Geosemantics:
Text Mining and Linked Data at the British Geological Survey

Jo Walsh – Senior Software Engineer, British Geological Survey
jowalsh@bgs.ac.uk

British Geological Survey (BGS)

The British Survey can claim to be the oldest such undertaking to have functioned continuously since its inception in May 1832 and formal establishment on 11 July 1835.

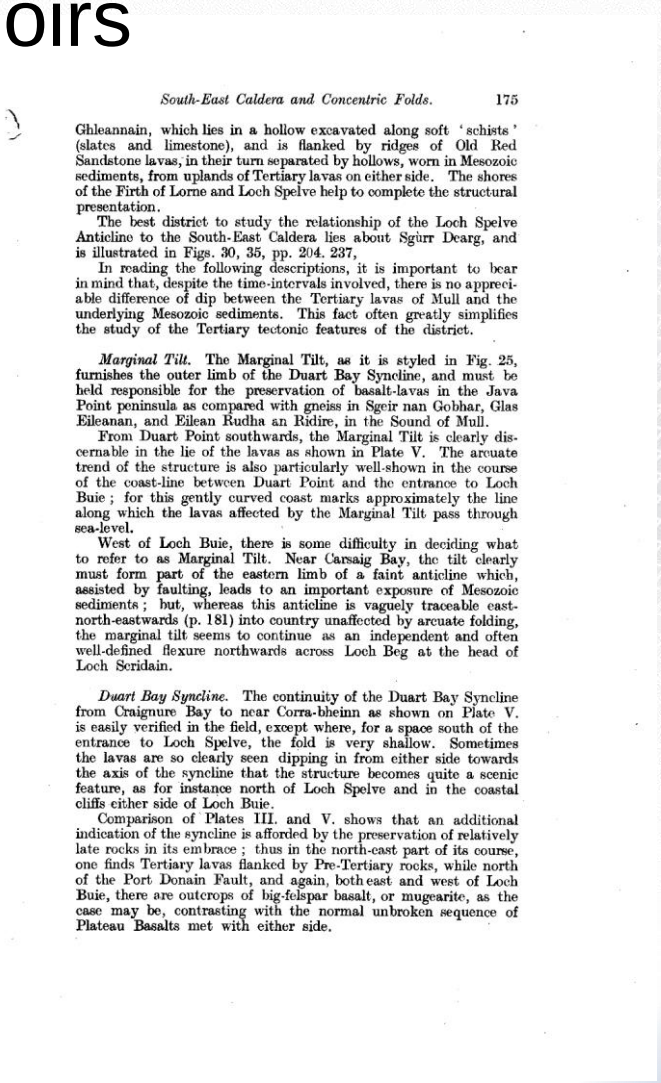
Sir Henry Thomas De la Beche, by William Brockedon, 1842 (© National Portrait Gallery, London).

Geosemantics

- Natural Language Processing with open source tools
- Named Entity Recognition model from labelled text
- Linked Open Data vocabularies for BGS resources
- Classifying relations between Named Entities

Geological Memoirs

The memoirs, compiled by BGS geologists, date from the late 1890s to present day and provide a comprehensive and detailed account of all aspects of the geology of an area.



South-East Caldera and Concentric Folds. 175

Ghleannain, which lies in a hollow excavated along soft 'schists' (slates and limestone), and is flanked by ridges of Old Red Sandstone lavas, in their turn separated by hollows, worn in Mesozoic sediments, from uplands of Tertiary lavas on either side. The shores of the Firth of Lorne and Loch Spelve help to complete the structural presentation.

The best district to study the relationship of the Loch Spelve Anticline to the South-East Caldera lies about Sgùrr Dearg, and is illustrated in Figs. 30, 35, pp. 204. 237,

In reading the following descriptions, it is important to bear in mind that, despite the time-intervals involved, there is no appreciable difference of dip between the Tertiary lavas of Mull and the underlying Mesozoic sediments. This fact often greatly simplifies the study of the Tertiary tectonic features of the district.

Marginal Tilt. The Marginal Tilt, as it is styled in Fig. 25, furnishes the outer limb of the Duart Bay Syncline, and must be held responsible for the preservation of basalt-lavas in the Java Point peninsula as compared with gneiss in Sgeir nan Gobhar, Glas Eileanan, and Eilean Rudha an Ridire, in the Sound of Mull.

From Duart Point southwards, the Marginal Tilt is clearly discernable in the lie of the lavas as shown in Plate V. The arcuate trend of the structure is also particularly well-shown in the course of the coast-line between Duart Point and the entrance to Loch Buie; for this gently curved coast marks approximately the line along which the lavas affected by the Marginal Tilt pass through sea-level.

West of Loch Buie, there is some difficulty in deciding what to refer to as Marginal Tilt. Near Carsaig Bay, the tilt clearly must form part of the eastern limb of a faint anticline which, assisted by faulting, leads to an important exposure of Mesozoic sediments; but, whereas this anticline is vaguely traceable east-north-eastwards (p. 181) into country unaffected by arcuate folding, the marginal tilt seems to continue as an independent and often well-defined flexure northwards across Loch Beg at the head of Loch Scridain.

Duart Bay Syncline. The continuity of the Duart Bay Syncline from Craignure Bay to near Corra-bheinn as shown on Plate V. is easily verified in the field, except where, for a space south of the entrance to Loch Spelve, the fold is very shallow. Sometimes the lavas are so clearly seen dipping in from either side towards the axis of the syncline that the structure becomes quite a scenic feature, as for instance north of Loch Spelve and in the coastal cliffs either side of Loch Buie.

Comparison of Plates III. and V. shows that an additional indication of the syncline is afforded by the preservation of relatively late rocks in its embrace; thus in the north-east part of its course, one finds Tertiary lavas flanked by Pre-Tertiary rocks, while north of the Port Donain Fault, and again, both east and west of Loch Buie, there are outcrops of big-felspar basalt, or mugearite, as the case may be, contrasting with the normal unbroken sequence of Plateau Basalts met with either side.

Geological memoirs

Scanned texts processed with optical character recognition software have been corrected by hand and annotated with identifiers for other BGS data (borehole logs, core samples)

turbidites are up to 5 mm thick, but are commonly discontinuous and lenticular, occurring as a series of low-amplitude, starved ripples.

The proportion of hemipelagite in the formation varies from 30 to 70 per cent. Individual beds range up to 5 cm thick. Although predominantly of the laminated variety, rare burrowed hemipelagites are locally present. They are most common in the top few metres of the formation, beneath the Derwenlas Formation.

Turbidite sandstone/mudstone couplets (type C), occurring either scattered or as thin bundles, are largely restricted to regions of interdigitation with the Caban Conglomerate and Ystrad Meurig Grits formations. Individual sandstones, in type Ci and Cii turbidites, are up to 20 cm and 5 cm thick, respectively.

Biostratigraphy

Graptolites from the Rhayader sheet area indicate that the Cwmere Formation ranges in age from the *persculptus* Biozone to the *magnus* Biozone. The Mottled Mudstone Member is entirely of *persculptus* Biozone age. The distribution and approximate thicknesses of individual biozones are depicted in Figure 19. Faunal lists from localities (Figures 18a and 21) in both the Cwmere Formation and the intercalations of Caban Conglomerate Formation are given in Tables 9 and 12.

On the north-western limb of the Tywi Anticline, north-east of the Afon Elan, the mudstones overlying facies of the Caban Conglomerate Formation range from possibly as low as upper *acuminatus* Biozone through to the *magnus* Biozone. There, in the Wye valley (Figure 19; Table 9), *atavus* Biozone graptolites have been found 25 m above the base of these mudstones in the lowest part of the A470 road section [SN 9799 6738 to SN 9794 6749]. Graptolites from the remainder of the A470 section and from a stratigraphically higher section in the River Wye at Ddole Farm [SN 9776 6737 to SN 9762 6745], have demonstrated that the *acinaces*, *cyphus* and *triangulatus* biozones are 75 m, 47 m and 20 m thick respectively (Table 9). The top 14 m of the mudstones contain *magnus* Biozone graptolites.

In the Llanilar sheet area, *magnus* Biozone graptolites also occur in the uppermost part of the Cwmere Formation, whilst upper *acuminatus* Biozone faunas have been proved in the uppermost part of the underlying tongue of Ystrad Meurig Grits Formation (Clwarel Goch facies) (Table 13). Although graptolite faunas have been found in the remainder of the formation (Table 13) none of the intervening biozones has been definitely proved.

The only shelly taxon from the Cwmere Formation in the district is a probable nowakiid tentaculitoid, *Nowakia gorgensis*, from the Wye valley [SN 9799 6741] east of Rhayader. The locality is reassigned here to the *acinaces* Biozone rather than the *atavus* Biozone reported by Tunnicliff (1989).

The mudstones in Nant Paradwys [SN 8918 6088 to SN 8933 6002] have yielded sphaeromorph acritarchs and indeterminate acanthomorphs of no biostratigraphical value (Appendix 2bii). More diverse assemblages including *Dicyotidium dictyotum*, *Lophosphaeridium parvearium*, *Multiplicisphaeridium fisheri* and *Tylotopalla caelamenicutis* were obtained from *magnus* Biozone mudstones at Ddole Farm [SN 9761 6747] (Appendix 2biii). They are referable either to the *M. fisherii* or to the *O. explanktonica* Biozone. Recycled acritarchs within the mudstones from both localities include *Acanthodiacrodium* spp. and *Vulcanisphaera* sp. from the early Ordovician and *Petrinosphaeridium vanofarrarium*, *Stellechinatum* cf. *brachyscolum* and *Villosacapsula irrorata*? from the mid- to late Ordovician.

Details

Mottled Mudstone Member

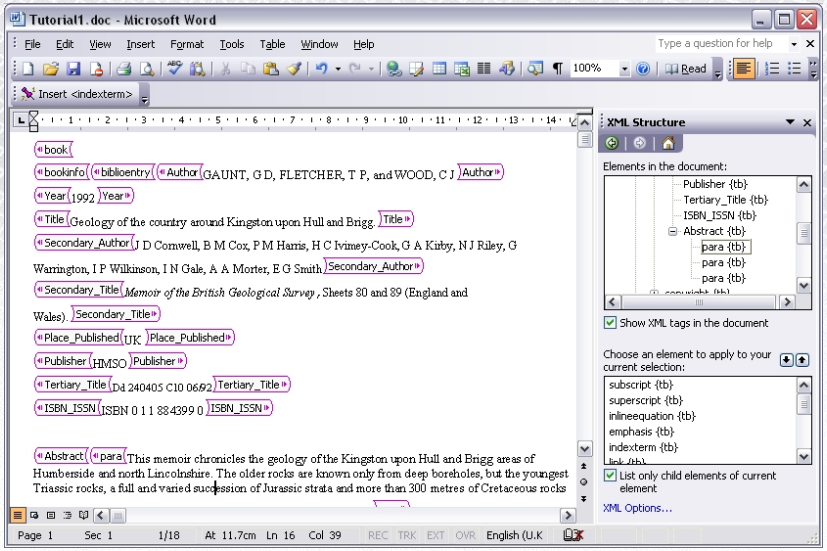
On the north-western limb of the Rhiwnant Anticline part of the Mottled Mudstone is exposed in Ffos y Rhest east of Llyn Carw (Figure 21b). Scattered sections in the member are common around the closure and on the south-eastern limb of the Rhiwnant Anticline. On Craig Cwm clyd [SN 8850 6250 to SN 8890 6248], on the north side of the Claerwen valley, the sharp basal contact with the Yr Allt Formation and the gradational contact with the overlying Dyffryn Flags are well exposed. Rare turbidite sandstone/mudstone couplets (types Di and Cii) with thin basal sandstones up to 2 cm thick are present in the top 1.5 m. A kilometre to the south-east on Craig y Bwlch [SN 8991 6205] (Figure 18a, locality 1), *Normalograptus* spp. were obtained 60 cm above the base of the member in an incomplete section (Table 12). On the south-eastern limb of the Rhiwnant Anticline, the member is about 10 m thick in Nant Paradwys [SN 8917 6097] (Figure 21c) and 6 m thick in Nant yr Ych [SN 8779 5993]. Both sections exhibit a gradational passage up into Dyffryn Flags.

Sections through the Mottled Mudstone Member on the north-western limb of the Tywi Anticline, where it is overlain by Cerrig Gwynion Grits, include a stream section [SN 9533 6450] north of Csom and Cerrig Gwynion quarry [SN 9710 6559] (Figure 20; Table 9; Plate 12a), both south-west of Rhayader. At the latter locality the following section is seen:

	Thickness m
Caban Conglomerate Formation, Cerrig Gwynion Grits facies	
Sandstone, medium-grained with sharp base	seen
Cwmere Formation, Mottled Mudstone Member	
Hemipelagite, pale and medium grey, colour banded, diffusely laminated and burrowed. Scattered mudstone turbidites with siltstone laminae up to 5 mm thick (type E) in upper 1.3 m. Turbidite sandstone/mudstone couplet (type Di) with basal 2 cm sandstone defines base	2.0
Hemipelagite as above	1.9
Upper leaf of *persculptus* Band: hemipelagite, dark grey, laminated with abundant irregular pyrite nodules up to 1 cm in diameter. Abundant *persculptus* Biozone graptolites are preserved in pyrite (Table 9)	0.08
Hemipelagite, pale and medium grey, colour-banded, burrowed and diffusely laminated	0.27
Lower leaf of *persculptus* Band: hemipelagite as in upper leaf	0.17
Hemipelagite, pale and medium grey, colour-banded, diffusely burrowed and laminated. Scattered pyrite nodules in upper part. Abundant interbedded turbidite mudstones with siltstone laminae up to 1 mm thick (type E) in basal 35 cm.	

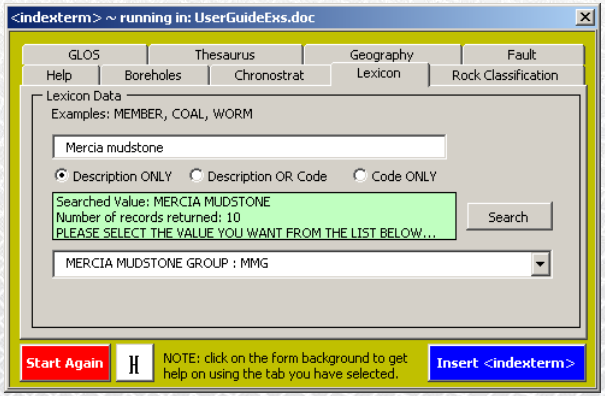
Textbase 2000

In the early 2000s, BGS ran a small project converting 30 regional reports including memoirs into XML using Word



Textbase 2000

This earlier project added not just syntactic information about the layout of the document, but also some semantic information, labelling words and phrases and linking them to BGS controlled vocabularies

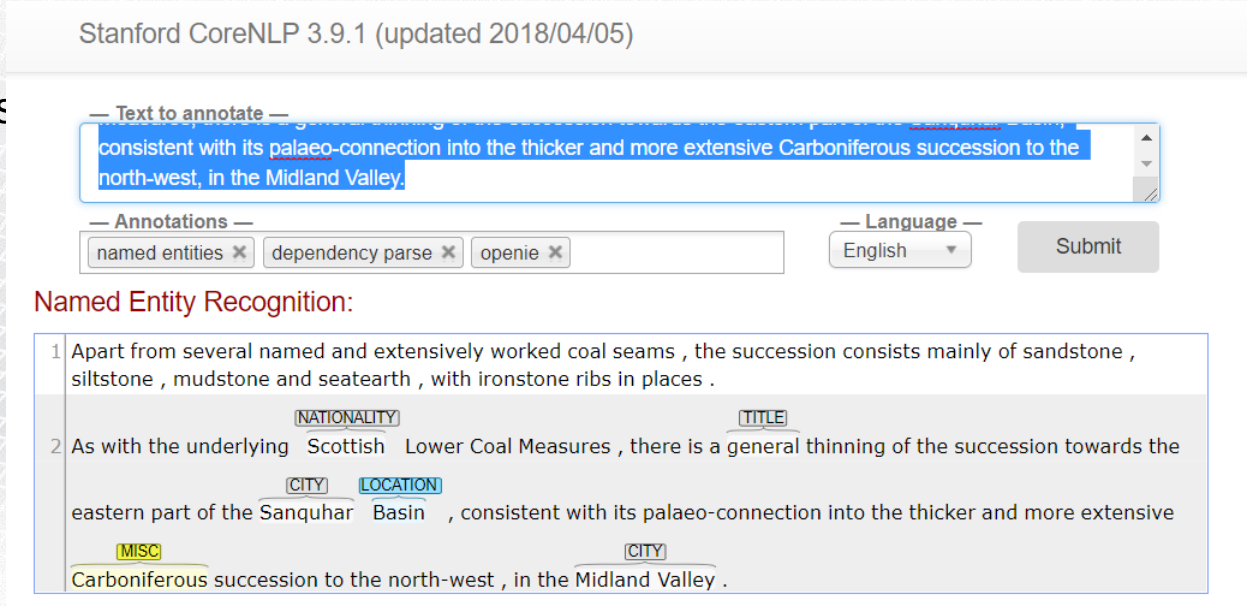


Textbase 2000

- From these manually-created annotations we extracted 2500 sentences
- Every term found in the BGS Lexicon of Named Rock units is labelled LEXICON
- Every reference to a geochronological or chronostratigraphic term is labelled CHRONOSTRAT
- We used this data to train a custom Named Entity Recognition model

Stanford CoreNLP NER

- Bas



Named Entity Recognition (NER) with default CoreNLP

CoreNLP training data - sample

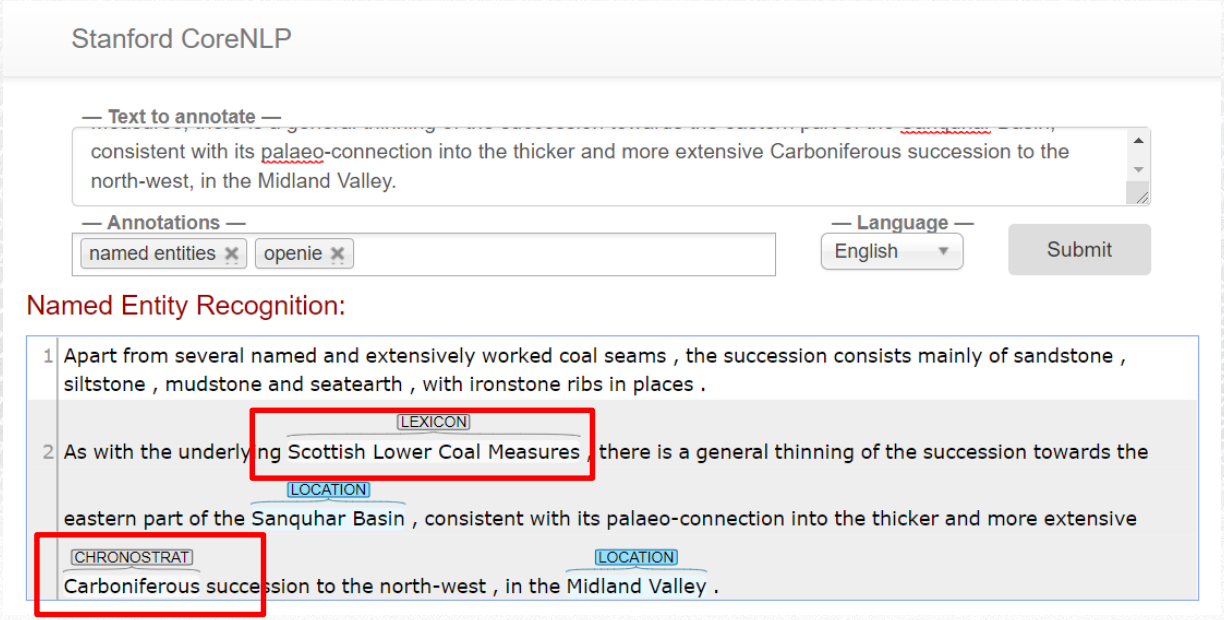
```
The     O
highly  O
irregular       O
middle  O
zone    O
is      O
underlain       O
by      O
Permian CHRONOSTRAT
and     O
Triassic        CHRONOSTRAT
rocks   O
;       O
the     O
Permian CHRONOSTRAT
strata  O
include O
Upper   CHRONOSTRAT
Permian CHRONOSTRAT
Zechstein       LEXICON
sedimentary     O
rocks   O
that    O
locally O
crop    O
out     O
in      O
the     O
study   O
area    O
.       O
```

Train NER model with CRFClassifier

```
java -cp stanford-ner.jar
edu.stanford.nlp.ie.crf.CRFClassifier -prop
bgs.3class.geo.prop
```

https://nlp.stanford.edu/software/crf-faq.shtml

Custom NER for CoreNLP

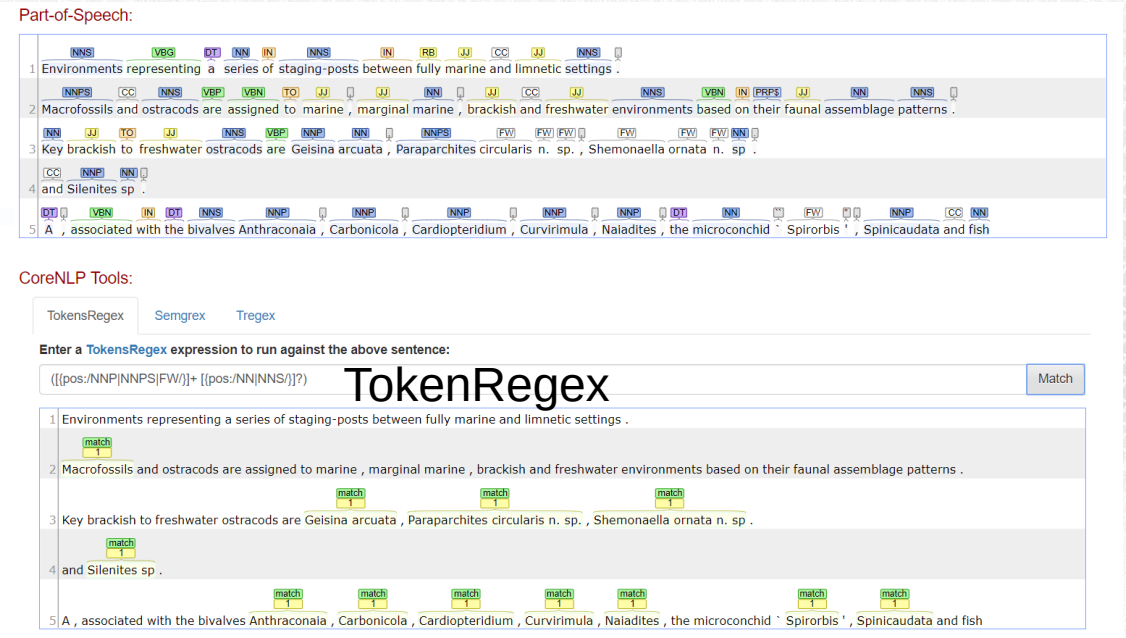


Named entities recognised with custom model added

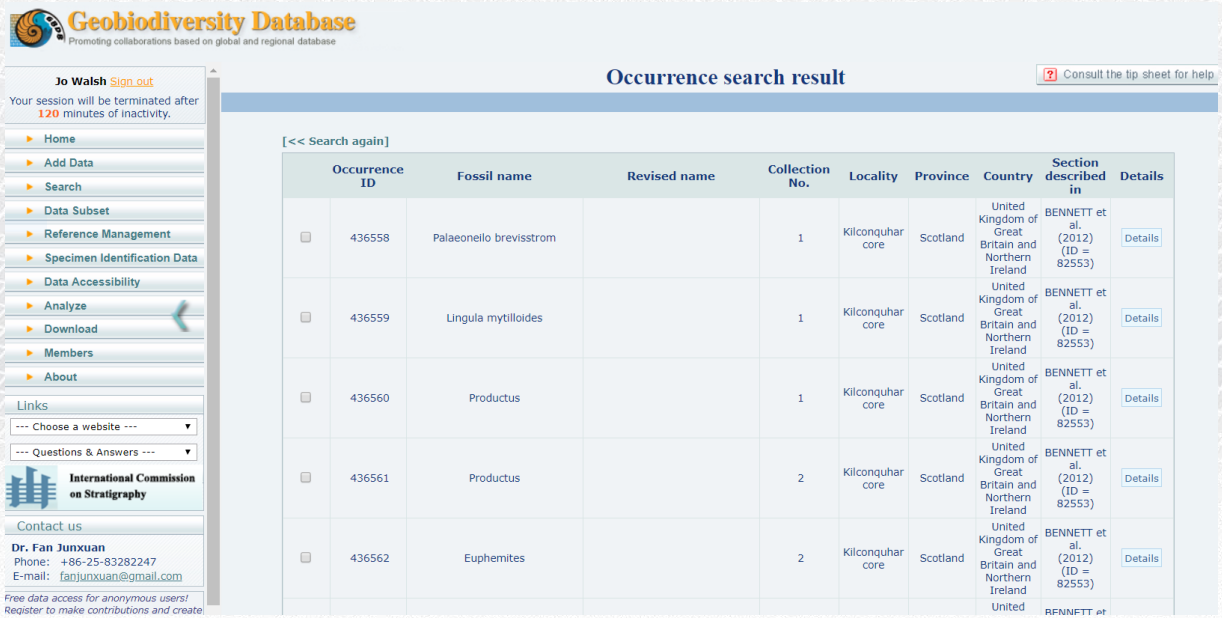
Adding additional vocabularies

- Given enough labelled examples, train a Named Entity Recognition model with a different vocabulary
- We had a few BIOZONEs in our training data but too few examples for the classifier to learn from

Rule-based approach using parts of speech



Or use manually extracted entities to recreate labelled sentences



Occurrence ID	Fossil name	Revised name	Collection No.	Locality	Province	Country	Section described in	Details
436558	Palaeoneilo brevisstrom		1	Kilconquhar core	Scotland	United Kingdom of Great Britain and Northern Ireland	BENNETT et al. (2012) (ID = 82553)	Details
436559	Lingula mytilloides		1	Kilconquhar core	Scotland	United Kingdom of Great Britain and Northern Ireland	BENNETT et al. (2012) (ID = 82553)	Details
436560	Productus		1	Kilconquhar core	Scotland	United Kingdom of Great Britain and Northern Ireland	BENNETT et al. (2012) (ID = 82553)	Details
436561	Productus		2	Kilconquhar core	Scotland	United Kingdom of Great Britain and Northern Ireland	BENNETT et al. (2012) (ID = 82553)	Details
436562	Euphemites		2	Kilconquhar core	Scotland	United Kingdom of Great Britain and Northern Ireland	BENNETT et al. (2012) (ID = 82553)	Details

Alternatives to CoreNLP

- spacy.io is a python-based toolkit for Natural Language Processing
- Nicer for integration with other python tools for machine learning and scientific computing
- We tried re-training spacy's NER with BGS data but couldn't get it to the same result quality
- USGS have used spacy to recognise dams and rivers in their dam removal project

Alternatives to CoreNLP

- Edinburgh Geoparser from Language Technology Group, School of Informatics, University of Edinburgh
- Uses a rule-based approach rather than a statistical model
- We used its “georesolution” phase to find coordinates for locations

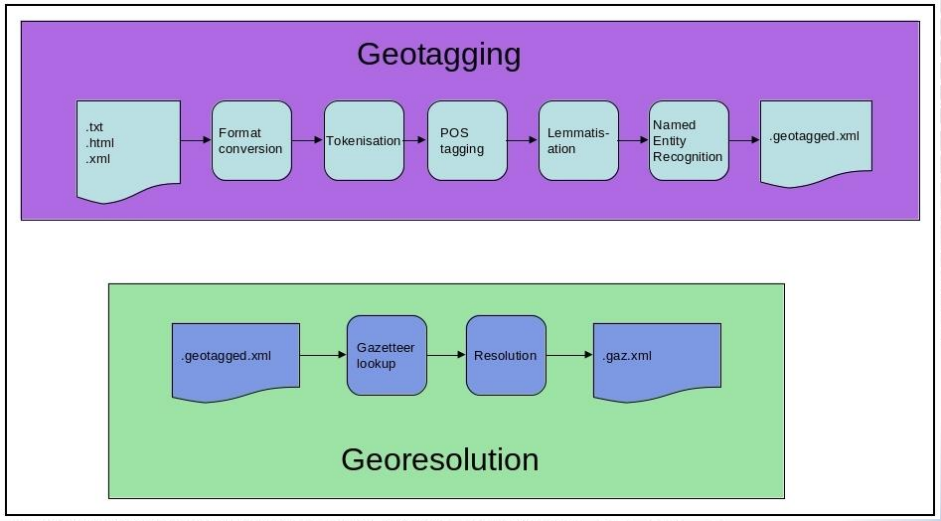


Image captions derived from sheet memoirs

Uploaded on:
14/02/2009 05:13:41

Type:
Digital Asset

File Size:
428.79 KB

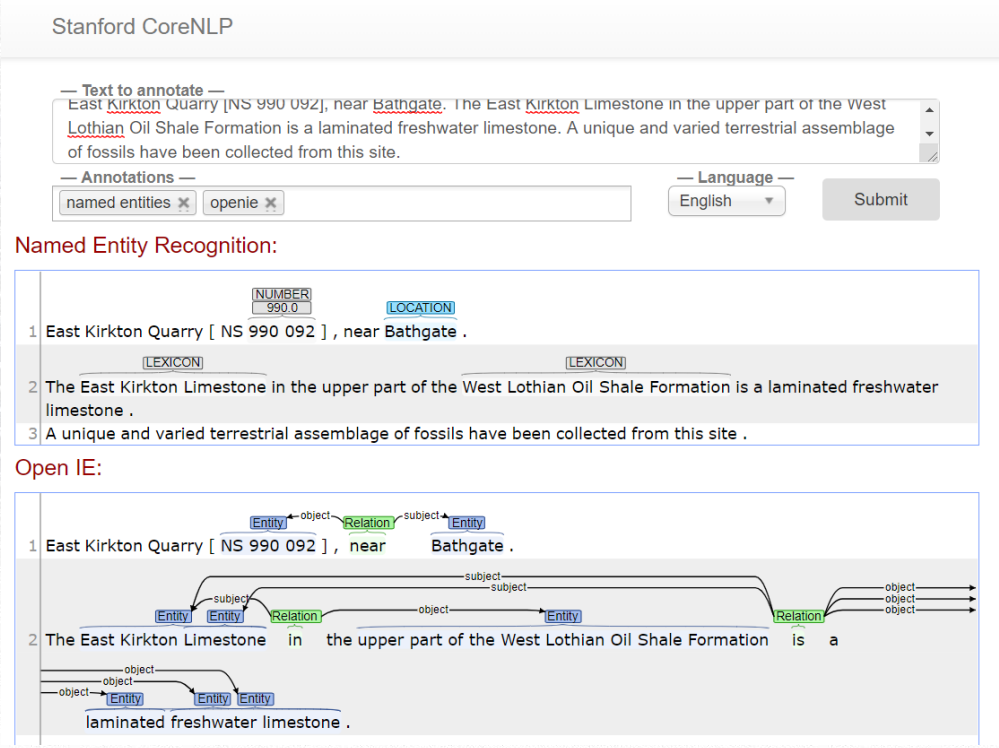
Dimensions:
1001 x 791 pixels

381 views

5 downloads

P number:	P002882
Old photograph number:	D05094
Caption:	East Kirkton Quarry, near Bathgate, West Lothian.
Description:	East Kirkton Quarry, near Bathgate, West Lothian. The East Kirkton Limestone in the upper part of the West Lothian Oil Shale Formation is a laminated freshwater limestone. A unique and varied terrestrial assemblage of fossils has been collected from this site, including Westlothiana lizziae, a vertebrate fossil intermediate between amphibians and reptiles.

Entities found with CoreNLP



Links to BGS vocabularies


```
{
  "_id": "B01613-001-p001-Geochronology:Division-0",
  "term": {
    "term": "Ordovician",
    "vocabulary": "Geochronology:Division",
    "url": "http://data.bgs.ac.uk/id/Geochronology/Division/O",
    "code": "O"
  },
  "text": {
    "match": "Ordovician",
    "offset": [
      38,
      48
    ],
    "snippet": "b'The district is mainly underlain by Ordovician and Silurian mudstones and sandstones. "
  },
  "doc": {
    "id": "B01613",
    "section": [
      "B01613",
      "001",
      "p001"
    ]
  }
}
```


JSON documents stored in and queried from a MongoDB document store

Lexicon of Named Rock Units

The BGS Lexicon of Named Rock Units — Result Details

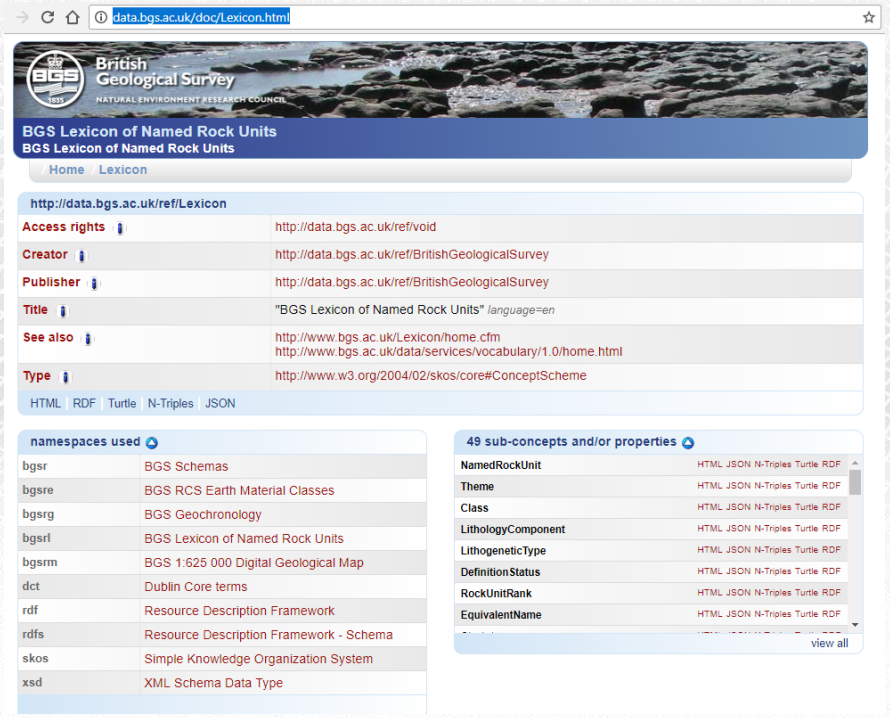
Bathgate Hills Volcanic Formation

Computer Code:	BHV	Preferred Map Code:	BHV
Status Code:	Full		
Age range:	Asbian Substage (CR) — Arnsbergian Substage (CG)		
Lithological Description:	Pyroclastic rocks and related basaltic lavas, mostly alkali olivine-basalts (basanite to hawaiite) and volcaniclastic rocks.		
Definition of Lower Boundary:	First occurrence of volcanic rocks above the Two Foot Coal in the West Lothian Oil-Shale Formation.		
Definition of Upper Boundary:	Highest occurrence of volcanic rocks, below the Roman Cement Limestone, Passage Formation.		
Thickness:	To 1000m		
Geographical Limits:	Bathgate Hills, west of a line from NT0066 to NT0980, extending westwards underground to a point west of the Rashiehill Borehole (NS839730).		
Parent Unit:	Bathgate Group (BATH)		
Previous Name(s):	Bathgate Lavas (-3215) Bathgate Volcanic Formation (-58)		
Alternative Name(s):	*none recorded or not applicable*		
Stratotypes:			
Type Area	Natural sections in the Bathgate Hills.		
Reference Section	Rashiehill Borehole (NS87SW/22), (Anderson, 1963), 2600ft (792.48m) to 3645ft (1110.99m).		
Reference Section	Couston Borehole (NS97SW/99) 110 fathoms (201.17m) to 298 fathoms (544.98m).		

Web-based

Human-readable

Linked Data Lexicon



Web-based

Machine-readable

“Pompidou Centre” data

A phrase coined by Peter Burnhill of EDINA, University of Edinburgh

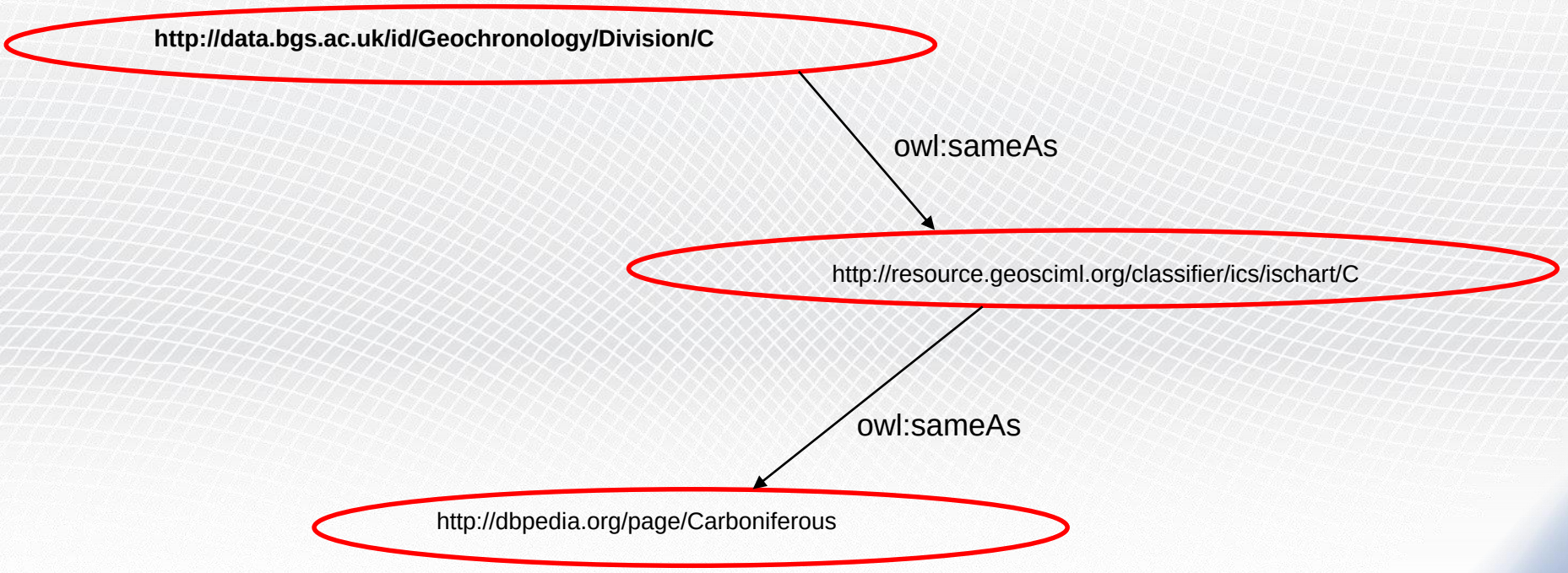
Data infrastructure is exposed on outside of the structure

Photo credit: copyright Pascal Poggi

“Use URIs as names for things”

- Internal code: *CAL*
- External code: *http://data.bgs.ac.uk/id/Lexicon/RockUnit/CAL*

Link to other things on the semantic web

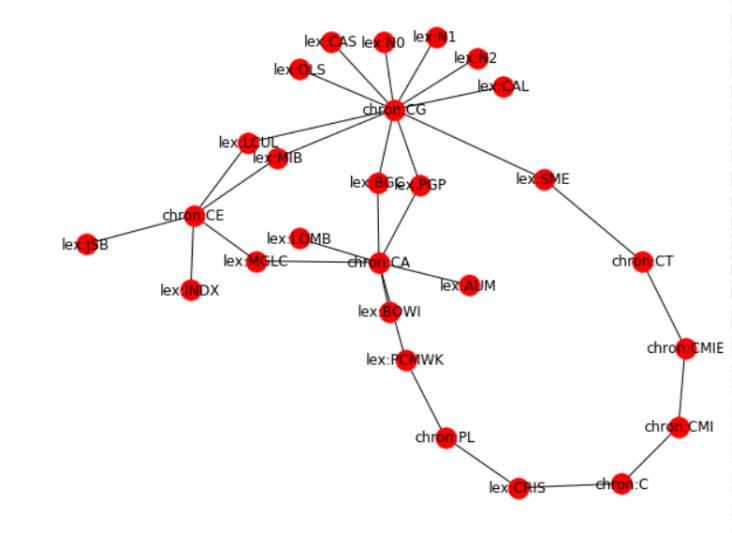


Triples all the way down

```
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2000/01/rdf-schema#label>    "\u77F3\u70AD\u7D00"@ja .
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/ontology/abstract>  "\u77F3\u70AD\u7D00\uFF08\u305B\u304D\u305F\u3093\u304D\u3001Carboniferous period\uFF09\u
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2000/01/rdf-schema#label>    "\u77F3\u70AD\u7EAA"@zh .
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/ontology/abstract>  "Karbon \n*  w sensie geochronologicznym \u2013 pi\u0105ty okres ery paleozoicznej. Trwa\
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2002/07/owl#sameAs>  <http://de.dbpedia.org/resource/Karbon> .
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/ontology/abstract>  "\u0627\u0644\u0639\u0635\u0631 \u0627\u0644\u0641\u062D\u0645\u064A \u0623\u0648 \u0627\
<http://dbpedia.org/resource/Carboniferous> <http://purl.org/dc/terms/subject>  <http://dbpedia.org/resource/Category:Carboniferous> .
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>   <http://www.w3.org/2002/07/owl#Thing> .
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/ontology/thumbnail> <http://commons.wikimedia.org/wiki/Special:FilePath/340Marect.jpg?width=300> .
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/ontology/abstract>  "Das Karbon ist in der Erdgeschichte das f\u00FCnfte chronostratigraphische System bzw. d
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2000/01/rdf-schema#comment>  "Le Carbonif\u00E8re est un syst\u00E8me g\u00E9ologique du Pal\u00E9ozo\u00EFque
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/ontology/abstract>  "Le Carbonif\u00E8re est un syst\u00E8me g\u00E9ologique du Pal\u00E9ozo\u00EFque. Cette
<http://dbpedia.org/resource/Carboniferous> <http://xmlns.com/foaf/0.1/isPrimaryTopicOf>    <http://en.wikipedia.org/wiki/Carboniferous> .
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2000/01/rdf-schema#comment>  "\u77F3\u70AD\u7D00\uFF08\u305B\u304D\u305F\u3093\u304D\u3001Carboniferous period
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/ontology/wikiPageExternalLink>  <http://jeb.biologists.org/cgi/content/abstract/201/8/1043> .
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2000/01/rdf-schema#label>    "Carbonif\u00E8re"@fr .
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/property/temp>  "14"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2000/01/rdf-schema#comment>  "El Carbon\u00EDfero es una divisi\u00F3n de la escala temporal geol\u00F3gica. E
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/ontology/wikiPageExternalLink>  <http://www.foraminifera.eu/querydb.php?&period=Carboniferous&aktion=suche> .
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2002/07/owl#sameAs>  <http://www.wikidata.org/entity/Q133738> .
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/ontology/abstract>  "Na escala de tempo geol\u00F3gico, o Carbon\u00EDfero ou Carb\u00F3nico \u00E9 o per\u00E
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2000/01/rdf-schema#comment>  "Das Karbon ist in der Erdgeschichte das f\u00FCnfte chronostratigraphische Syste
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/ontology/abstract>  "El Carbon\u00EDfero es una divisi\u00F3n de la escala temporal geol\u00F3gica. Es el per
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2000/01/rdf-schema#comment>  "Karbon \n*  w sensie geochronologicznym \u2013 pi\u0105ty okres ery paleozoiczne
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2000/01/rdf-schema#comment>  "Nella Scala dei tempi geologici, il Carbonifero \u00E8 il quinto dei sei periodi
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2002/07/owl#sameAs>  <http://ja.dbpedia.org/resource/\u77F3\u70AD\u7D00> .
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2000/01/rdf-schema#label>    "\u041A\u0430\u043C\u0435\u043D\u043D\u043E\u0443\u0433\u043E\u043B\u044C\u043D\u
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2002/07/owl#sameAs>  <http://fr.dbpedia.org/resource/Carbonif\u00E8re> .
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/ontology/wikiPageID>    "5401"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2002/07/owl#sameAs>  <http://cs.dbpedia.org/resource/Karbon> .
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2002/07/owl#sameAs>  <http://d-nb.info/gnd/4126367-4> .
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/ns/prov#wasDerivedFrom>  <http://en.wikipedia.org/wiki/Carboniferous?oldid=741782135> .
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/property/o> "32.5"^^<http://www.w3.org/2001/XMLSchema#double> .
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2000/01/rdf-schema#label>    "Carbon\u00EDfero"@es .
<http://dbpedia.org/resource/Carboniferous> <http://dbpedia.org/ontology/wikiPageExternalLink>  <http://www.stratigraphy.org/gssp.htm> .
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2000/01/rdf-schema#label>    "Karbon"@de .
<http://dbpedia.org/resource/Carboniferous> <http://www.w3.org/2002/07/owl#sameAs>  <http://eu.dbpedia.org/resource/Karbonifero> .
```

Knowledge Graph

Fashionable term for linked data extracted from text

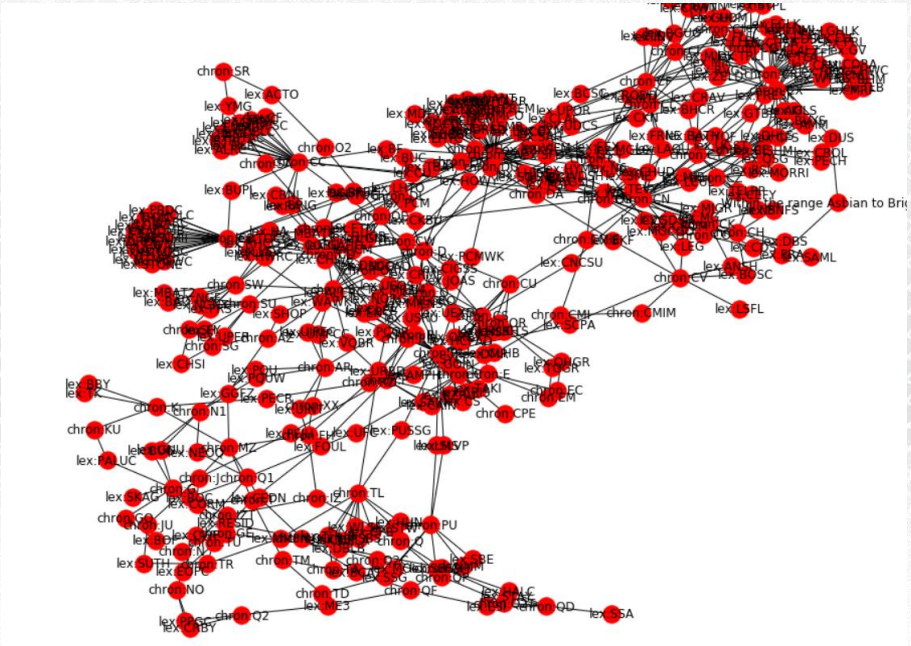


Named entities extracted from a BGS Palaeontology report

Sum of all shortest paths through a graph of connections between min/max ages of rock types and ICS hierarchy of geochronological eras

rdflib, networkx, matplotlib

Is this knowledge?



Same process;

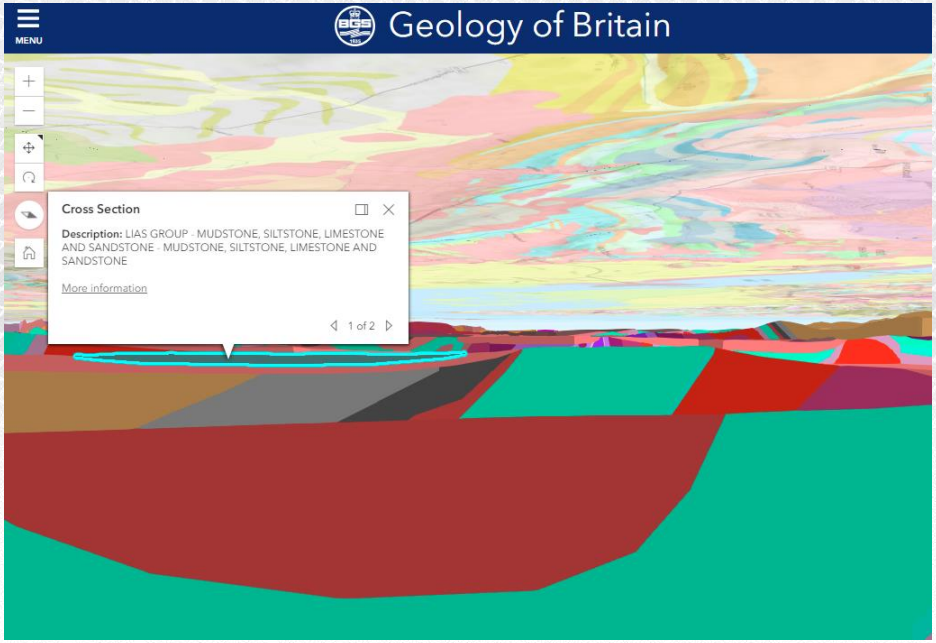
Graph inferred from several hundred page map sheet memoir

Links to other resources

- Other kinds of BGS data resources are already linked to the Lexicon and Geochronology
- Thin section imagery, borehole scans
- Deposited datasets have some manually-added tags
- We can automatically add more to new incoming data
- Also a lot of un-mined text content in the Lexicon itself

Other linked resources

BGS 3D geological models are linked to the rock units



BGS memoirs – entities visualised

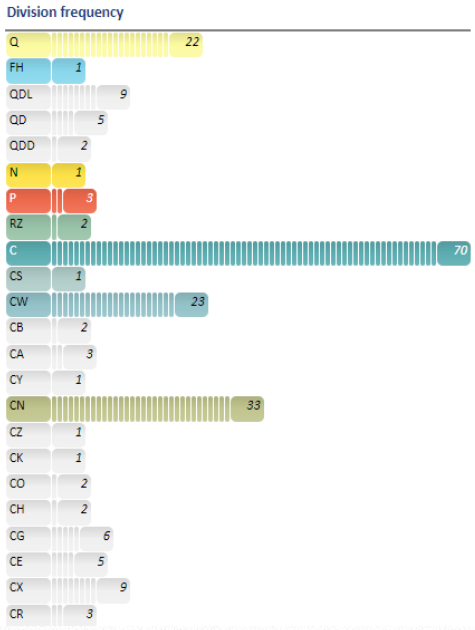
- We ran the Named Entity Recognition over 60 BGS memoirs in cleaned-up XML format
- We extracted references to Chronostratigraphic terms, to Lexicon of Named Rock Units, and locations
- We also ran a selection of PDF scans of palaeontology reports
- Even with lower-quality OCR the results are promising

Geochronology extracted from text

32 Geochronological Divisions found in *B01881*

Divisions matched sort by Age (young-old)

Code	Name	Min Age (MYBP)	Max Age (MYBP)	Matches
Q	Quaternary Period	0 ±0	2.588 ±0	22
FH	Phanerozoic Eon	0 ±0	541 ±1	1
QDL	Loch Lomond Stadial (Younger Dryas)	0.0118 ±0	0.01265 ±0	9
QD	Devensian Stage	0.0118 ±0	0.116 ±0	5
QDD	Dimlington Stadial	0.0147 ±0	0.026 approx. ± 0	2
N	Neogene Period	2.588 ±0	23.03 ±0	1
P	Permian Period	252.2 ±0.5	298.9 ±0.2	3
RZ	Palaeozoic Era	252.2 ±0.5	541 ±1	2
C	Carboniferous Period	298.9 ±0.2	358.9 ±0.4	70
CS	Stephanian Stage	302 approx. ± 0	308 approx. ± 0	1
CW	Westphalian Stage	308 approx. ± 0	319 approx. ± 0	23
CB	Duckmantian Substage	315.2 ±0.2	318 approx. ± 0	2
CA	Langsettian Substage	318 approx. ± 0	319 approx. ± 0	3
CY	Yeadonian Substage	319 approx. ± 0	320 approx. ± 0	1
CN	Namurian Stage	319 approx. ± 0	329 approx. ± 0	33
CZ	Marsdenian Substage	320 approx. ± 0	321.5 approx. ± 0	1
CK	Kinderscoutian Substage	321.5 approx. ± 0	322 approx. ± 0	1
CO	Alportian Substage	322 approx. ± 0	323 approx. ± 0	2
CH	Chokierian Substage	323 approx. ± 0	324 approx. ± 0	2
CG	Arnsbergian Substage	324 approx. ± 0	328 approx. ± 0	6
CE	Pendleian Substage	328 approx. ± 0	329 approx. ± 0	5
CX	Brigantian Substage	329 approx. ± 0	330.9 approx. ± 0	9
CR	Asbian Substage	330.9 approx. ± 0	337 approx. ± 0	3
CI	Chadian Substage	344.5 approx. ± 0	346.7 ±0.4	1



Mentions of rock unit types

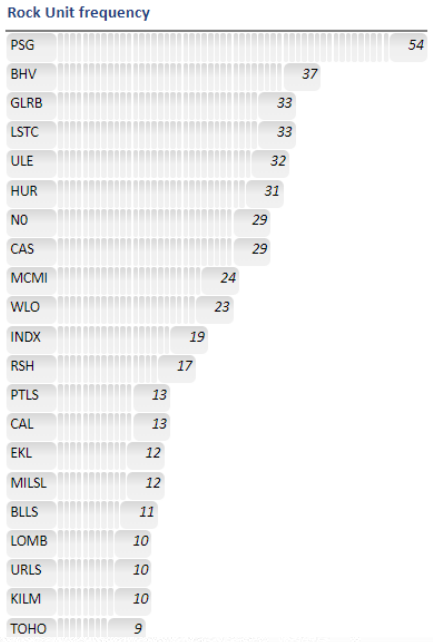
BGS TextMining :: Document Report — Home | Document Report | Vocabulary Report | Location Report

Document | Geochronology | Lexicon Rock Units | Locations | Correlations

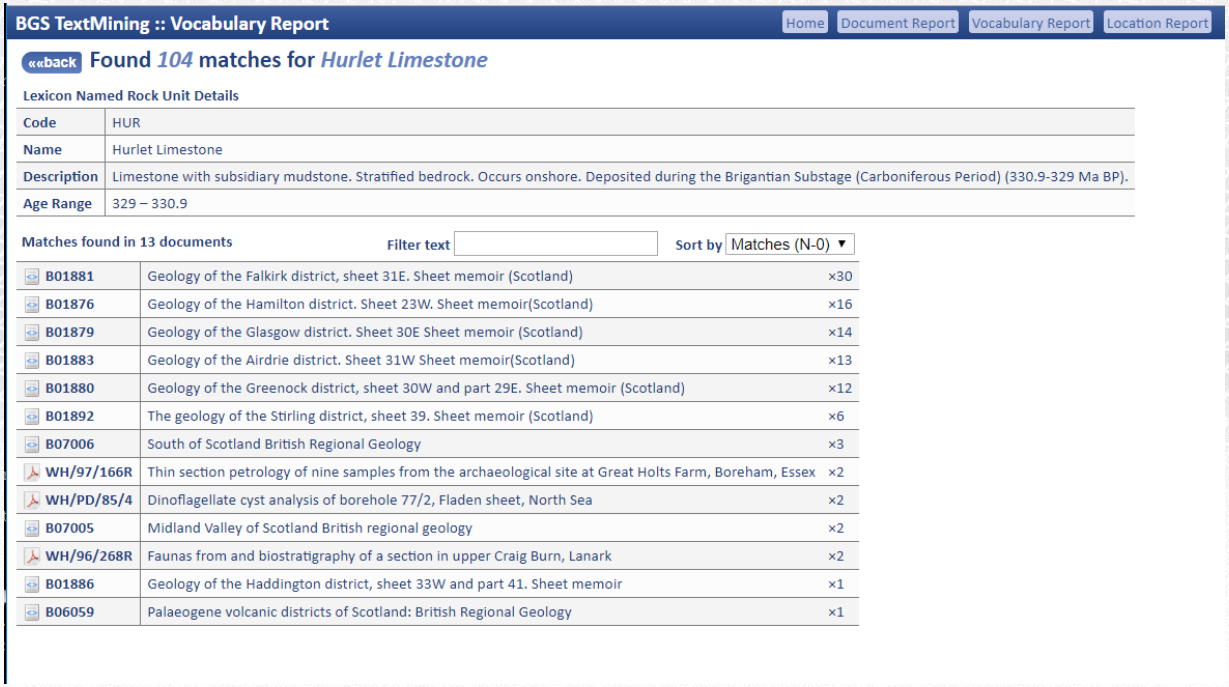
138 Lexicon Rock Units found in *B01881*

Lexicon matches sort by Match Count (9-1)

Code	Name	Min Age (MYBP)	Max Age (MYBP)	Matches
PSG	Passage Formation (Wenlock Series)	427.4	433.4	54
BHV	Bathgate Hills Volcanic Formation	324	337	37
GLRB	Lower Limestone	145	152.1	33
LSTC	Limestone Coal	329	330.9	33
ULE	Upper Limestone	145	152.1	32
HUR	Hurlet Limestone	329	330.9	31
N0	No. 0 Marine Band	324	328	29
CAS	Castlecary Limestone	324	328	29
MCMI	Middle Coal Measures (Dungannon, Northern Ireland)	318	319	24
WLO	West Lothian Oil-Shale Formation	329	337	23
INDX	Index Limestone [Duplicate Code: Use ILS]	328	329	19
RSH	Raeburn Shale	329	346.7	17
PTLS	Petershill Limestone	329	330.9	13
CAL	Calmy Limestone	324	328	13
EKL	East Kirkton Limestone	329	346.7	12
MILSL	Mill Coal (South Lancashire)	315.2	318	12
BLLS	Blackhall Limestone	329	330.9	11
LOMB	Lowstone Marine Band	318	319	10
URLS	Under Limestone	329	330.9	10
KILM	Kiltongue Musselband Coal	318	319	10
TOHO	Top Hosie Limestone	328	330.9	9



Rock unit types across documents



BGS TextMining :: Vocabulary Report

Home | Document Report | Vocabulary Report | Location Report

«back Found *104* matches for *Hurlet Limestone*

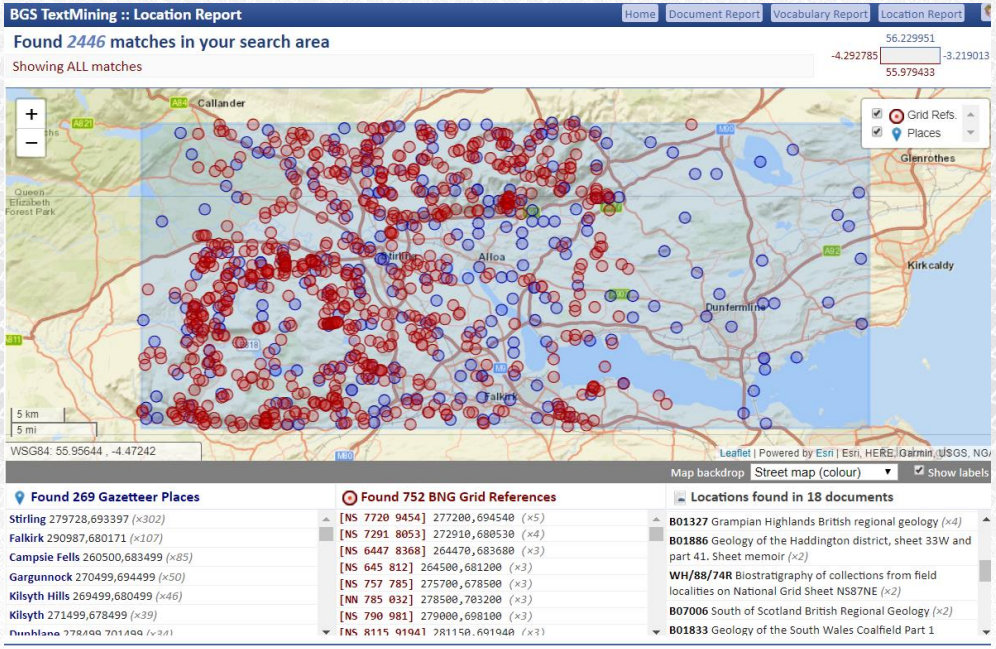
Lexicon Named Rock Unit Details

Code	HUR
Name	Hurlet Limestone
Description	Limestone with subsidiary mudstone. Stratified bedrock. Occurs onshore. Deposited during the Brigantian Substage (Carboniferous Period) (330.9-329 Ma BP).
Age Range	329 – 330.9

Matches found in 13 documents Filter text Sort by Matches (N-0)

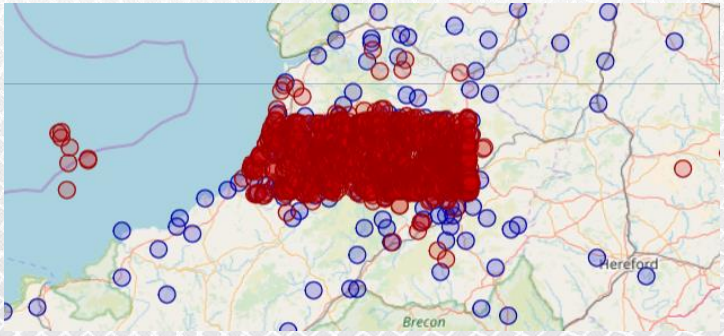
B01881	Geology of the Falkirk district, sheet 31E. Sheet memoir (Scotland)	×30
B01876	Geology of the Hamilton district. Sheet 23W. Sheet memoir(Scotland)	×16
B01879	Geology of the Glasgow district. Sheet 30E Sheet memoir (Scotland)	×14
B01883	Geology of the Airdrie district. Sheet 31W Sheet memoir(Scotland)	×13
B01880	Geology of the Greenock district, sheet 30W and part 29E. Sheet memoir (Scotland)	×12
B01892	The geology of the Stirling district, sheet 39. Sheet memoir (Scotland)	×6
B07006	South of Scotland British Regional Geology	×3
WH/97/166R	Thin section petrology of nine samples from the archaeological site at Great Holts Farm, Boreham, Essex	×2
WH/PD/85/4	Dinoflagellate cyst analysis of borehole 77/2, Fladen sheet, North Sea	×2
B07005	Midland Valley of Scotland British regional geology	×2
WH/96/268R	Faunas from and biostratigraphy of a section in upper Craig Burn, Lanark	×2
B01886	Geology of the Haddington district, sheet 33W and part 41. Sheet memoir	×1
B06059	Palaeogene volcanic districts of Scotland: British Regional Geology	×1

Memoir overview

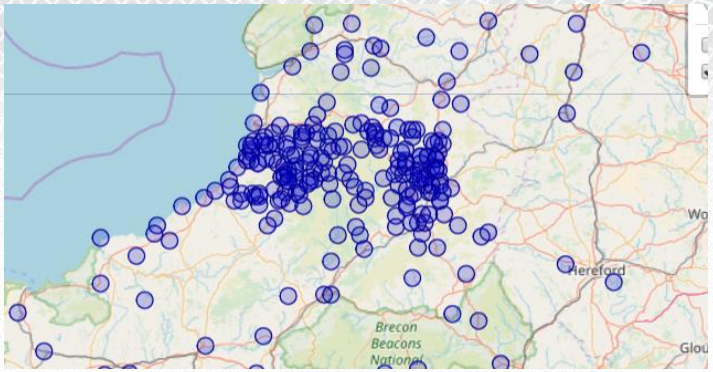


Explore references to places

Coordinates found in text:



Place-names recognised in text:



Correlations between entities

Lexicon Rock Unit :: Gazetteer Place

Rock Unit	Distances	Place
Top Hosie Limestone *Lexicon NamedRockUnit TOHO*	100 100 100 100 100 100 100	**Scotland** *270946,768055*
Johnstone Shell Bed *Lexicon NamedRockUnit JSB*	76 76 76 76 76 76 76	**Scotland** *270946,768055*
Top Hosie Limestone *Lexicon NamedRockUnit TOHO*	78	**Central Coalfield** *437222,317148*

Geochronological Division :: Gazetteer Place

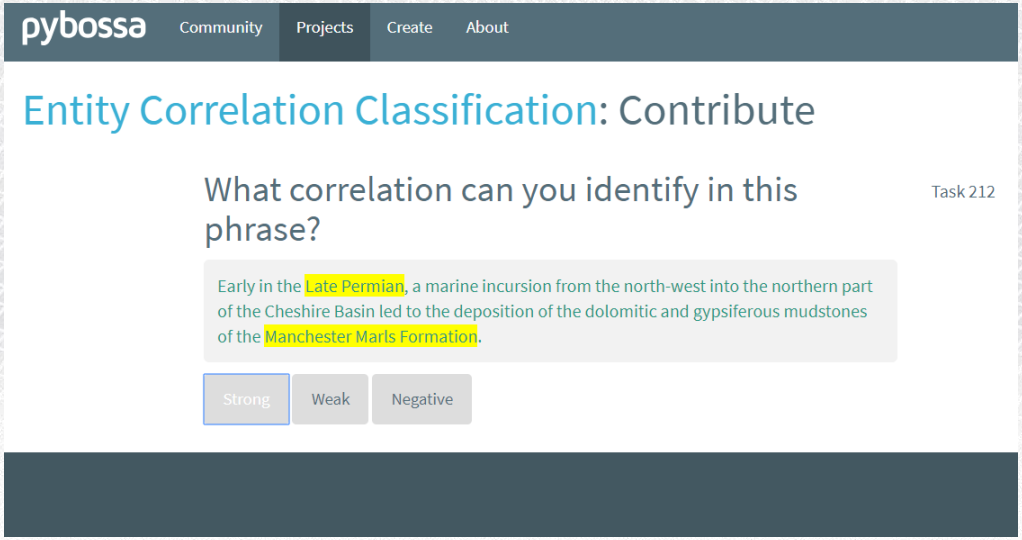
Division	Distances	Place
Carboniferous *Geochronology Division C*	4	**South Africa** *367891,445613*
Namurian *Geochronology Division CN*	94	**South Africa** *367891,445613*
Namurian *Geochronology Division CN*	18	**Central Scotland** *309846,725208*

Next annotation task

- We can see co-occurrences of terms but want to know when they are most meaningful
- We want a model classify positive and negative correlations between named entities in sentences
- Natural History Museum in Oslo have done a similar project with excellent results
- Resulting in the ability to infer fossil ages from correlations of taxa with chronostratigraphic names

pybossa task to classify relations

pybossa.org - like an open source GalaxyZoo



We will need several annotators to label several thousand sentences for a decent model

Training a TextCategorizer

- NHM in Oslo built with own classification model
- Feeding skipgram word embeddings from fasttext into an LSTM created with scikit-learn
- We are planning to start with the TextCategorizer package in spacy.io and work from there

Extracting tables and figures

- GeoDeepDive is a project at University of Madison, Wisconsin, US
- GeoDeepDive searches mainly journal publications but holds some geological survey archives
- With Stanford CoreNLP with Tesseract OCR they scan and analyse millions of short publications
- Blackstack is a new component of the GeoDeepDive project which takes a machine learning approach to identifying figures, maps and tables
- It's a work in progress!
- https://geodeepdive.org
- https://github.com/UW-Deepdive-Infrastructure/blackstack

Conclusions

"You have to have a human in the loop; what matters is where you put the human in the loop"

- *Claire Grover, Language Technology Group, University of Edinburgh*

- NER tools are accessible without NLP expertise
- Semantic web is reviving with a simpler approach
- Potential to generate scientific discoveries…
- … but are we often discovering the obvious?

Geosemantics at BGS Informatics
Project leader: Rachel Heaven: reh@bgs.ac.uk
PhD student: Ike Nkisi-Orji: i.o.nkisi-orji@rgu.ac.uk