EC Project 610829

A Decarbonisation Platform for Citizen Empowerment and Translating Collective Awareness into Behavioural Change

D2.2.1: Text Analytics Tools for Environmental Information Extraction

04 October 2014

Version: 1.0

Version history

Version	Date	Author	Comments
0.1	15/09/2014	Diana Maynard	Initial version.
0.2	23/09/2014	Arno Scharl	Section 4
0.3	03/10/2014	Diana Maynard	changes following OU comments

Peer reviewed by: Harith Alani (OU)

Dissemination Level: PU – Public

This document is part of the DecarboNet research project, which receives funding from the European Union's 7th Framework Programme for research, technology development and demonstration (Grant Agreement No 610829; ICT-2013.5.5 CAPS Collective Awareness Platforms for Sustainability and Social Innovation).

Executive Summary

This deliverable provides a report to accompany the three web services for environmental information extraction delivered. The web services provide tools to perform entity disambiguation, recognition of environmental terms, and extraction of environmental indicators respectively. Since the services are still in development, and this is only the first version, users are able to just make use of the web service; the final version will be made open source.

The report explains how to use the web services, describes the applications and the underlying natural language processing tools used, and details some initial experiments carried out to evaluate the performance of these tools. Finally, it provides some information about ongoing work and further possible improvements to be made.

Table of Contents

1. Introduction 4
2. ClimaTerm: a web service for term recognition 4
   2.1. Ontologies 4
      2.1.1. GEMET 5
      2.1.2. REEGLE clean energy and climate glossary 5
   2.2. Combining the terms 6
   2.3. Linguistic processing of terms 6
      2.3.1. Terms within hashtags 6
      2.3.2. Named entities 7
      2.3.3. Restrictions on POS tags 7
   2.4. Validation experiments 7
   2.5. Climate corpus 7
   2.6. Energy corpus 10
   2.7. Fracking corpus 11
   2.8. Software availability 11
3. Web service for indicators 12
   3.1. Extracting climate change indicators 12
   3.2. Software availability 14
4. Recognyze: a service for named entity recognition 15
   4.1. Analysing Named Entities 15
   4.2. Software availability 16
5. Summary and further work 17

A. List of Figures 18

B. List of Tables 18

C. List of Abbreviations 19

D. References 20

1. Introduction

This deliverable describes the three web services provided for knowledge extraction in DecarboNet, and gives some more detailed information about the underlying technology of the applications, along with some first experimental results to test their accuracy. The web services enable other project members to access the extraction services so that they can use them within the project for experimentation. They require no technical skills and can therefore be used by partners from any WP.

The work described here is also related to WP1, WP3 and WP4. In D1.3 the planned architecture and APIs for the knowledge extraction web services were described, which have now been implemented here. The data used for the development and experiments was collected using the tools developed in WP3 for filtering the Twitter stream in real time. WP4 uses the annotation service to track environment-related tweets over time, posted by individual users around Earth Hour campaigns (see D6.2.1). This helps to study the evolution of engagement with environment topics before, during, and after Earth Hour.

In the rest of this document, we describe each of the three tools in turn: term recognition, indicator recognition and entity disambiguation, and finish with some plans for their further development.

2. ClimaTerm: a web service for term recognition

This web service aims to annotate documents with terms related to climate change. We have investigated various relevant ontologies available as Linked Open Data and chosen the two which appear to be the most relevant: GEMET and REEGLE, as described below. The web service takes as input a document or set of documents, and outputs those documents as XML files annotated with term and URI information. The underlying application is developed in GATE[1] and contains the following processing stages:

- linguistic pre-processing: tokenisation, sentence splitting, part-of-speech tagging, morphological analysis
- term extraction: matching against known terms, plus some recognition of morphological and synonym variants
- export as XML (inline annotation)

2.1. Ontologies

In this section, we describe the two ontologies we make use of for the term annotation.

[1] http://gate.ac.uk

2.1.1. GEMET

GEMET (GEneral Multilingual Environmental Thesaurus)[2] is the reference vocabulary of the European Environment Agency (EEA) and its Network (Eionet). It was conceived as a "general" thesaurus, aiming to define a common general language, a core of general terminology for the environment, and contains 5208 terms originating from a number of different thesauri. From this, we extracted all the terms along with their label and URI, as in the example entry below:

label=air pollutant

URI=http://www.eionet.europa.eu/gemet/concept/263

2.1.2. REEGLE clean energy and climate glossary

The REEGLE glossary[3] contains 2527 terms related to climate change in RDF format and a SPARQL endpoint. We extracted the URI, prefLabel, and scopeNote information from this ontology, as shown in the example entry below:

prefLabel: crop yield increase

URI: http://reegle.info/glossary/1400

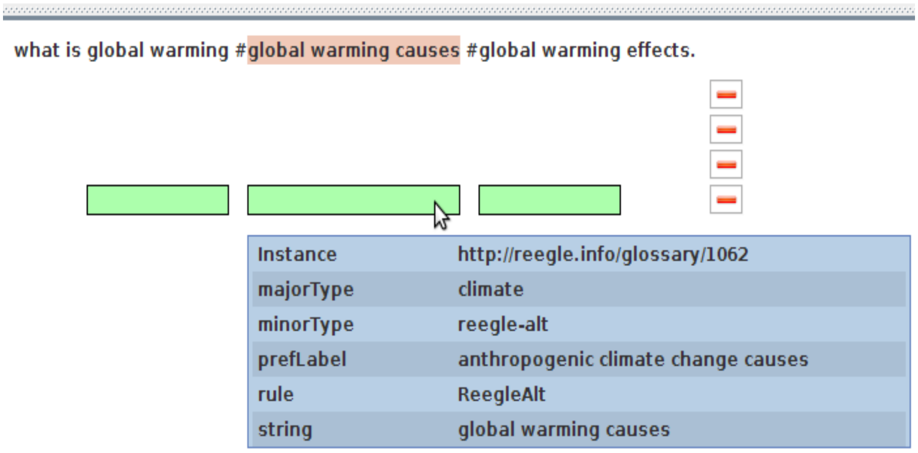
scopeNote: how and where yields might increase due to climate change.

Note that not all the entries have a scopeNote. We also extracted an additional 965 terms listed as "alternative labels" to the main terms. For example, "wind power frequency changers" is the alternative label for the term "windpower inverters". In most cases, these are synonymous or close-to-synonymous terms. Further work will consist of extending the list with morphological and morphosyntactic variants.

Figure 1 shows the term "global warming causes" found in a tweet and annotated with respect to the Reegle glossary. The green boxes show the various terms found in that tweet (e.g. "global warming"). The features depicted in the popup window (in blue) give the URI of the term (instance), the type of term (climate-related, and coming from the altLabel property in the ontology), the preferred label of that term (prefLabel), the rule fired (for debugging purposes) and the original string.

[2] http://www.eionet.europa.eu/gemet/

[3] http://www.reegle.info/glossary



what is global warming #global warming causes #global warming effects.

Instance	http://reegle.info/glossary/1062
majorType	climate
minorType	reegle-alt
prefLabel	anthropogenic climate change causes
rule	ReegleAlt
string	global warming causes

Figure 1: Annotation of a term variant in GATE

2.2. Combining the terms

The two ontologies contain overlapping sets of terms. Given that REEGLE contains mostly more specific terms, we decided to prefer these over GEMET terms. However, we prefer GEMET terms over the alternative terms derived from REEGLE[4]. We also prefer the longest match in any case (so a longer GEMET term would take preference over a shorter REEGLE term). We conducted some experiments to check the validity of the terms annotated in the text, by manually annotating a small set and also by comparing with our generic term extractor tool, TermRaider[5]. Details are given in Section 2.4. First, however, we compared the results with a small evaluation set in order to improve the performance in an iterative fashion, as described in Section 2.3 below.

2.3. Linguistic processing of terms

Our initial application, as described above, achieved excellent precision but only moderate recall when compared with the gold standard set. We found that a small number of missing terms accounted for many cases, for instance "global warming" was not in either ontology. Furthermore, a large number of missing terms were due to hashtags where a multiword term was combined into a single word and was therefore not recognised, for example #palmoil. Other missing terms included morphological variants of multi-word terms. We added some extra terms to the list based on some top-ranked terms found using TermRaider, and added some further pre-processing components to the application, as follows.

2.3.1. Terms within hashtags

First, we added some hashtag pre-processing to re-tokenise hashtags according to their constituent words, using the tool developed in (Maynard et al. 2014). This enables for example the term "palm oil" to be matched against the text "#palmoil", as depicted in the screenshot in Figure 2. Here we can see the span of the original

[4] We could also change the behaviour to return all terms that match, i.e. to return multiple URIs for a term, if desired.

[5] https://gate.ac.uk/projects/arcomem/TermRaider.html

hashtag (in blue, and denoted by the row “Terms#Hashtag”), the terms found within the hashtag (in green, and denoted by the row “Terms#Term”) and the new tokens (in red, and denoted by the row “Terms#Token”). The original hashtag *#palmoilhumanrights* has been correctly tokenised into four words, and then two terms have been found, each containing two words (*palm oil* and *human rights*). Without the retokenisation first, we could not expect to make correct term identification.

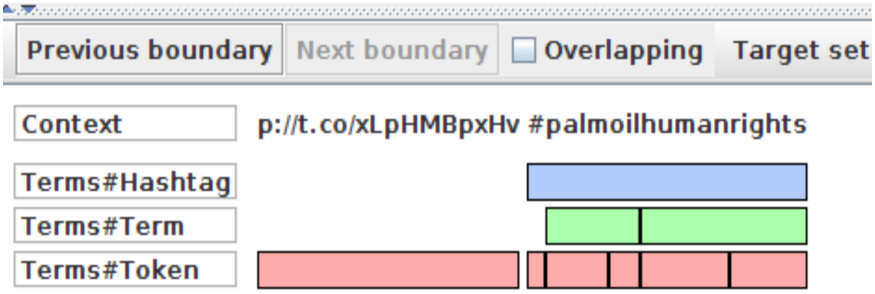


Figure 2: Screenshot of a decomposed hashtag in GATE

2.3.2. Named entities

Some of the terms in GEMET and REEGLE are named entities (manly names of organisations, such as “World Wildlife Fund”). Because these are already recognised and disambiguated by Recognyze (see Section 4) we do not also recognise them here since we consider terms and named entities to be mutually exclusive. We therefore restrict the matching to prevent these being identified here.

2.3.3. Restrictions on POS tags

Finally, we added some restrictions such that terms which are not part of noun phrases should also not be included. For example “global warming causes”, where “causes” is a plural noun, could be a relevant term, but if “causes” is a verb, then it is not part of the term. This does bring some additional issues, however, since the POS tagging is not perfect, and also due to homographs such as “lead”, but initial experiments show it to be a worthwhile tradeoff.

2.4. Validation experiments

In this section we describe some experiments we conducted to check the validity of extracted terms. We used three different corpora, all collected via the MWCC and then manually annotated by a student and verified by one of the developers. These corpora and the experiments are described in the following 3 sections.

2.5. Climate corpus

The first corpus was a set of 455 tweets from MWCC[6] about climate change and palm oil. Table 2 shows the evaluation results for each term set (GEMET only, REEGLE only and the combined set) compared with the gold standard. In all the tables shown here, the overlap column shows the number of annotations that were partially correct, i.e. where the annotation was correct but the span of the annotation was wrong. This occurs where a term contains more or fewer words than the gold

[6] http://www.ecoresearch.net/climate/

standard, e.g. recognising “climate” instead of “climate change”, and is scored (as is usual) by a half weight.

Term Set	Match	Missing	Spurious	Overlap	P	R	F1
GEMET	659	787	85	77	84.96	45.80	59.51
REEGLE	471	1033	12	19	**95.72**	31.55	47.46
Combined	779	686	104	58	85.87	**53.05**	**65.582**

Table 1: Evaluation of different term sets on climate corpus

As expected, we find the Precision is high for all 3 sets, since we expect these terms to be valid. The only reason they might not be is either through manual annotation error, ambiguity (i.e. not relevant in this particular context) or because these were subjective annotations according to the human annotator. Looking at the terms which were annotated as spurious (Table 3), we see that it is not always clear whether they should be considered as terms in this context, e.g. “crime”. One or two are ambiguous, e.g. “lead” can be a verb or a noun, but it should only be annotated as a term if it is a noun, as the two have different meanings. This can be resolved by e.g. only matching against the term list if the term found has the correct part-of-speech (in this case a noun). Others are not relevant in this context although they are in the climate ontology, e.g. “video”. Of the missing terms, we found that morphological variants of multi-word terms accounted for many, and we are working on solutions to this.

Left Context	Term	Right Context
	crime	confirms global warming/climate
eco #climate change #	**SDGs**	http://t.co/uhjhcqylUt
#arctic #climate #	**science**	@derbyuni #cryosphere
buying deforested palm oil amid	**pressure**	http://t.co/rRFya1sNmy #sustainability
climate change. gilligans	**island**	
in Jakarta to discuss increasing	**demand**	for sustainable palm oil.
oil...captivity....entertainment....	**trapping**	http://t.co/4cA6hdsu34
Climate buzzwords over	**time**	: ozone hole, el
Green Energy: Parody	**video**	exposes P&G'
The social pre	**CoP**	of climate change #UNFCCC
falling palm oil price,	**premium**	for 'sustainable' product
Indonesia should take the	**lead**	in responsible palm oil
#Solar The	**history**	of climate change http://t.co/qrU2XJljpx

climate change. Instead we	**need**	to more ... http://t.co/d5hCbFrm7p

Table 2: Spurious terms identified in the corpus

We also compared the effect of using ANNIE (Cunningham et al. 2002) and TwitIE (Bontcheva et al. 2013) as the main pre-processing part of the application. ANNIE is the default application in GATE, whereas TwitIE is designed to work specifically on tweets, and offers some advantages such as hashtag recognition, normalisation of some common abbreviations, and better processing of irregular capitalisation. Table 3 shows the comparison on the combined term set: we see that while precision is slightly lower (due to the more relaxed parameters of TwitIE on non-standard text), the recall and F1 are higher. We therefore use TwitIE as the main pre-processing component in all further evaluations and in the web service.

Term Set	Match	Missing	Spurious	Overlap	P	R	F1
Combined - ANNIE	679	797	61	47	**89.26**	46.13	60.82
Combined - TwitIE	779	686	104	58	85.87	**53.05**	**65.58**

Table 3: Comparison of ANNIE and TwitIE as pre-processor on the climate corpus

Next, we performed the same evaluation using only the high-ranked terms as gold standard, and using only the high- and medium-ranked terms, since we were not confident that the low- and medium-ranked terms were really valid. Of the original 1523 terms, 1154 were marked with high confidence, 320 terms with medium and 40 with low confidence. The results for the experiment are shown in Table 4. As expected, the Recall increased but the Precision decreased as we restricted the terms in the gold standard set to be higher quality. This indicates also that we might need a second pass over the annotated gold standard data to check the quality of the annotations. However, removing the low confidence annotations from the set does not improve the application significantly. For the remaining evaluations, we used the full set of terms from the manual annotation.

Term Set	Match	Missing	Spurious	Overlap	P	R	F1
H	655	467	233	32	72.93	**58.15**	64.71
H+M	775	665	111	34	**86.09**	53.73	**66.17**
H+M+L	779	686	104	58	85.87	53.05	65.58

Table 4: Evaluation against high- and medium-ranked terms in climate corpus

Finally, we compared the annotations found in GEMET and Reegle against those found by TermRaider, which performs single and multi-word term recognition based on tf.idf and other statistical measures. In the climate change corpus, when using the whole TermRaider output as the gold standard, we achieved overall a Precision of 98.05%, Recall of 26.14% and an F1 of 41.28% using the combined term application. This shows that in general, similar kinds of errors are probably made by our corpus-based application, TermRaider, as by our ontology-based applications. We also compared TermRaider output with the manually annotated gold standard, getting a Precision of 45.64%, Recall of 74.49% and an F1 of 56.60%.

Since TermRaider ranks candidate terms (essentially Noun Phrases) in order of termhood, and then applies a threshold to only consider the top-ranked terms, we split the extracted list from TermRaider into 3 sections: top, middle and bottom. We would expect high correlation between the top-ranked terms and the GEMET/REEGLE terms. We would also expect to find many terms in the middle and bottom sections of the TermRaider list that were not present in GEMET/REEGLE. Table 5 shows some examples of terms in different sections of the TermRaider list, while Table 6 shows some examples of terms found only in TermRaider, REEGLE and GEMET respectively.

High	Medium	Low
solar energy	unmanned aircraft	world cup
hydraulic fracturing	medical center	mental illness
ice sheet	cell phone	heart disease

Table 5: High, medium and low-ranked terms extracted by TermRaider

TermRaider only	GEMET only	REEGLE only
Arctic biodiversity	agriculture	sustainability
abrupt climate change	deforestation	anthropogenic climate change
renewable energy	Antarctica	geothermal
evolution	biofuel	biodiesel
shark	ecology	palm oil industry

Table 6: Terms unique to each application

2.6. Energy corpus

The second corpus consisted of 413 tweets about energy-related issues, collected via the MWCC using the keywords “home display”, “energy monitor” and “energy-bill”. We performed the original term extraction experiment on this corpus, to see how it differed on a slightly different subdomain. From the results shown in Table 5, we see that while REEGLE has the best Precision, it has terrible recall, only finding 52 matches. This was partly because terms such as “energy” were not in REEGLE, and occurred hundreds of times in the corpus. All the results are a little bit lower than the climate corpus in the first experiment, possibly also because the climate corpus was used for some initial development.

Term Set	Match	Missing	Spurious	Overlap	P	R	F1
GEMET	448	1008	135	72	77.65	36.26	49.44
REEGLE	48	1582	0	8	**92.86**	03.17	06.14
Combined	559	1004	103	75	80.94	**36.42**	**50.23**

Table 7: Results of term extraction in the energy corpus

2.7. Fracking corpus

Finally, we performed the same experiment on a third corpus of 352 tweets, collected using the keywords "fracking", "arctic site" and "drill". The results are shown in Table 6 and are slightly higher than those on the energy corpus, albeit with lower Precision. This is probably due to increased term ambiguity.

Term Set	Match	Missing	Spurious	Overlap	P	R	F1
GEMET	337	716	176	88	63.39	33.39	43.74
REEGLE	69	1066	12	6	**82.76**	06.31	11.73
Combined	572	491	137	78	77.64	**53.55**	**63.38**

Table 8: Results of term extraction in the fracking corpus

2.8. Software availability

The term extraction application is available via a web service at:

http://services.gate.ac.uk/decarbonet/term-recognition

The web service is publicly available; the final version will be made open source. It takes a document as input, and outputs the text asa JSON document of standoff annotations with term and URI information.

The term recognition web service requires the text to be processed to be passed using one of the following three request parameters.

Parameter	Supported Request	Description
text	GET or POST	Plain text to process
url	GET or POST	The URL of a document to process
file	POST	A file to process.

The response from the service is a simple XML document containing just two elements: Document and Term. There is a single Document element which acts as the root element of the response document. Within the Document element is the processed text. Any word or phrase annotated by the application is encapsulated in a

Term element. Each Term element has an Instance attribute which is the URI of an ontology instance which is equivalent to the annotated text. For example, processing the text "cars pollute by emitting carbon dioxide" results in the following document (simplified to only show those attributes listed above):

```
<Document>
<TermInstance="http://www.eionet.europa.eu/gemet/concept/1148">cars
</Term> pollute by emitting
<Term Instance="http://reegle.info/glossary/1064">carbon dioxide</Term>
</Document>
```

3. Web service for indicators

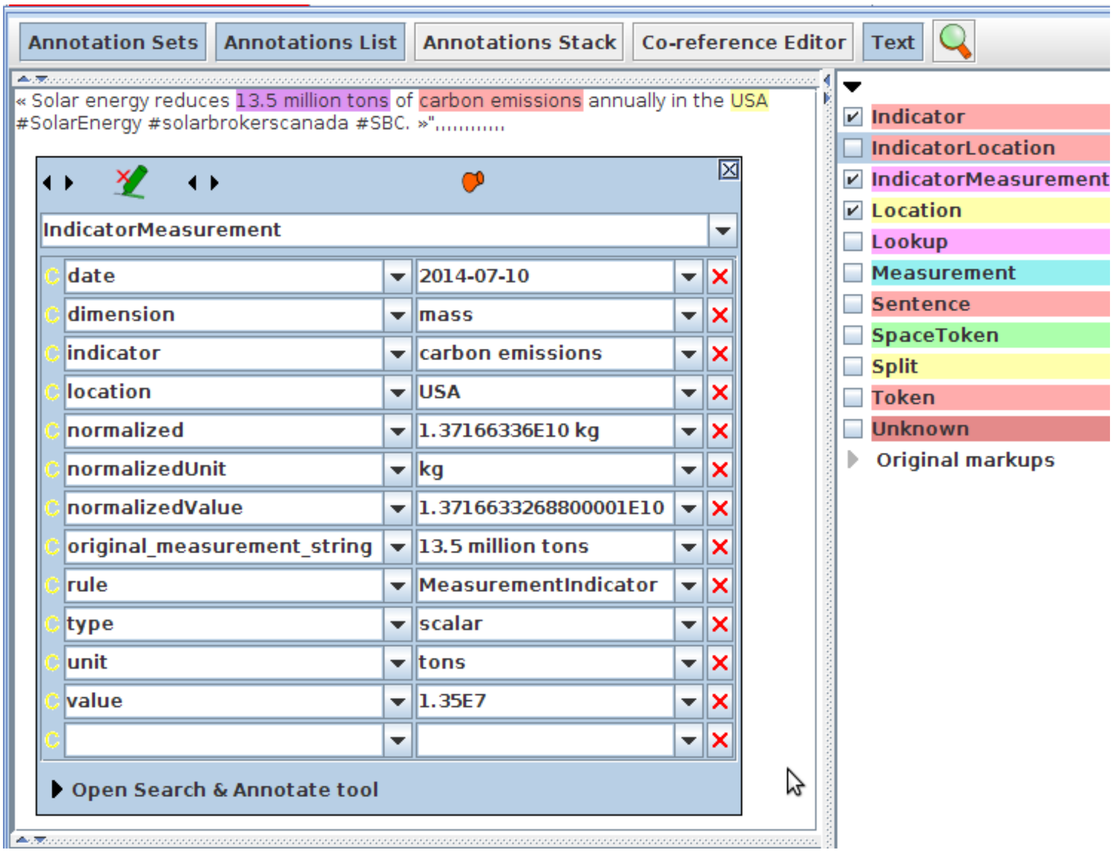
This web service complements the term recognition service ClimaTerm, by identifying indications of the presence of a quantitative measurement related to climate change. For example, this might include changes in mortality rates for a country or population, percentage decrease in forest areas and so on.

3.1. Extracting climate change indicators

The application aims to extract useful indicators of climate change such as “energy use”, “carbon pollution”, etc. for particular locations, together with measurable effects such as percentages, measurements etc. It uses a manually compiled list of indicator seed terms (e.g. “energy use”) plus processing resources from the core GATE tools to extract location and dates (via ANNIE), measurements and percentages (via our Measurements plugin) and additional terms (using TermRaider). The measurements are also normalised to their SI unit, so that the same measurements in different systems (e.g. acres and square metres) can be equated. This normalisation process is described more fully in (Cunningham et al, 2011).

Following the creation of the initial seed list of 42 terms, we then started an incremental process to retrieve further terms, by extracting relevant tweets from the MWCC using the terms as keywords (to ensure relevance), and then processing the tweets with the application and investigating the other top-ranked terms found in those tweets using TermRaider. New terms were added to the list and the process repeated.

Examples of the output of the application are shown in Figures 3 and 4. Figure 3 shows a measurement with its normalisation information. Figure 4 shows a percentage with also the date and normalised date information.



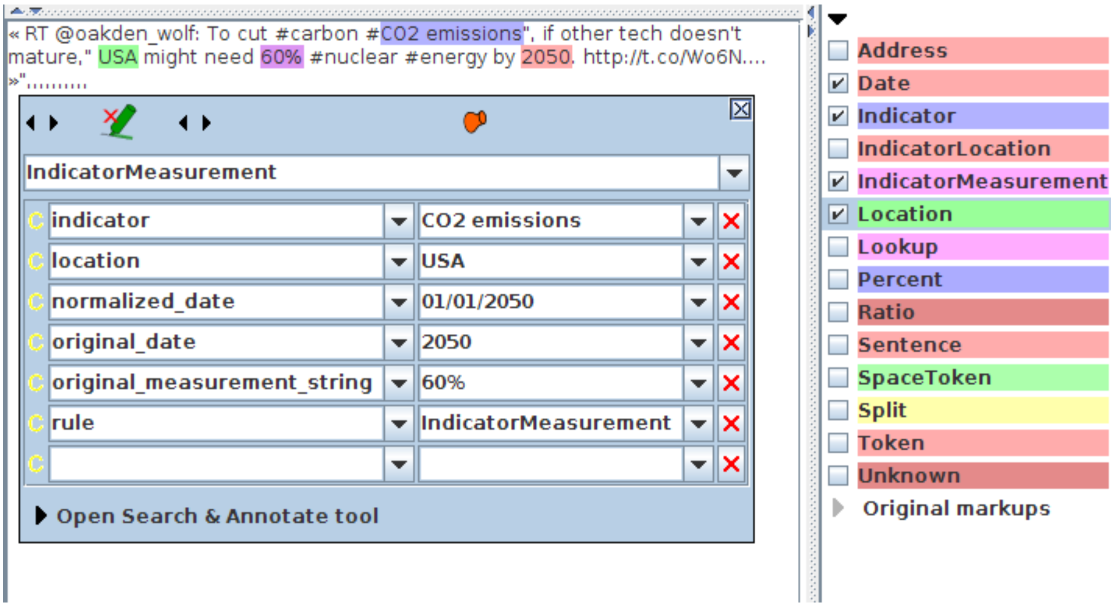
« Solar energy reduces 13.5 million tons of carbon emissions annually in the USA #SolarEnergy #solarbrokerscanada #SBC. »",..........

IndicatorMeasurement

date	2014-07-10
dimension	mass
indicator	carbon emissions
location	USA
normalized	1.37166336E10 kg
normalizedUnit	kg
normalizedValue	1.3716633268800001E10
original_measurement_string	13.5 million tons
rule	MeasurementIndicator
type	scalar
unit	tons
value	1.35E7

☑ Indicator
□ IndicatorLocation
☑ IndicatorMeasurement
☑ Location
□ Lookup
□ Measurement
□ Sentence
□ SpaceToken
□ Split
□ Token
□ Unknown
Original markups

Figure 3: Screenshot of a tweet annotated with Indicator information in GATE



« RT @oakden_wolf: To cut #carbon #CO2 emissions", if other tech doesn't mature," USA might need 60% #nuclear #energy by 2050. http://t.co/Wo6N.... »"..........

IndicatorMeasurement

indicator	CO2 emissions
location	USA
normalized_date	01/01/2050
original_date	2050
original_measurement_string	60%
rule	IndicatorMeasurement

□ Address
☑ Date
☑ Indicator
□ IndicatorLocation
☑ IndicatorMeasurement
☑ Location
□ Lookup
□ Percent
□ Ratio
□ Sentence
□ SpaceToken
□ Split
□ Token
□ Unknown
Original markups

Figure 4: Screenshot of a tweet annotated with indicator and date information in GATE

The application checks for the presence of the following in the tweet:

- an indicator (via gazetteer lookup)

- a location
    - Locations are first identified in the body of the text or in hashtags via ANNIE
    - If no Location is found, the usernames are checked to see if a location is contained there (e.g. @USAToday, @age_uk).
    - Failing this, the tweet metadata is checked for the presence of a value of target_location and this is used instead)
- a measurement or percentage (via the Measurements plugin and ANNIE respectively)
- a date
    - The body of the tweet is first checked for mention of a date.
    - The date in the text is normalised, so that they are all represented in the same format (DD-MM-YYYY) and so that relative dates are represented as absolute dates with respect to today's date. For example, if today is 1 September 2014, a mention of "tomorrow" in the text will be represented as 02-09-2014. A feature also tells us whether the date is in the past, present or future.
    - If this fails, the tweet metadata is checked for the presence of a date.

The application has not been formally evaluated yet. However, initial observation of the results shows that there are a few cases where items are missing, largely because of more complex grammar structure (for instance, coordinations of amounts and dates are not always correctly dealt with). Other errors are fairly rare, but generally due to errors in the subcomponents,e .g. if a Location is wrongly identified or missing. In general, the results are of high quality, however.

3.2. Software availability

The first version of the climate change indicator application is available via a web service at:

http://services.gate.ac.uk/decarbonet/indicators/

The web service is publicly available; the final version will be made open source. The service takes a document as input, and outputs the text as an XML file annotated with term and URI information.

The indicators web service requires the text to be processed to be passed using one of the following three request parameters. The response from the service is a simple XML document containing just two elements: Document and Term. There is a single Document element which acts as the root element of the response document. Within the Document element is the processed text. Any word or phrase annotated by the application is encapsulated in an IndicatorMeasurement element. IndicatorMeasurement elements include a number of attributes which relate to the indicator, the location, measurement, and date.

For example, processing the text "to cut carbon CO2 emissions if other tech doesn't mature USA might need 60 percent nuclear energy by 2050" results in the following document (simplified to only show those attributes listed above):

<Document>to cut carbon CO2 emissions if other tech doesn't mature USA might need
<IndicatorMeasurement original_measurement_string="60 percent" original_date="2050"
indicator="CO2 emissions" location="USA"
normalized_date="01/01/2050">60 percent</IndicatorMeasurement> nuclear energy by 2050
</Document>

4. Recognyze: a service for named entity recognition

Which organizations tend to have a negative reputation among environmental stakeholders? Who are the most visible climate change activists, and what are mainstream media associating with their recent public appearances?

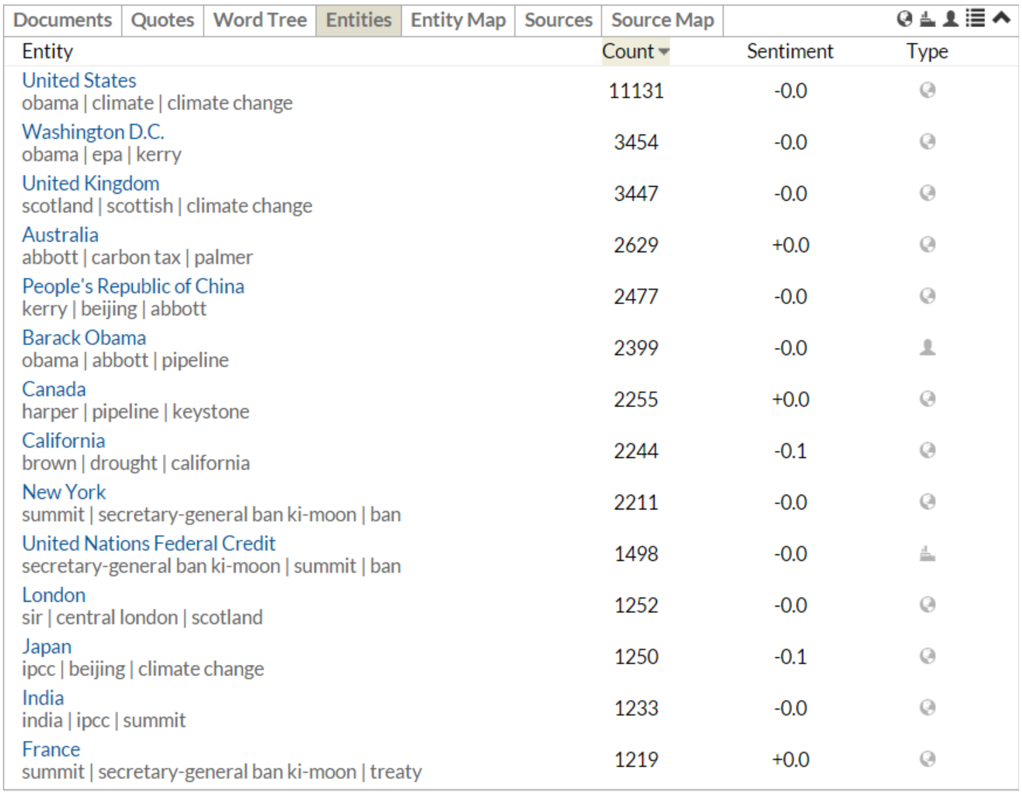
For properly answering such communication questions, the Media Watch for Climate Change (MWCC) (Scharl et al. 2013) has been updated with a named entity recognition and resolution component called Recognyze (Scharl et al. 2014; Weichselbraun et al. 2014) that draws upon structured external knowledge repositories such as DBpedia.org, Freebase.com and GeoNames.org to identify and disambiguate named entities (organizations, persons and locations), assigning confidence values to align them with the items contained in the external knowledge repositories.

Applying the Recognyze component to annotate the knowledge repository of the Media on Climate Change helps to better understand environmental networks and the dynamic relations among the actors in these networks.

4.1. Analysing Named Entities

In contrast to other approaches, Recognyze does not apply machine learning and therefore does not require training corpora or iterative learning steps. Analytics components extract relevant company names as well as contextual and structural information, which is then used for named entity linking and ranking. While the literature reports higher accuracies for some methods that apply machine learning techniques, Recognyze is more flexible than these approaches since it:

- is not limited to a particular knowledge source;
- does not require training or annotated training corpora, but can be deployed for any domain or language as long as appropriate linked data resources such as DBpedia are available; and
- offers a good overall performance even with connected to very large knowledge bases.



Entity	Count	Sentiment	Type
United States obama | climate | climate change	11131	-0.0	🌐
Washington D.C. obama | epa | kerry	3454	-0.0	🌐
United Kingdom scotland | scottish | climate change	3447	-0.0	🌐
Australia abbott | carbon tax | palmer	2629	+0.0	🌐
People's Republic of China kerry | beijing | abbott	2477	-0.0	🌐
Barack Obama obama | abbott | pipeline	2399	-0.0	👤
Canada harper | pipeline | keystone	2255	+0.0	🌐
California brown | drought | california	2244	-0.1	🌐
New York summit | secretary-general ban ki-moon | ban	2211	-0.0	🌐
United Nations Federal Credit secretary-general ban ki-moon | summit | ban	1498	-0.0	🏛
London sir | central london | scotland	1252	-0.0	🌐
Japan ipcc | beijing | climate change	1250	-0.1	🌐
India india | ipcc | summit	1233	-0.0	🌐
France summit | secretary-general ban ki-moon | treaty	1219	+0.0	🌐

Figure 5: Locations, people and organizations in association with "climate change" in 2014 Anglo-American news media coverage

Initial evaluations show that Recognyze successfully disambiguates and grounds named entities in settings where a lot of similarly named alternatives and collisions occur – for example, ambiguous names or acronyms of organizations such as WWF, which stands for both the Worldwide Fund for Nature and the World Wrestling Federation. Depending on the evaluation corpus used, Recognyze yields a recall of 0.72 for identifying the most relevant organization in an article, and an F1 measure of up to 0.63 for named entity linking, without source-specific optimizations or human interventions (Weichselbraun et al. 2014).

4.2. Software availability

The first version of the Recognyze web service is available at http://triple-store.ai.wu.ac.at/. Given a text input, the Recognyze service returns a set of named entities, together with their start and end positions within the input text. Under the hood, Recognyze makes use of open data portals such as DBPedia and GeoNames for its queries, returning predefined subsets (property-wise) of respective entities. Note that service usage is limited to 100 requests per day (max. 1MB data transfer per request).

When querying, a search profile to search within must be provided. A search profile describes a domain from the real world; currently the following set of domains exists:

{en,de}.organization.ng

Organizations in English and German, taken from DBpedia. Returns type.

{en,de}.people.ng

Person names in English and German, taken from DBpedia. Returns type.

{en,de,fr}.geo.50000.ng

Geolocations (cities, countries) with a population larger than 50000, taken from GeoNames. Returns type.

Passing multiple profiles at once is also supported by the API.

The REST interface can easily be accessed via the open source webLyzard API at https://github.com/weblyzard/weblyzard_api. For more information on the API, please consult the documentation at http://weblyzard-api.readthedocs.org/en/latest/.

Recognyze returns a JSON list object of all entities found. For each entity found, the service returns the entity type, the associated search profile (see above), the entity's occurrences within the given text (start, end, sentence, surface form), the confidence of the correctness of the entity, the public key where the entity links to (e.g. http://sws.geonames.org/4990729), as well as extra properties where available.

5. Summary and further work

In this deliverable we have described the tools we have developed for environmental information extraction. This includes tools to perform entity disambiguation, recognition of environmental terms, and extraction of environmental indicators, realised as web services. We have experimented with different ontologies for the term extraction, and after some preliminary evaluations have made improvements to the performance of the tools for term and indicator recognition by incorporating further natural language processing. For example, terms found in the existing ontologies are not sufficient to use for extraction alone as they may occur in different forms in the text, e.g. as part of hashtags and with different lexical realisations. Furthermore, these terms are ambiguous and care has to be taken to resolve these issues in order to avoid overgeneration of recognised terms. Initial results are promising and will be used by other project partners; however, further improvements are still necessary and are ongoing. In terms of Recognyze, future work will focus on (i) further improving the disambiguation performance by considering more complex structural knowledge in the linking process, (ii) optimizing and evaluating disambiguation profiles that work with publicly available sources such as DBpedia, and (iii) providing evaluations for other entity types such as people and locations.

A. List of Figures

Figure 1: Annotation of a term variant in GATE 6

Figure 2: Screenshot of a decomposed hashtag in GATE 7

Figure 3: Screenshot of a tweet annotated with Indicator information in GATE 13

Figure 4: Screenshot of a tweet annotated with indicator and date information in GATE 13

Figure 5: Locations, people and organizations in association with “climate change” in 2014 Anglo-American news media coverage 16

B. List of Tables

Table 1: Evaluation of different term sets on climate corpus 8

Table 2: Spurious terms identified in the corpus 9

Table 3: Comparison of ANNIE and TwitIE as pre-processor on the climate corpus .. 9

Table 4: Evaluation against high- and medium-ranked terms in climate corpus 9

Table 5: High, medium and low-ranked terms extracted by TermRaider 10

Table 6: Terms unique to each application 10

Table 7: Results of term extraction in the energy corpus 11

Table 8: Results of term extraction in the fracking corpus 11

C. List of Abbreviations

Abbreviation	**Explanation**
CA	Consortium agreement
DoW	Decription of work, i.e. GA - Annex I
EC	European commission
GA	Grant agreement
IP	Intellectual property
IPR	Intellectual property rights
PC	Project coordinator
PMB	Project management board
SC	Scientific Coordinator
PO	Project officer
PSB	Project steering board
DM	Data Manager
AB	Advisory board
WP	Work package

D. References

[Cunningham et al. 2002] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[Bontcheva et al. 2013] K. Bontcheva, L. Derczynski, A. Funk, M.A. Greenwood, D. Maynard, N. Aswani. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013).

[Maynard et al. 2014] Diana Maynard and Mark A. Greenwood. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. Proc. of LREC 2014, Reykjavik, Iceland, May 2014.

[Cunningham et al. 2011] H. Cunningham, V. Tablan, I. Roberts, M. A. Greenwood, & N. Aswani (2011). Information extraction and semantic annotation for multi-paradigm information management. In Current Challenges in Patent Information Retrieval(pp. 307-327). Springer Berlin Heidelberg.

[Scharl et al. 2013] Arno Scharl, Alexander Hubmann-Haidvogel, Albert Weichselbraun, Heinz-Peter Lang, Marta Sabou (2013). Media Watch on Climate Change -- Visual Analytics for Aggregating and Managing Environmental Knowledge from Online Sources, 46th Hawaii International Conference on System Sciences, pp. 955-964.

[Scharl et al. 2014] Scharl, A., Kamolov, R., Fischl, D., Rafelsberger, W. and Jones, A. (2014). Visualizing Contextual Information in Aggregated Web Content Repositories. 9th Latin American Web Congress (LA-WEB 2014). Ouro Preto, Brazil: Forthcoming.

[Weichselbraun et al. 2014] Weichselbraun, A., Streiff, D. and Scharl, A. (2014). Linked Enterprise Data for Fine Grained Named Entity Linking and Web Intelligence. 4th International Conference on Web Intelligence, Mining and Semantics (WIMS-2014). Thessaloniki, Greece: ACM Press.

DecarboNet Consortium

The Open University
Walton Hall
Milton Keynes MK7 6AA
United Kingdom
Tel: +44 1908652907
Fax: +44 1908653169
Contact person: Jane Whild
E-mail: h.alani@open.ac.uk

Waag Society
Piet Heinkade 181A
1019HC Amsterdam
The Netherlands
Tel: +31 20 557 98 14
Fax: +31 20 557 98 80
Contact person: Tom Demeyer
E-mail: tom@waag.org

MODUL University Vienna
Am Kahlenberg 1
1190 Wien
Austria
Tel: +43 1320 3555 500
Fax: +43 1320 3555 903
Contact person: Arno Scharl
E-mail: scharl@modul.ac.at

WWF Schweiz
Hohlstrasse 110
8004 Zürich
Switzerland
+41 442972344
Contact person: Christoph Meili
E-mail: Christoph.Meili@wwf.ch

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
United Kingdom
Tel: +44 114 222 1930
Fax: +44 114 222 1810
Contact person: Kalina Bontcheva
E-mail: K.Bontcheva@dcs.shef.ac.uk

Green Energy Options
Main Street, 3 St Mary's Crt
Hardwick CB23 7QS
United Kingdom
+44 1223850210
+44 1223 850 211
Contact person: Simon Anderson
E-mail: simon@greenenergyoptions.co.uk

Wirtschaftsuniversität Wien
Welthandelsplatz 1
1020 Wien
Austria
Tel: +43 31336 4756
Fax: +43 31336 774
Contact person: Kurt Hornik
E-mail: kurt.hornik@wu.ac.at