Data and text mining

ENVIRONMENTS and EOL: identification of Environment Ontology terms in text and the annotation of the Encyclopedia of Life

Evangelos Pafilis1,*, Sune P. Frankild2, Julia Schnetzer3,4, Lucia Fanini1, Sarah Faulwetter1, Christina Pavloudi1, Katerina Vasileiadou1, Patrick Leary5, Jennifer Hammock6, Katja Schulz6, Cynthia Sims Parr6,†, Christos Arvanitidis1 and Lars Juhl Jensen2,*

1Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Crete, Greece, 2Disease Systems Biology Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, 3Max Planck Institute for Marine Microbiology, Bremen, Germany, 4Jacobs University gGmbH, School of Engineering and Sciences, Bremen, Germany, 5Marine Biological Laboratory, Woods Hole, MA 02543, USA and 6National Museum of Natural History, Smithsonian Institution, Washington, DC 20013-7012, USA

*To whom correspondence should be addressed.
†Present address: National Agricultural Library, USDA, Beltsville, MD 20705-2351, USA

Associate Editor: Jonathan Wren

Received on November 13, 2014; revised on December 23, 2014; accepted on January 18, 2015

Abstract

Summary: The association of organisms to their environments is a key issue in exploring biodiversity patterns. This knowledge has traditionally been scattered, but textual descriptions of taxa and their habitats are now being consolidated in centralized resources. However, structured annotations are needed to facilitate large-scale analyses. Therefore, we developed ENVIRONMENTS, a fast dictionary-based tagger capable of identifying Environment Ontology (ENVO) terms in text. We evaluate the accuracy of the tagger on a new manually curated corpus of 600 Encyclopedia of Life (EOL) species pages. We use the tagger to associate taxa with environments by tagging EOL text content monthly, and integrate the results into the EOL to disseminate them to a broad audience of users.

Availability and implementation: The software and the corpus are available under the open-source BSD and the CC-BY-NC-SA 3.0 licenses, respectively, at http://environments.hcmr.gr

Contact: pafilis@hcmr.gr or lars.juhl.jensen@cpr.ku.dk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

The Encyclopedia of Life (EOL; http://eol.org/) is a web resource offering biodiversity knowledge summaries of the world’s species to a vast audience (Parr et al., 2014). It currently aggregates content from more than 250 providers. These include textual descriptions about the biology, such as habitat, of more than 900 000 taxa.

The Environment Ontology (ENVO) project aims to provide a controlled, structured vocabulary to support annotation of organisms with environmental descriptors (Buttigieg et al., 2013). The ontology comprises ~1600 terms and is part of recommended (meta-)genomic metadata standards (Yilmaz et al., 2011). Having the environmental information contained in EOL annotated in the form of ENVO terms, rather than as free text, would enhance search capabilities and enable users to easily compile summary statistics on, for example, the ecological distribution of any taxa. However, manually annotating all EOL entries with ENVO terms would be
part of the curated corpus, we added 142 synonyms to the dictionary. Based on this analysis and false negatives found in the development all words occurring more than 100 times in untagged text segments. Of our evaluation set (seesupplementary information), and inspected habitat and ecology sections of 1,342,968 EOL pages that are not part EOL. To find important synonyms missing in ENVO, we tagged the for all names that appeared more than 2000 times in Medline and for homonymy. We created a block list of such names by inspecting text noun forms, we automatically generated plural and adjective forms. Synonym of the term. Because ENVO usually lists only the singular primary name, an exact synonym, a narrow synonym or a related its meaning, by ranking the terms based on whether the name was the tissues, which are better captured by other ontologies. Names con- tion), excluding broad synonyms, obsolete terms, terms describing names and synonyms from the OBO file (seesupplementary informa-

4 Performance evaluation
Since the ENVIRONMENTS tagger recognizes names within text and links them to ENVO terms, we benchmarked both aspects of its performance. To quantify to which extent the tagger recognizes the same text fragments as the annotators, we calculated precision and recall at the mention level, considering both exact and partially overlapping matches as true positives. On the evaluation part of the corpus, this resulted in 87.8% precision and 77.0% recall, corresponding to an F1 score of 82.0%. For the matches that were considered true positive for the recognition task, we further evaluated if the tagger linked them to the same ENVO terms as the annotators did. In 87.1% of cases, the tagger and the annotator agreed on at least one ENVO term (see supplementary information).

5 Annotation of EOL
To realize the full potential of any text-mining system, it is impor-
tant that it is adopted by the broad community and that its results are disseminated to the intended end users. To this end, we have integrated the ENVIRONMENTS tagger with the EOL web resource to provide users with ENVO terms for each taxon. Each month, we rerun the tagger on all English text in environment-related sections of EOL (see supplementary information). As of October 2014, this gave rise to 1,077,522 annotations of ENVO terms for 234,582 EOL taxa.

We make these environment annotations available to end users in three different ways. First, we show them within the EOL taxon web pages, which provide links to the relevant paragraphs in the textual descriptions for each ENVO term (Fig. 1). Second, they can be queried through the web interface or the application program-
ing interface of the new EOL/Traitbank semantic web data repository of organismal traits (http://eol.org/traitbank). Third, the full annotation dataset can be downloaded in tab-delimited format from http://download.jensenlab.org/EOL/.

6 Future work
The ENVIRONMENTS tagger is applicable to other large sources of text than EOL. For example, it can be applied to text fields, such as isolation source, in Genbank (Hirschman et al., 2008). Combined with the SPECIES tagger (Pafilis et al., 2013), it can also be used to
extract species–environment pairs from the scientific literature, like legacy biodiversity literature (Gwinn and Rinaldo, 2009).

**Funding**

The Encyclopedia Of Life Rubenstein Fellows Program [CRDF EOL-33066-13/E33066], the LifeWatchGreece Research Infrastructure [384676-94/GSRT/NSRF(C&E)] and the Novo Nordisk Foundation Center for Protein Research [NNF14CC0001].

Conflict of Interest: none declared.

**References**


