EC Project 610829

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

D2.1: Data Acquisition Infrastructure

31 March 2014

Version: 0.4

Version history

Version	Date	Author	Comments
0.1	09/03/2014	A. Scharl	Initial Version
0.2	18/03/2014	H.-P. Lang	Unstructured Sources; Process Description
0.3	25/03/2014	A. Scharl	Document Revision
0.4	29/03/2014	A. Scharl	Document Revision

Peer reviewed by: Harith Alani, The Open University

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 document summarizes the output of Task 2.1, Acquiring Structured and Unstructured Data. Although there is a trend towards combining multiple data sources, integrated approaches tapping into multiple structured, unstructured, and social sources at various levels of abstraction are still scarce. DecarboNet addresses this shortcoming and builds upon the Media Watch on Climate Change[1] when developing the required methods to contextualise environmental knowledge and track collective awareness in a granular manner.

T2.1 utilizes and extends existing open source libraries including the (i) *extensible Web Retrieval Toolkit* (eWRT),[2] a modular open-source Python API to retrieve social data from Web sources, and (ii) the TwitIE plugin[3] for the GATE open source natural language processing platform:

- **Unstructured Sources**. We have developed a dynamic crawling method with improved metadata handling and superior caching control, customized the content acquisition services to the large-scale awareness campaigns of WP6, and provided an API to exporting data in either JSON or XML format.

- **Structured Sources.** The initial version of the DecarboNet Linked Open Data repository contains data sets from DBpedia,[4] Geonames[5] and Freebase.[6] Some of these data sets are simply dumps of the triples, while others have required a more elaborate process including slicing, cleaning and fixing steps. Initially, we have decided to use the Sesame RDF repository, but will also explore alternative storage tools depending on the size and complexity of the harvested data.

- **Social Sources.** DecarboNet requires user-generated content to capture threaded dialogs and identify trends in public discourse. In addition to utilizing the Twitter Streaming API for on-going monitoring, we have added support of the Search API for ad-hoc queries and historical data.[7]

A configuration interface guides the setup and specifies the workflow on how to gather and process these sources (this includes topic definition, the specification of mirroring intervals for crawled resources and RSS feeds, and the access parameters for social media APIs). Deliverable 2.1 summarizes the required specifications and provides a detailed account on the composition of the evolving DecarboNet knowledge repository.

[1] www.ecoresearch.net/climate
[2] www.weblyzard.com/ewrt
[3] gate.ac.uk/wiki/twitie.html
[4] www.dbpedia.org
[5] www.geonames.org
[6] www.freebase.com
[7] dev.twitter.com/docs/using-search

Table of Contents

1. Introduction

The presented work utilizes and extends the *extensible Web Retrieval Toolkit* (eWRT),[8] a modular open-source Python API to retrieve social data from Web sources such as Delicious, Flickr, Yahoo! and Wikipedia, including various helper classes for effective caching and data management (Weichselbraun et al., 2013). The toolkit is developed by MOD and a group of academic partners. In addition to components for content acquisition and caching, it also provides low-level natural language processing functionalities such as phonetic string similarity measures and methods for string normalization.

In addition to eWRT, the pre-processing of unstructured social media content for DecarboNet also uses a GATE open-source plugin for social media analysis called TwitIE (Bontcheva et al., 2013).[9] The plugin was developed and released as part of the TrendMiner and uComp research projects. In particular, we make use of the tweet tokeniser (identifies individual words, user names, hashtags, URLs, etc), language identification components, part-of-speech tagger, and tweet normaliser. The latter normalises expressions like “2moro” into their full lexical variants - i.e., “tomorrow”.

2. Data Acquisition

To provide citizens with a continuous supply of relevant digital content, DecarboNet provides a portfolio of scalable methods to acquire, aggregate and process a rich stream of information from unstructured, structured and social (user profiles, social network data, user-generated content such as blogs, folksonomies) data sources.

2.1. Unstructured Sources

This category includes Web content and the archives of core and associate partners. Unstructured content is well covered by existing technologies of the consortium (Scharl et al., 2013), The existing platform had to be extended and improved in several ways to accommodate the goals of DecarboNet:

- Development of a more dynamic crawling architecture based on scrapy[10] to replace the previously used HTTrack.[11] The new architecture offers improved metadata handling (e.g., when gathering social media content via OpenGraph), superior caching control, and is better suited for deployment in distributed scenarios. It also facilitates the collection of embedded hyperlink references.

[8] www.weblyzard.com/ewrt
[9] gate.ac.uk/wiki/twitie.html
[10] www.scrapy.org
[11] www.httrack.com

- For customizing the content acquisition services to the requirements of the large-scale awareness campaign use case (WP6), we also improved the computation of content relevance as well as the monitoring and quality control infrastructure. For this purpose, we adopted the NAGIOS-based *Open Monitoring Distribution*[12] to (i) track the quantity of gathered documents, (ii) report errors that occur during data collection and data analysis, (iii) scan logfiles for errors and ensure that they only happen once, (iv) monitor the backup process.

- An Application Programming Interface (API) to allow exporting data in either JSON or XML format, for reusing the gathered social media content in third-party applications.

2.2. Structured Sources

The DecarboNet collective awareness platform will utilize linked open data sources to semantically enrich unstructured data, and gather sustainability indicators from other open data sources such as EuroStat,[13] United Nations,[14] World Bank[15] and the U.S. Open Data Repository[16] – e.g., CO2 (metric tons), power consumption (kWh) or energy use (kg oil equivalent) as baselines for comparing individual consumption patterns to national, European and global averages.

Open Data Acquisition

DecarboNet acquires open data by making use of the data interfaces of the various open data sources mentioned above, which typically offer very diverse exchange formats – ranging from large CSV/XLS dumps to APIs that return JSON-encoded data and, in very rare cases, to SPARQL endpoints. We will continue to refine the DecarboNet suite of data acquisition methods in order to ensure compatibility with a variety of input data types. Whenever necessary, the harvested data will be mapped into appropriate ontological models, and the resulting semantic metadata will be stored in dedicated triple stores. Initially, we have decided to use the Sesame RDF repository, but will also explore alternative storage tools depending on the size and complexity of the harvested data.

Linked Data Cleanup and Preprocessing

The initial version of the DecarboNet Linked Open Data repository contains datasets from multiple sources: DBpedia, Geonames, and Freebase. Some of these datasets are simply dumps of the triples, while others have required a more elaborate process including slicing, cleaning and fixing steps. For Geonames, it was enough to use the dumps generated by Python scripts, as it contained only geographical entities (Places), and hardly any embedded links. Each DBpedia dataset contains all the

[12] www.omdistro.org
[13] epp.eurostat.ec.europa.eu
[14] data.un.org
[15] data.worldbank.org
[16] www.data.gov

triples related to a single entity type in a single language (slice by language and entity type). DBpedia was organized from the beginning around several main types of entities (People, Organisations, Places, Works, Species, Things, Buildings, Planets), and we decided that it makes sense for the initial DecarboNet repository to keep this distinction of classes. We have used a series of bash and Python scripts to create the datasets and upload them to a Sesame triple store. We removed malformed URLs, bad formatting or encoding, and some of the internal links that were not needed (cleaning and fixing steps). While it can be argued that these pre-processing steps might cause some loss of contextual information, we have not removed any disambiguation links or references to other DBpedia-supported languages. All DBpedia datasets were based on the recent 3.9 version, which also uses a type inference mechanism to check data types (Paulheim et al, 2013). For Freebase, we used the BaseKB build, which was provided in a cleaned and preprocessed state.

2.3. Social Sources

DecarboNet aims to integrate citizen-provided content, social graphs, resource annotations and other types of metadata. WWF and the Associate Partners of DecarboNet maintain large online communities. Harvesting their users' lifestreams reflects their online activities and allows us to enrich dynamic user models, which will in turn help identify users across networks.

When developing the DecarboNet data acquisition processing pipeline, we have placed special emphasis on its flexibility and scalability. Distributed, focused on-demand crawling (T2.1) captures user-generated content, threaded dialogs and trends in public discourse patterns. In addition to replacing HTTrack with scrapy (see above), we have also included support of the Twitter Search API, complementing the Streaming API that is used for the ongoing monitoring. Individual and collective profiles based on dynamic user models (T4.1) guide the data acquisition process and help to customize information services to the needs of individual citizens and stakeholder groups alike.

3. Data Acquisition Process and Configuration

A configuration interface streamlines the setup of all sources and the workflow on how to process them. This includes the topic configuration, the specification of different mirroring intervals for crawled resources and RSS feeds, and the access parameters for social media APIs. The architecture is scalable and straightforward to extend. Processing resources in the form of virtual machines can be added on the fly. Tasks are being distributed by celery[17] and added to RabbitMQ,[18] from which the workers get pending tasks and process them.

For feed mirroring, the data acquisition system scans crawled HTML pages for RSS streams while detecting duplicates and resolving timezone issues. Once the mirroring of all documents is complete, the use case relevance is being calculated based on

[17] www.celeryproject.org

[18] www.rabbitmq.com

the set of regular expressions contained in the DecarboNet input filter. The data acquisition component requires the following specifications:

- **Regexp.** A term or set of regular expressions to search for.
- **Weight.** This value reflects the importance of a Regexp. We have decided to set this value to 1.0 for standard terms, and to 10.0 for blacklist items.
- **Term Lists for Social Media Sources.** The peculiarities of the APIs of social media platforms require a more specific approach:
    1. *Unique Single Words.* Social media sources will be searched for these terms and will be added to the DecarboNet repository *without* any further relevance checks. Use only single words, because most social media APIs do not support multi-word search queries.
    2. *Ambiguous Single Words.* Social media sources will be searched for these more general terms, but will be subjected to the Whitelist RegExp (see below). This two-step process ensures high recall (capturing a large number of documents in phase 1) and high precision (filtering out irrelevant documents using the combination of white- and blacklists in phase 2).
- **Whitelist Regexp**. Only articles matching this list are being added to the DecarboNet portal. The Whitelist Regexp is applied to most documents (news, Web sites, as well as social media content stemming from the search for “ambiguous single words”). The only exception are unique single words from social media, if the option to add all documents from a specific source is chosen.
- **Blacklist Regexp.** This list is used to filter out irrelevant articles.
- **Topic_Name** groups input terms, with the option to make them accessible via the topic navigator of the Web portal (list of topics on the left). This allows using a different set of topic_name groups for each individual source.
- **Comment.** To explain or discuss the chosen term or Regexp.
- **Keyword Stoplist.** This list elicits terms that should be ignored in the keyword computation and therefore will not show up in the ‘Keyword Graph’ or ‘Associations’ windows.

4. DecarboNet Knowledge Repository

The efforts to adapt and extend the existing content aggregation platform to the use case requirements included:

- the development of a German version of the *Media Watch on Climate Change*, currently scheduled for public release in the second quarter of 2014 at www.ecoresearch.net/climate/de,

- the integration of new content sources; e.g. Web sites of Swiss, Austrian and German municipalities to track "Public Lighting of Municipalities", the Earth Hour 2014 focus theme of WWF Switzerland,[19] and

- the revision of the input filter as outlined in Table 1 – including use case-specific topics such as "earth hour" and "palm oil".[20]

Table 1: Topic definition for the awareness campaign use case

Topic Definition	Number of Regular Expressions
wwf.de.social_media.general	18
wwf.de.social_media.specific	99
wwf.de.whitelist	113
wwf.en.social_media.general	34
wwf.en.social_media.specific	71
wwf.en.whitelist	110

The new sample to provide **German content streams** for the DecarboNet collective awareness platform comprises the following sources:

- **News Media:** CH (244), AT (172), DE (63).
- **Social Media:** German Twitter, Facebook, Google+ and YouTube postings.
- **NGOs:** ch.wwf.ngo (197); climate.unfccc (1250).
- **Municipalities:** CH (2318), AT (616), DE (753).

This complements the sources of the **English portal**, which has also been extended to accommodate specific use case requirements:

- **News Media:** 163 (US, CA, UK, ZA, AU, NZ).
- **Social Media:** English Twitter, Facebook, Google+ and YouTube postings.
- **NGOs:** ch.wwf.ngo (197); climate.unfccc (1250); us.npo (32).
- **Fortune:** Top 1000 U.S. companies in terms of revenue.

[19] earthhour.wwf.ch/de/earthhour

[20] www.wwf.ch/de/hintergrundwissen/wald/bedrohung/palmoelsoja

5. Conclusion

By providing the required knowledge base, the data acquisition services outlined in this deliverable are an important step in achieving the goals of DecarboNet. The methods have been deployed and will be continuously refined and aligned with semantic methods to pre-process, integrate, annotate and classify the gathered content. To enrich the gathered unstructured content and metadata with semantics in T2.2, DecarboNet will use extra-linguistic knowledge. Pursuing this enrichment process in a manual manner is prohibitively expensive and unsustainable, since linked data vocabularies typically contain millions of instances. DecarboNet will therefore deploy LOD-based semantic enrichment algorithms that are scalable to be able to process large repositories, and tailored to the environmental science domain to achieve the required precision.

A. List of Figures

none

B. List of Tables

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

[Piccolo el al, 2013] Piccolo, L.S.G. et al (2013) Designing to Promote a New Social Affordance for Energy Consumption. Proc. of 12th IFIP Conference on e-Business, e-Services, e-Society. "Collaborative, trusted and privacy aware e/m-services" - I3E' 2013, 213-225

[Bontcheva et al 2013] Bontcheva, K., Derczynski, L., et al. (2013): An Open-Source Information Extraction Pipeline for Microblog Text. *Int'l Conference on Recent Advances in Natural Language Processing* (RANLP-2013). Hissar, Bulgaria. 83-90.

[Paulheim et al 2013] Paulheim, H. and Bizer, C. (2013). "Type Inference on Noisy RDF data," The Semantic Web - *12th International Semantic Web Conference (ISWC-2013),* Proceedings, Part I. Sydney, NSW, Australia: Springer. 510–525.

[Scharl et al 2013] Scharl, A., Hubmann-Haidvogel, A., et al. (2013). "From Web Intelligence to Knowledge Co-Creation – A Platform to Analyze and Support Stakeholder Communication", *IEEE Internet Computing*, 17(5): 21-29.

[Weichselbraun et al 2013] Weichselbraun, A., Scharl, A. and Lang, H.-P. (2013). Knowledge Capture from Multiple Online Sources with the Extensible Web Retrieval Toolkit (eWRT). *7th International Conference on Knowledge Capture (K-CAP 2013).* Banff, Canada: ACM: 129-132.

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

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

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

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

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

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

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