in floors, ceilings and roofs uding trussed rafter roofs) for dwellings', shed by TRADA, available from Chiltern se, Stocking Lane, Hughenden Valley, High ombe, Bucks HP14 4ND.

native guidance is available in BS EN 1995-2004 Design of timber structures with its UK onal Annex and additional guidance given in Published Document PD 6693-1:2012 and BS 8103-3:2009 Structural design of low-buildings, Code of practice for timber floors roofs for housing.

Hous

2B2 Table 1 or fixec within t adequa house

Guidan is giver manua Specifi from 5(West Y

ble 1 **Areas at risk from house longhorn beetle**

graphical area
e Borough of Bracknell Forest the parishes of Sandhurst and Crowthorne.
Borough of Elmbridge
e District of Hart, the parishes of Hawley and Yateley
District of Runnymede
Borough of Spelthorne
Borough of Surrey Heath
e Borough of Rushmoor, the area of the former district of Farnborough
Borough of Woking

n 3: Surface water drainage

tion gives guidance on the ce water drainage systems. It is he drainage of small catchments s areas up to 2 hectares. For the ms serving larger catchments, ld be made to BS EN 752-4 3.36).

water drainage should discharge or other infiltration system ble.

ge to a watercourse may require a he Environment Agency, who may discharge. Maximum flow rates by provision of detention basins 3.35).

ther forms of outlet are not scharge should be made to a

systems

ewers carry both foul water and combined systems) in the same ey do the sewerage undertaker ce water to discharge into the ewer has enough capacity to take (see Approved Document H1 . Some private sewers (drains han one building that have not by the sewerage undertaker) also water and surface water. If a d as a combined system does not apacity, the surface water should arate system with its own outfall.

circumstances, where a sewer is combined system and has sufficient ate drainage should still be provided Document H5).

water drainage connected to ers should have traps on all inlets.

nfall intensities

rainfall intensities of 0.014 litres/ y be assumed for normal rnatively the rainfall intensity ed from Diagram 2.

ow levels of surface flooding could of buildings the rainfall intensities ined from BS EN 752-4 (see).

here is evidence of a liability from sewers, or levels in the he site make gravity connection surface water lifting equipment (see Approved Document H1 to 2.12).

Layout

3.11 Refer to paragraphs 2. Approved Document H1.

Depth of pipes

3.12 Refer to paragraphs 2.27 and 2.28 of Approved Document H1.

Pipe gradients and sizes

3.13 Drains should have enough capacity carry the flow. The capacity depends on the and gradients of the pipes.

3.14 Drains should be at least 75mm diam Surface water sewers (serving more than or building) should have a minimum size of 10 Diagram 3 shows the capacities of drains o various sizes at different gradients. Howeve capacity can be increased by increasing the gradient, or by using larger pipes.

3.15 75mm and 100mm rainwater drains s be laid at not less than 1:100. 150mm drain sewers should be laid at gradients not less 1:150 and 225mm drains should be laid at gra not less than 1:225. For minimum gradients larger pipes see BS EN 752-4 (see paragraph

Diagram 3 **Discharge capacities of rainwater drains running full**

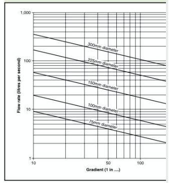


Diagram 8 **General air supply to a combustion appliance** (for sizes see Sections 2, 3 and 4)

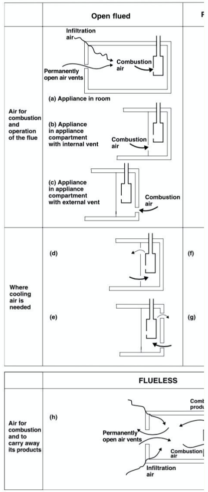

Open flued
Infiltration air
Combustion air
Permanently open air vents
(a) Appliance in room
Air for combustion and operation of the flue
(b) Appliance in appliance compartment with internal vent
Combustion air
(c) Appliance in appliance compartment with external vent
Combustion air
(d)
(f)
Where cooling air is needed
(e)
(g)
FLUELESS
Air for combustion and to carry away its products
(h)
Permanently open air vents
Comb prod
Combustion air
Infiltration air

Table 1.1 List of the approved documents and what they cover

Dwellings		Other buildings	
New	Existing[1]	New	Existing[1]
A: Structure			
B: Fire safety, Volume 1: Dwellings		**B:** Fire safety, Volume 2: Buildings other than dwellings	
C: Site preparation and resistance to contaminants and moisture			
D: Toxic substances			
E: Resistance to the passage of sound			
F: Ventilation			
G: Sanitation, hot water safety and water efficiency			
H: Drainage and waste disposal			
J: Combustion appliances and fuel storage systems			
K: Protection from falling, collision and impact			
L: Conservation of fuel and power **L1A** New dwellings	**L:** Conservation of fuel and power **L1B** Existing dwellings	**L:** Conservation of fuel and power **L2A** New buildings other than dwellings	**L:** Conservation of fuel and power **L2B** Existing buildings other than dwellings
M: Access to and use of buildings Volume 1: Dwellings		**M:** Access to and use of buildings Volume 2: Buildings other than dwellings	
P: Electrical safety – dwellings[2]		**P:** No approved document	
ement		**Q:** No requirement	
high-speed electronic communications networks			

hanges of use are covered in Table A2 in Volume 2.
e for other buildings if the supply is shared with a dwelling.

F1(2), R39, R44

Requirement F1(2) and regulations 39 and 44

This section deals with the requirements of Part F1(2) of Schedule 1 and regulations 39 and 44 of the Building Regulations 2010.

Requirement

Requirement	*Limits on application*
F1. (2) Fixed systems for mechanical ventilation and any associated controls must be commissioned by testing and adjusting as necessary to secure that the objective referred to in sub-paragraph (1) is met.	Requirement F1 does not apply to a building or space within a building: a. into which people do not normally go; b. which is used solely for storage; or c. which is a garage used solely in connection with a single dwelling.

Regulations

Information about ventilation

39. (1) This regulation applies where paragraph F1(1) of Schedule 1 imposes a requirement in relation to building work.

(2) The person carrying out the work shall not later than five days after the work has been completed give sufficient information to the owner about the building's ventilation system and its maintenance requirements so that the ventilation system can be operated in such a manner as to provide adequate means of ventilation.

Commissioning

44. (1) This regulation applies to building work in relation to which paragraph F1(2) of Schedule 1 imposes a requirement, but does not apply to the provision or extension of any fixed system for mechanical ventilation or any associated controls where testing and adjustment is not possible.

(2) This regulation also applies to building work in relation to which paragraph L1(b) of Schedule 1 imposes a requirement, but does not apply to the provision or extension of any fixed building service where testing and adjustment is not possible or would not affect the energy efficiency of that fixed building service.

(3) Where this regulation applies the person carrying out the work shall, for the purpose of ensuring

dings for stairs

ll buildings

For means of escape requirements, refer also to Approved Document B: Volume 1 – Dwellinghouses, and Volume 2 – Buildings other than dwellinghouses.

At the top and bottom of every flight, provide landings the width and length at least as great as the smallest width of the flight (see Diagram 1.6).

A landing:

a. may include part of the floor of the building

b. should be kept clear of permanent obstructions

c. may have doors to cupboards and ducts that open over a landing at the top of a flight, as shown in Diagram 1.7, but only when they are kept shut or locked shut when under normal use.

See para 1.21

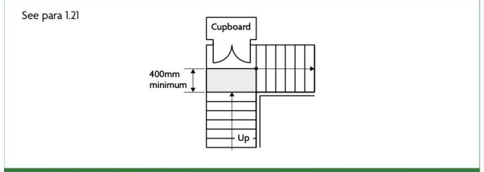


Diagram 1.7 Cupboard onto landing

Table B1 Indoor air pollutants guidance values[1][2]

Pollutant	Exposure limit	Exposure time	Guidance
Carbon monoxide (CO)	100mg/m^3	15-minute average	WHO, 2010
	30mg/m^3	1-hour average	WHO, 2010
	35mg/m^3 (occupational exposure)	8-hour average	HSE, 2020
Nitrogen dioxide (NO_2)	200µg/m^3	1-hour average	WHO, 2010
	40µg/m^3	1-year average	WHO, 2010
Formaldehyde (CH_2O)	100µg/m^3	30-minute average	WHO, 2010
	10µg/m^3	1-year average	PHE, 2019
TVOC[3]	300µg/m^3	8-hour average	ECA, 1992/WHO, 2010
Ozone	100µg/m^3		DETR, 1994

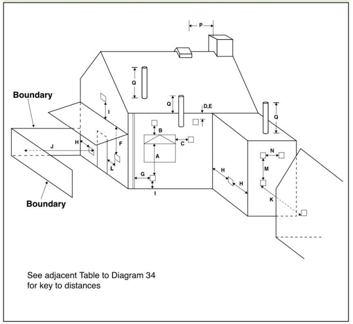
NOTES:

1. No safe levels can be recommended for benzene or trichloroethylene so they have not been considered in the definition of ventilation rates in buildings. The best strategy for reducing their concentration indoors may be to control them at source.
2. Even if the designer and builder choose to reduce volatile organic compound (VOC) levels in buildings by controlling them at source, the ventilati
3. The total volatile organic compound (T concentrations and should not be used for the purposes of ventilation control s be used where justified in accordance w

B3 As an alternative to using TVOC, the by robust independent evidence. Pub *Volatile Organic Compounds (VOCs)* to be more complex than testing aga

Where the Health and Safety Executiv followed in preference to the guidanc

Diagram 34 **Location of outlets from flues serving gas appliances**



See adjacent Table to Diagram 34 for key to distances

her means. The siting, spacing an e access points will depend on the pth and size of the runs.

47 The provisions described bel rmal methods of rodding (which n e direction of flow) and not mecha clearing.

48 Access points should be one of four types. bles 11 and 12 show the depth at which each pe should be used and the recommended mensions it should have. The dimensions ould be increased at junctions if they do not ow enough space for branches. The types are:

a. on or near the head of each drain run, and

b. at a bend and at a change of gradient, and

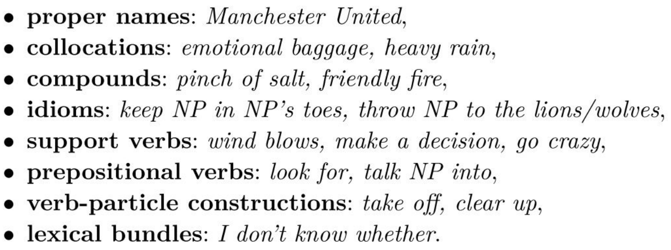
c. at a change of pipe size (but see below if it is at a junction), and

d. at a junction unless each run can be cleared from an access point (some junctions can only be rodded through from one direction).

able 11 **Minimum dimensions for access fittings and inspection chambers**

		Internal sizes		Cover sizes	
Type	**Depth to invert from cover level (m)**	**Length x width (mm x mm)**	**Circular (mm)**	**Length x width (mm x mm)**	**Circular (mm)**
Rodding eye		As drain but min. 100			Same size as pipework[1]
ccess fitting					
mall 150 diam.	0.6 or less, except where situated in a chamber				Same size as access fitting
150 x 100		150 x 100	150	150 x 100[1]	
rge 225 x 100		225 x 100	225	225 x 100[1]	
nspection chamber					
shallow	0.6 or less	225 x 100	190[2]	–	190[1]
	1.2 or less	450 x 450	450	Min. 430 x 430	430
deep	> 1.2	450 x 450	450	Max. 300 x 300[3]	Access restricted to max. 350[3]

1. Multi-Word Expressions



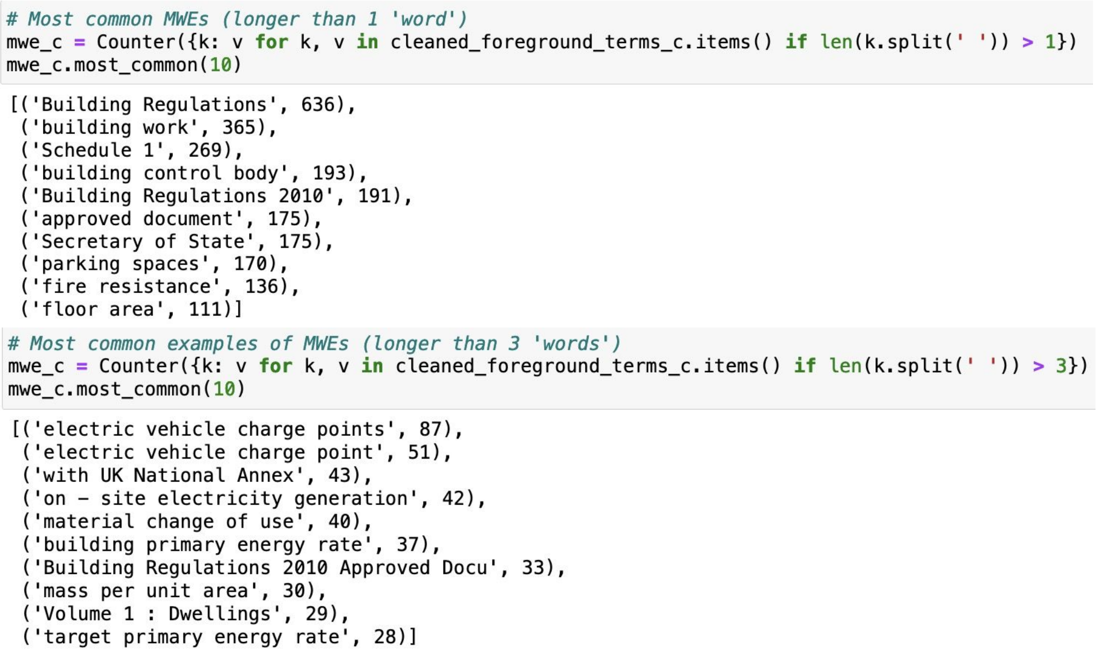
- **proper names**: *Manchester United*,
- **collocations**: *emotional baggage, heavy rain*,
- **compounds**: *pinch of salt, friendly fire*,
- **idioms**: *keep NP in NP's toes, throw NP to the lions/wolves*,
- **support verbs**: *wind blows, make a decision, go crazy*,
- **prepositional verbs**: *look for, talk NP into*,
- **verb-particle constructions**: *take off, clear up*,
- **lexical bundles**: *I don't know whether*.

Source: Villavicencio and Idiart (2019)

Many *'entities'* in the regulations consist of multiple words.

- Just think of the different types of wall, roof, beam, etc.
- All of these can have plural, acronyms, alternative labels, alternative spelling, etc.

58.36% of the unique filtered SPaR.txt concepts are MWEs.



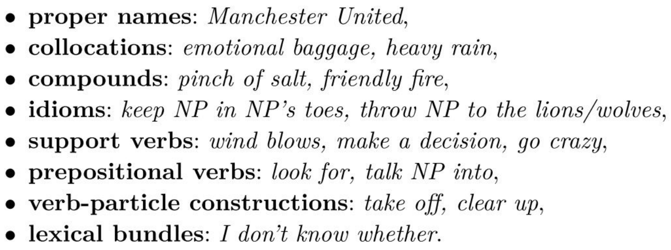
```
# Most common MWEs (longer than 1 'word')
mwe_c = Counter({k: v for k, v in cleaned_foreground_terms_c.items() if len(k.split(' ')) > 1})
mwe_c.most_common(10)
```

```
[('Building Regulations', 636),
 ('building work', 365),
 ('Schedule 1', 269),
 ('building control body', 193),
 ('Building Regulations 2010', 191),
 ('approved document', 175),
 ('Secretary of State', 175),
 ('parking spaces', 170),
 ('fire resistance', 136),
 ('floor area', 111)]
```

```
# Most common examples of MWEs (longer than 3 'words')
mwe_c = Counter({k: v for k, v in cleaned_foreground_terms_c.items() if len(k.split(' ')) > 3})
mwe_c.most_common(10)
```

```
[('electric vehicle charge points', 87),
 ('electric vehicle charge point', 51),
 ('with UK National Annex', 43),
 ('on - site electricity generation', 42),
 ('material change of use', 40),
 ('building primary energy rate', 37),
 ('Building Regulations 2010 Approved Docu', 33),
 ('mass per unit area', 30),
 ('Volume 1 : Dwellings', 29),
 ('target primary energy rate', 28)]
```

1. Multi-Word Expressions



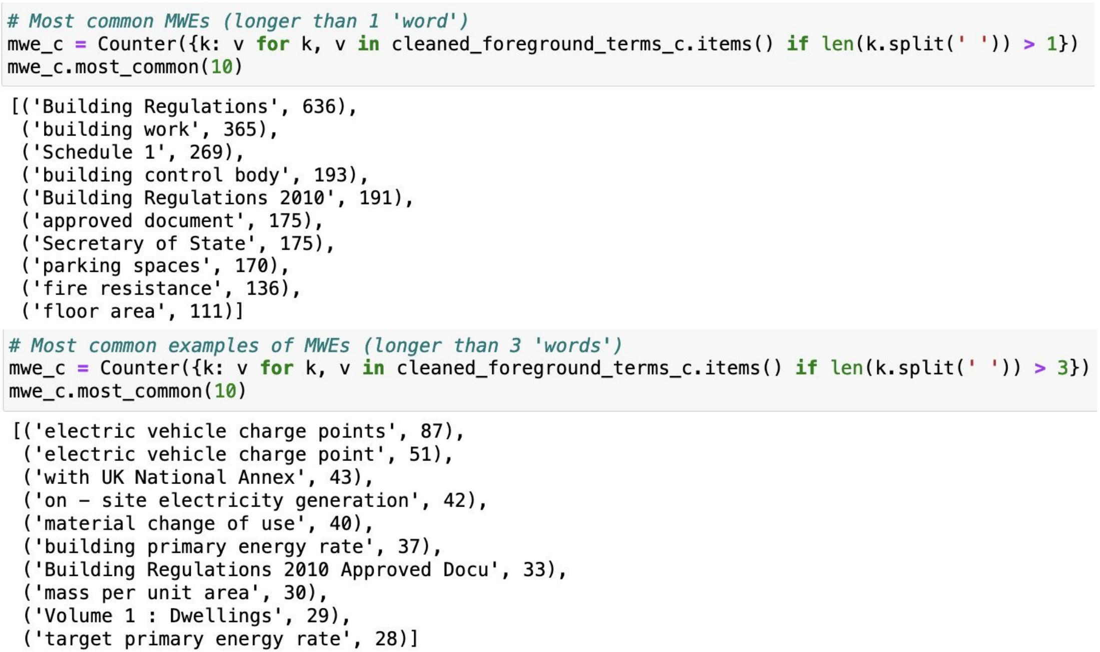
- **proper names**: *Manchester United*,
- **collocations**: *emotional baggage, heavy rain*,
- **compounds**: *pinch of salt, friendly fire*,
- **idioms**: *keep NP in NP's toes, throw NP to the lions/wolves*,
- **support verbs**: *wind blows, make a decision, go crazy*,
- **prepositional verbs**: *look for, talk NP into*,
- **verb-particle constructions**: *take off, clear up*,
- **lexical bundles**: *I don't know whether*.

Source: Villavicencio and Idiart (2019)

Many *'entities'* in the regulations consist of multiple words.

- Just think of the different types of wall, roof, beam, etc.
- All of these can have plural, acronyms, alternative labels, alternative spelling, etc.

58.36% of the unique filtered SPaR.txt concepts are MWEs.



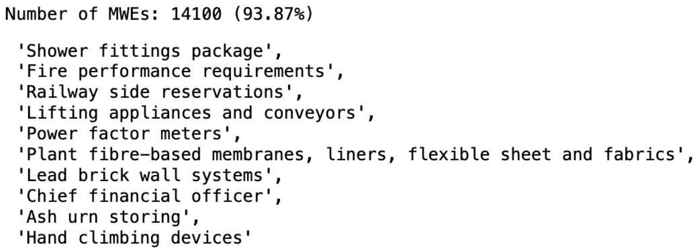
```
# Most common MWEs (longer than 1 'word')
mwe_c = Counter({k: v for k, v in cleaned_foreground_terms_c.items() if len(k.split(' ')) > 1})
mwe_c.most_common(10)
```

```
[('Building Regulations', 636),
 ('building work', 365),
 ('Schedule 1', 269),
 ('building control body', 193),
 ('Building Regulations 2010', 191),
 ('approved document', 175),
 ('Secretary of State', 175),
 ('parking spaces', 170),
 ('fire resistance', 136),
 ('floor area', 111)]
```

```
# Most common examples of MWEs (longer than 3 'words')
mwe_c = Counter({k: v for k, v in cleaned_foreground_terms_c.items() if len(k.split(' ')) > 3})
mwe_c.most_common(10)
```

```
[('electric vehicle charge points', 87),
 ('electric vehicle charge point', 51),
 ('with UK National Annex', 43),
 ('on - site electricity generation', 42),
 ('material change of use', 40),
 ('building primary energy rate', 37),
 ('Building Regulations 2010 Approved Docu', 33),
 ('mass per unit area', 30),
 ('Volume 1 : Dwellings', 29),
 ('target primary energy rate', 28)]
```

Notably, Uniclass mostly MWEs →



```
Number of MWEs: 14100 (93.87%)

 'Shower fittings package',
 'Fire performance requirements',
 'Railway side reservations',
 'Lifting appliances and conveyors',
 'Power factor meters',
 'Plant fibre-based membranes, liners, flexible sheet and fabrics',
 'Lead brick wall systems',
 'Chief financial officer',
 'Ash urn storing',
 'Hand climbing devices'
```

2. Identify which terms are candidates for concepts

Which terms are relevant to the building domain?

Foreground corpus:
Merged Approved Documents (MAD)

Background corpus:
5 EU regulations concerning medical devices

1. Similar style of text, yet different domains
2. Not an extreme size difference (<10x)

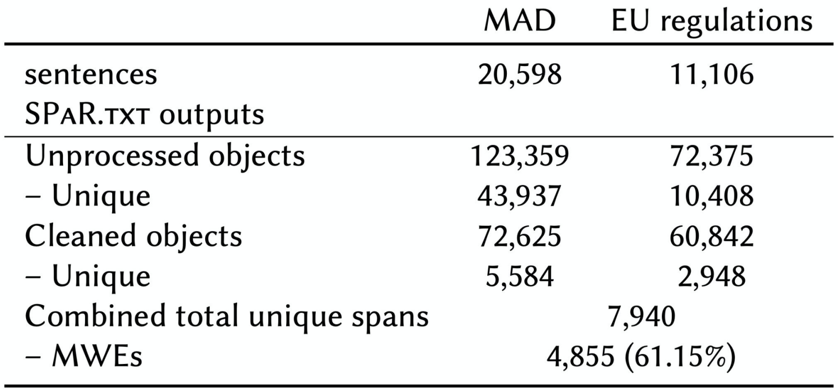
Also, both corpora openly available.

MAD top 10

('building', 1824),
('buildings', 934),
('guidance', 887),
('requirements', 641),
('Building Regulations', 636),
('dwelling', 500),
('work', 490),
('dwellings', 385),
('building work', 365),
('document', 351)

EU med top 10

('device', 1885),
('devices', 1696),
('manufacturer', 1587),
('notified body', 1199),
('information', 799),
('Member States', 686),
('Commission', 652),
('requirements', 604),
('Regulation', 596),
('market', 475)



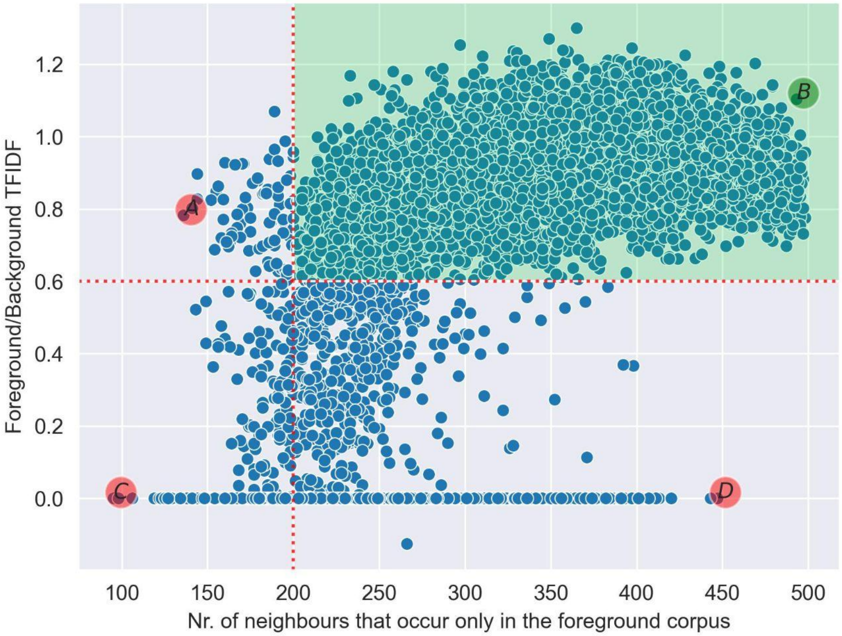
	MAD	EU regulations
sentences	20,598	11,106
SPaR.txt outputs		
Unprocessed objects	123,359	72,375
– Unique	43,937	10,408
Cleaned objects	72,625	60,842
– Unique	5,584	2,948
Combined total unique spans	7,940	
– MWEs	4,855 (61.15%)	

2. Simple domain classification

Term Frequency-Inverse Document Frequency (TF-IDF) metric:

$$TF\text{-}IDF(t) = log(1 + \frac{fc_t}{fc_t + bc_t}) * log(avgIDF_t) \tag{1}$$

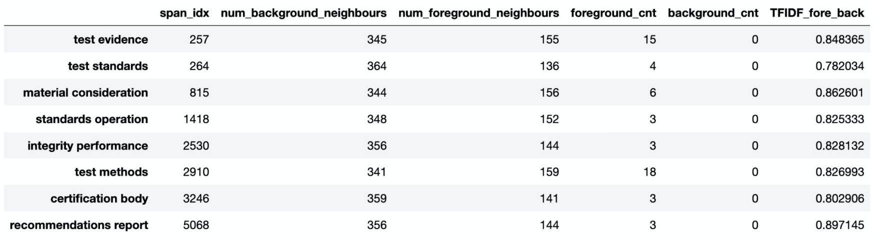
with fc_t the number of times term t occurs in the foreground corpus, bc_t the background corpus count, and $avgIDF$ the averaged IDF weight over the subword tokens of term t.



Expected domains and example terms:

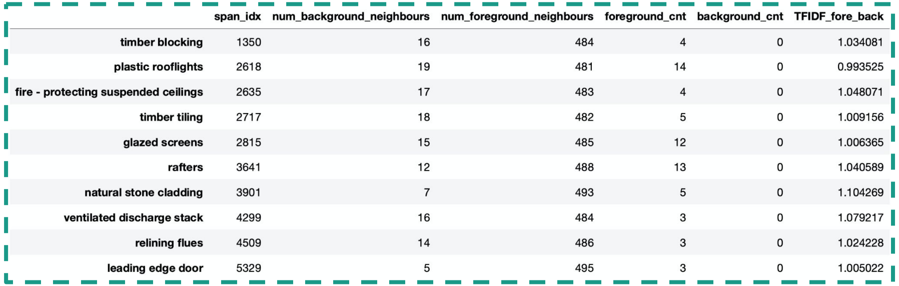
A) general domain:	test standards
B) building regulations:	natural stone cladding
C) medical domain:	clinical investigation
D) general domain:	artificial opening

2. Top examples domain classification



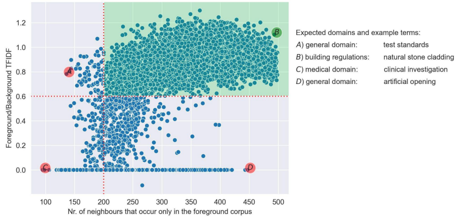
	span_idx	num_background_neighbours	num_foreground_neighbours	foreground_cnt	background_cnt	TFIDF_fore_back
test evidence	257	345	155	15	0	0.848365
test standards	264	364	136	4	0	0.782034
material consideration	815	344	156	6	0	0.862601
standards operation	1418	348	152	3	0	0.825333
integrity performance	2530	356	144	3	0	0.828132
test methods	2910	341	159	18	0	0.826993
certification body	3246	359	141	3	0	0.802906
recommendations report	5068	356	144	3	0	0.897145

Expecting general domain

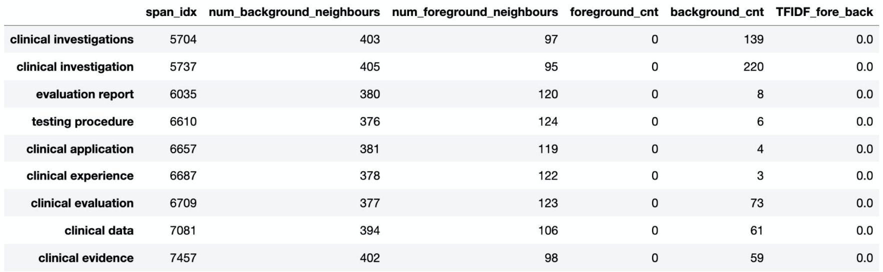


	span_idx	num_background_neighbours	num_foreground_neighbours	foreground_cnt	background_cnt	TFIDF_fore_back
timber blocking	1350	16	484	4	0	1.034081
plastic rooflights	2618	19	481	14	0	0.993525
fire - protecting suspended ceilings	2635	17	483	4	0	1.048071
timber tiling	2717	18	482	5	0	1.009156
glazed screens	2815	15	485	12	0	1.006365
rafters	3641	12	488	13	0	1.040589
natural stone cladding	3901	7	493	5	0	1.104269
ventilated discharge stack	4299	16	484	3	0	1.079217
relining flues	4509	14	486	3	0	1.024228
leading edge door	5329	5	495	3	0	1.005022

Expecting candidate for concepts within building regulation domain

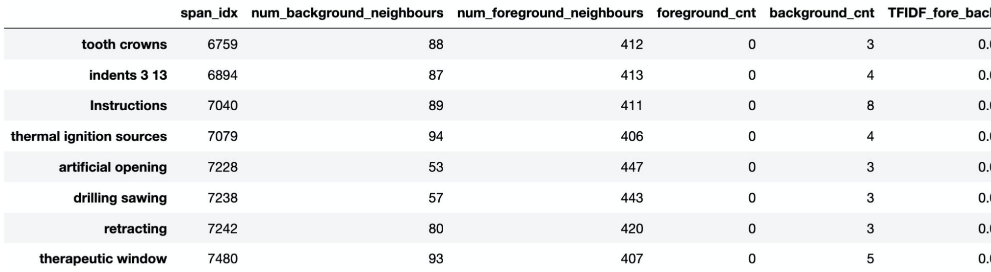


Expecting medical device domain



	span_idx	num_background_neighbours	num_foreground_neighbours	foreground_cnt	background_cnt	TFIDF_fore_back
clinical investigations	5704	403	97	0	139	0.0
clinical investigation	5737	405	95	0	220	0.0
evaluation report	6035	380	120	0	8	0.0
testing procedure	6610	376	124	0	6	0.0
clinical application	6657	381	119	0	4	0.0
clinical experience	6687	378	122	0	3	0.0
clinical evaluation	6709	377	123	0	73	0.0
clinical data	7081	394	106	0	61	0.0
clinical evidence	7457	402	98	0	59	0.0

Expecting general domain



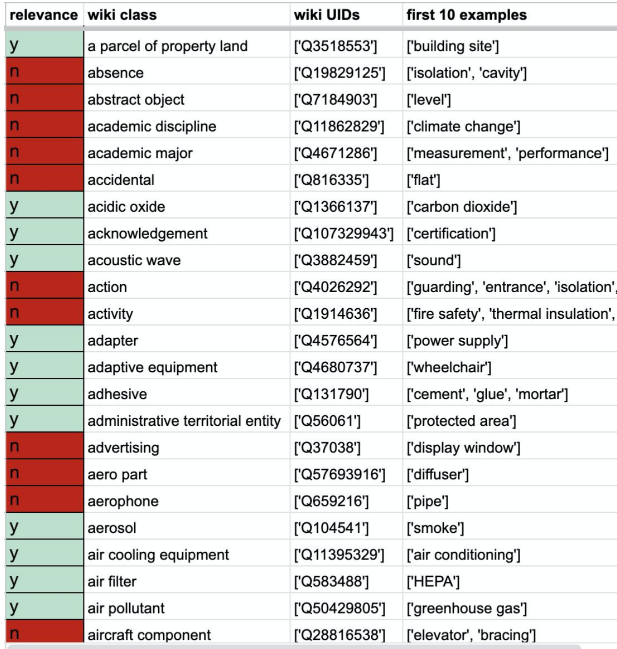
	span_idx	num_background_neighbours	num_foreground_neighbours	foreground_cnt	background_cnt	TFIDF_fore_bac
tooth crowns	6759	88	412	0	3	0.
indents 3 13	6894	87	413	0	4	0.
Instructions	7040	89	411	0	8	0.
thermal ignition sources	7079	94	406	0	4	0.
artificial opening	7228	53	447	0	3	0.
drilling sawing	7238	57	443	0	3	0.
retracting	7242	80	420	0	3	0.
therapeutic window	7480	93	407	0	5	0.

4,958 domain and 2,982 general/out-of-domain terms

3. Internal and external linking

Which of those 5K '*candidates for concepts*' occur in other vocabularies?

- Uniclass
  - only 598 (4%) of the 15K terms occur verbatim in the 1.274 pages of the UK Merged Approved documents (MAD)
- WikiData
  - 29% of our 5K domain terms found in WikiData
  - Many WikiData classes and definitions irrelevant
  - Annotate 1.2K WikiData classes (46% irrelevant)
  - When only retaining relevant WikiData matches, 13% of our candidate concepts found in WikiData

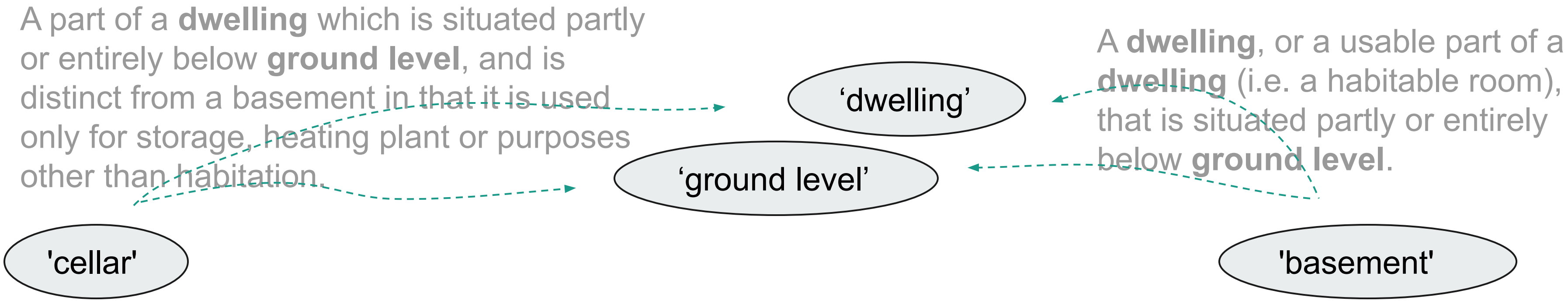


relevance	wiki class	wiki UIDs	first 10 examples
y	a parcel of property land	['Q3518553']	['building site']
n	absence	['Q19829125']	['isolation', 'cavity']
n	abstract object	['Q7184903']	['level']
n	academic discipline	['Q11862829']	['climate change']
n	academic major	['Q4671286']	['measurement', 'performance']
n	accidental	['Q816335']	['flat']
y	acidic oxide	['Q1366137']	['carbon dioxide']
y	acknowledgement	['Q107329943']	['certification']
y	acoustic wave	['Q3882459']	['sound']
n	action	['Q4026292']	['guarding', 'entrance', 'isolation',
n	activity	['Q1914636']	['fire safety', 'thermal insulation',
y	adapter	['Q4576564']	['power supply']
y	adaptive equipment	['Q4680737']	['wheelchair']
y	adhesive	['Q131790']	['cement', 'glue', 'mortar']
y	administrative territorial entity	['Q56061']	['protected area']
n	advertising	['Q37038']	['display window']
n	aero part	['Q57693916']	['diffuser']
n	aerophone	['Q659216']	['pipe']
y	aerosol	['Q104541']	['smoke']
y	air cooling equipment	['Q11395329']	['air conditioning']
y	air filter	['Q583488']	['HEPA']
y	air pollutant	['Q50429805']	['greenhouse gas']
n	aircraft component	['Q28816538']	['elevator', 'bracing']

3. Internal and external linking

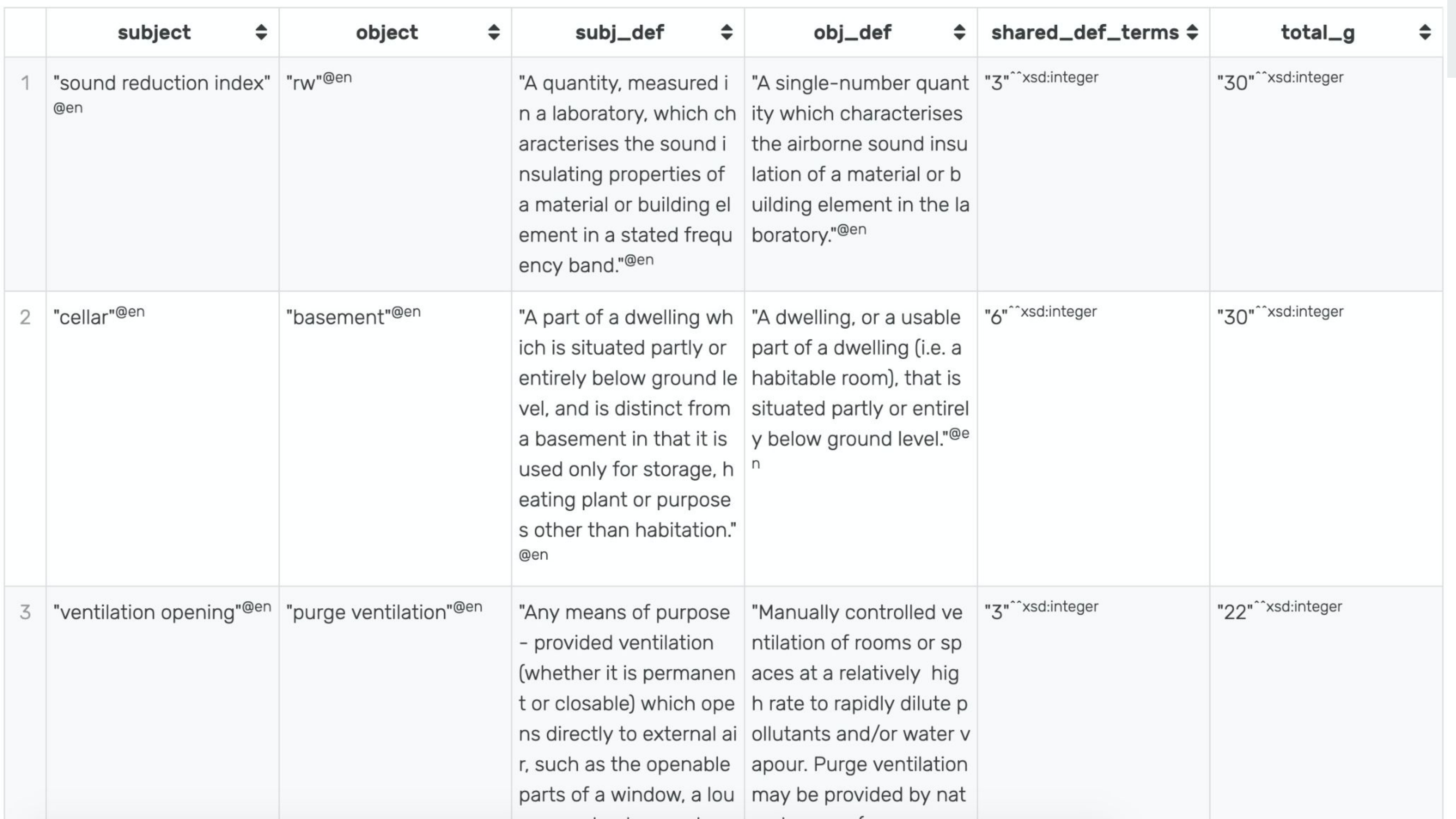
So far matches all based on exact overlap, we add:

- Morphological similarity
  - e.g., ‘*structural element*’ is morphologically similar to ‘*element of structure*’.
- Semantic similarity based on distributed representations
  - 5 nearest neighbours based on avg. weighted embedding of spans
- Potential acronyms and antonym-based similarity
- Number of shared terms among definitions:



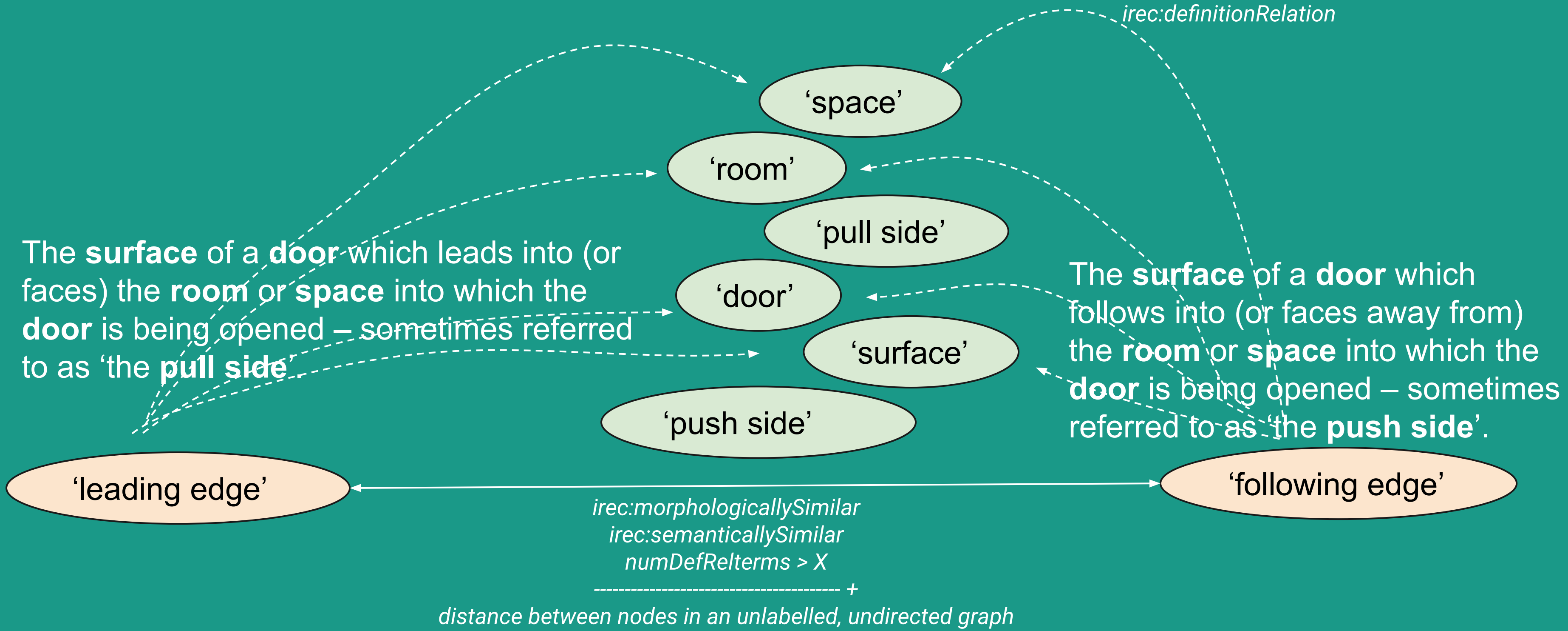
A part of a **dwelling** which is situated partly or entirely below **ground level**, and is distinct from a basement in that it is used only for storage, heating plant or purposes other than habitation.

A **dwelling**, or a usable part of a **dwelling** (i.e. a habitable room), that is situated partly or entirely below **ground level**.



	subject	object	subj_def	obj_def	shared_def_terms	total_g
1	"sound reduction index"@en	"rw"@en	"A quantity, measured in a laboratory, which characterises the sound insulating properties of a material or building element in a stated frequency band."@en	"A single-number quantity which characterises the airborne sound insulation of a material or building element in the laboratory."@en	"3"^^xsd:integer	"30"^^xsd:integer
2	"cellar"@en	"basement"@en	"A part of a dwelling which is situated partly or entirely below ground level, and is distinct from a basement in that it is used only for storage, heating plant or purposes other than habitation."@en	"A dwelling, or a usable part of a dwelling (i.e. a habitable room), that is situated partly or entirely below ground level."@en	"6"^^xsd:integer	"30"^^xsd:integer
3	"ventilation opening"@en	"purge ventilation"@en	"Any means of purpose - provided ventilation (whether it is permanent or closable) which opens directly to external air, such as the openable parts of a window, a lou	"Manually controlled ventilation of rooms or spaces at a relatively high rate to rapidly dilute pollutants and/or water vapour. Purge ventilation may be provided by nat	"3"^^xsd:integer	"22"^^xsd:integer

Using the KG to elucidate salient terms



irec:definitionRelation

‘space’

‘room’

‘pull side’

‘door’

‘surface’

‘push side’

The **surface** of a **door** which leads into (or faces) the **room** or **space** into which the **door** is being opened – sometimes referred to as ‘the **pull side**’.

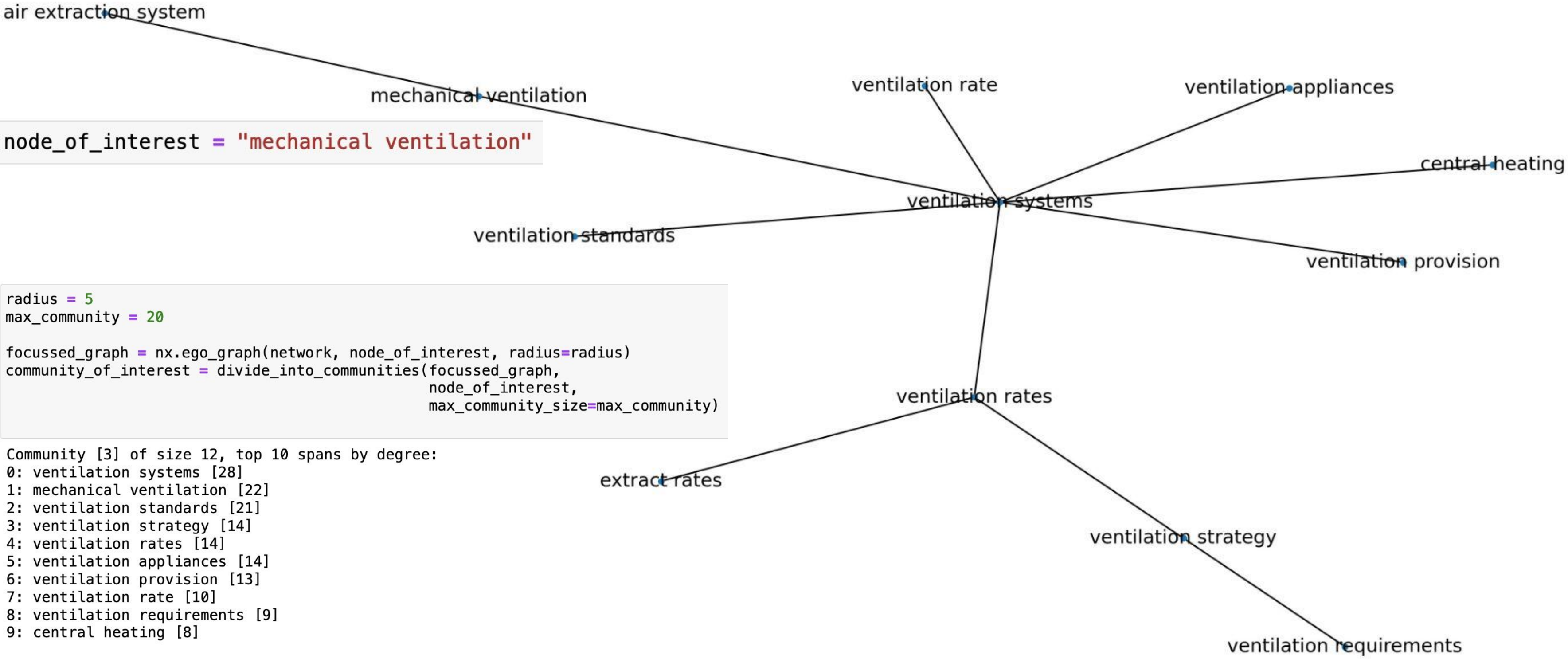
The **surface** of a **door** which follows into (or faces away from) the **room** or **space** into which the **door** is being opened – sometimes referred to as ‘the **push side**’.

‘leading edge’

‘following edge’

irec:morphologicallySimilar
irec:semanticallySimilar
numDefRelterms > X
-- +
distance between nodes in an unlabelled, undirected graph

Minimum Spanning Tree



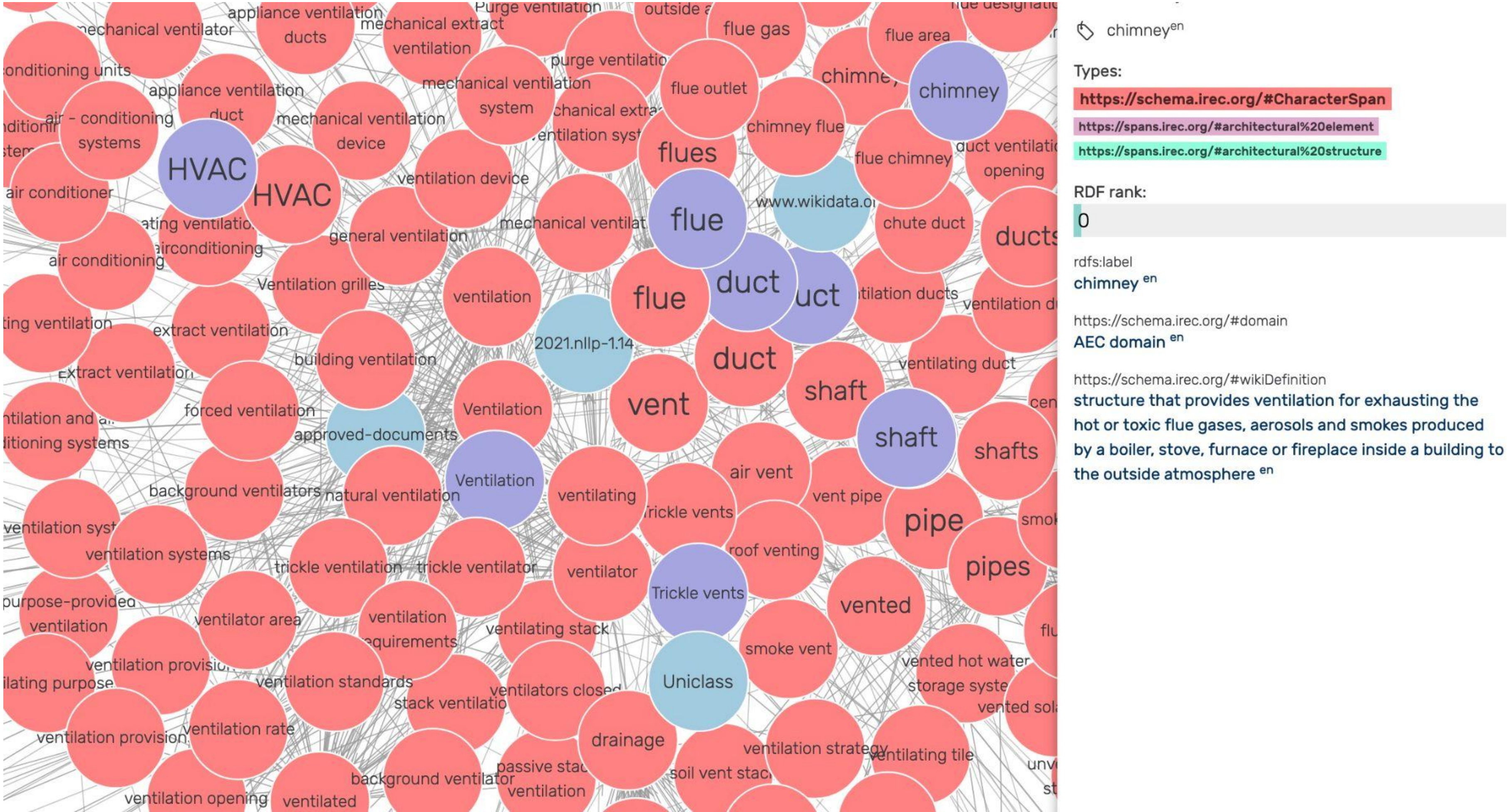
```
node_of_interest = "mechanical ventilation"
```

```
radius = 5
max_community = 20

focussed_graph = nx.ego_graph(network, node_of_interest, radius=radius)
community_of_interest = divide_into_communities(focussed_graph,
                                                node_of_interest,
                                                max_community_size=max_community)
```

```
Community [3] of size 12, top 10 spans by degree:
0: ventilation systems [28]
1: mechanical ventilation [22]
2: ventilation standards [21]
3: ventilation strategy [14]
4: ventilation rates [14]
5: ventilation appliances [14]
6: ventilation provision [13]
7: ventilation rate [10]
8: ventilation requirements [9]
9: central heating [8]
```

GraphDB



chimney^{en}

Types:

https://schema.irec.org/#CharacterSpan

https://spans.irec.org/#architectural%20element

https://spans.irec.org/#architectural%20structure

RDF rank:

0

rdfs:label

chimney ^{en}

https://schema.irec.org/#domain

AEC domain ^{en}

https://schema.irec.org/#wikiDefinition

structure that provides ventilation for exhausting the hot or toxic flue gases, aerosols and smokes produced by a boiler, stove, furnace or fireplace inside a building to the outside atmosphere ^{en}

Manual vs automated term collection

Manual

Issues include:

- 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

Benefits include:

- Complete control over terms and relations that are part of the KG

Support from automated term collection

Benefits include:

→ Source and provenance of terms can be tracked in the KG

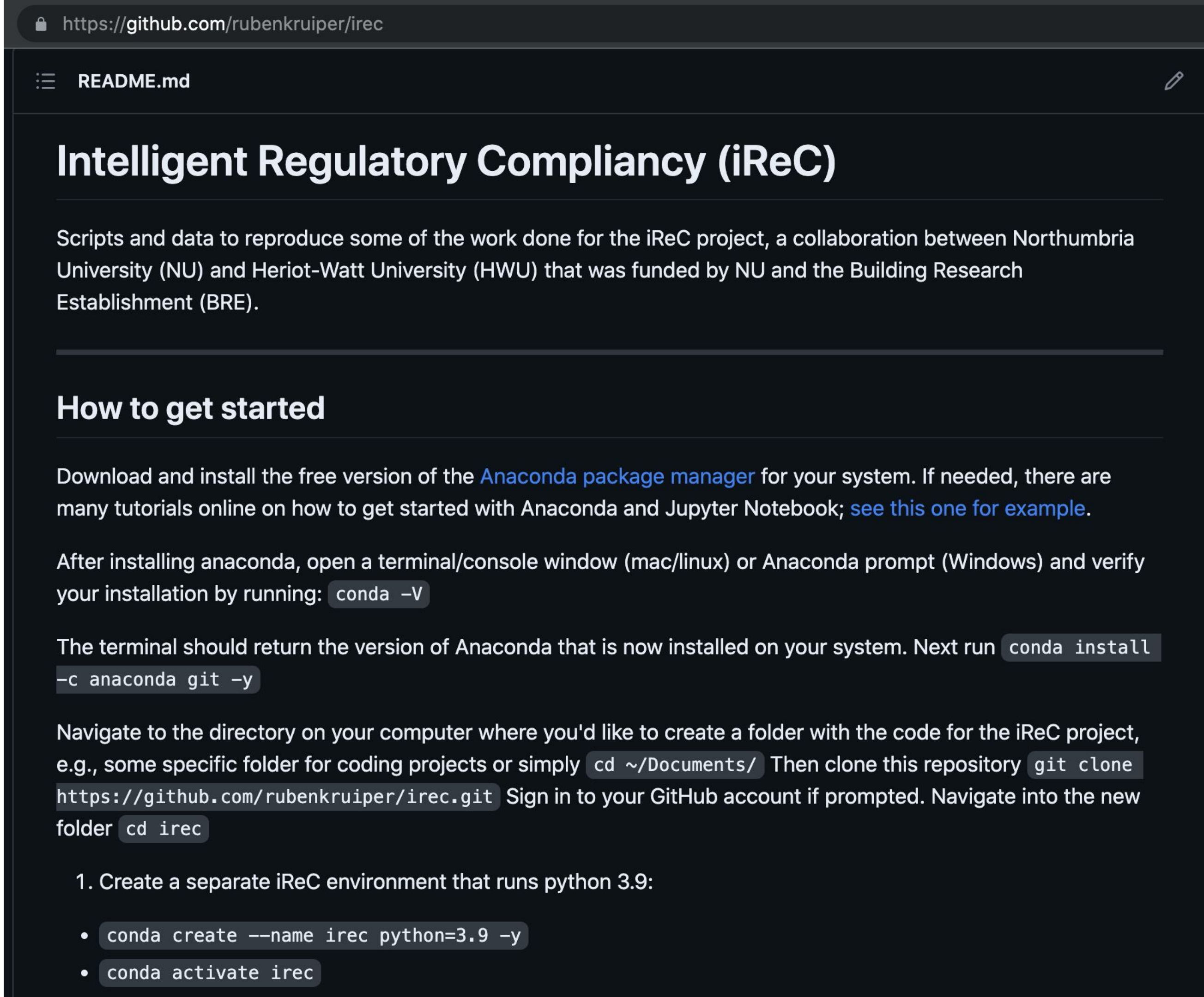
→ Scalable approach (excl. some span-span metrics), can be assumed to be reasonably comprehensive if input is representative

→ Easy to identify related terms, especially when definitions are present (even from less reliable sources like WikiData)

Issues include:

→ Contains noise, mostly the type of noise a human annotator has to filter

Thank you!



Intelligent Regulatory Compliancy (iReC)

Scripts and data to reproduce some of the work done for the iReC project, a collaboration between Northumbria University (NU) and Heriot-Watt University (HWU) that was funded by NU and the Building Research Establishment (BRE).

How to get started

Download and install the free version of the Anaconda package manager for your system. If needed, there are many tutorials online on how to get started with Anaconda and Jupyter Notebook; see this one for example.

After installing anaconda, open a terminal/console window (mac/linux) or Anaconda prompt (Windows) and verify your installation by running: `conda -V`

The terminal should return the version of Anaconda that is now installed on your system. Next run `conda install -c anaconda git -y`

Navigate to the directory on your computer where you'd like to create a folder with the code for the iReC project, e.g., some specific folder for coding projects or simply `cd ~/Documents/` Then clone this repository `git clone https://github.com/rubenkruiper/irec.git` Sign in to your GitHub account if prompted. Navigate into the new folder `cd irec`

1. Create a separate iReC environment that runs python 3.9:

- `conda create --name irec python=3.9 -y`
- `conda activate irec`