EXTRACT

Interactive extraction of environment metadata and term suggestion for metagenomic sample annotation

https://extract.hcmr.gr | extract@hcmr.gr

Evangelos Pafilis[1*], Pier Luigi Buttigieg[2], Barbra Ferrell[3], Emiliano Pereira[4], Julia Schnetzer[4,5], Christos Arvanitidis[1], Lars Juhl Jensen[6*]

[1] Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Crete, Greece, [2] Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Germany, [3] Delaware Biotechnology Institute, Newark, Delaware, USA, [4] Max Planck Institute for Marine Microbiology, Bremen, Germany, [5] Jacobs University gGmbH, School of Engineering and Sciences, Bremen, Germany, [6] Disease Systems Biology Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; *correspondence: pafilis@hcmr.gr, lars.juhl.jensen@cpr.ku.dk

Abstract

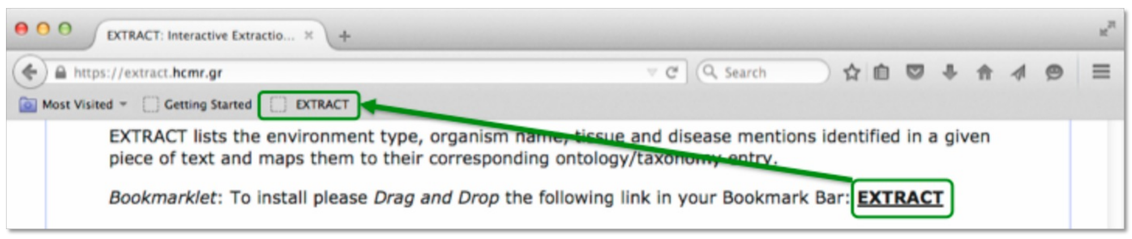
Substantial progress has been made on standardizing environment metadata of samples [1, 2]. However, manual annotation is highly labor intensive and requires familiarity with the terminologies used.

We present a web-based tool, EXTRACT, which helps curators identify and extract standard-compliant terms by combining tools for NER of environments [3], organisms [4], tissues [5] and diseases [6].

Evaluators found the system to be intuitive, well documented and accurate enough to be helpful. Compared to fully manual curation, EXTRACT sped up annotation by 15–25% and helped curators identify more terms. EXTRACT is available at https://extract.hcmr.gr/.

Availability

EXTRACT is installed by *Drag and Drop* of the bookmarklet from https://extract.hcmr.gr/ to your browser bookmark bar.

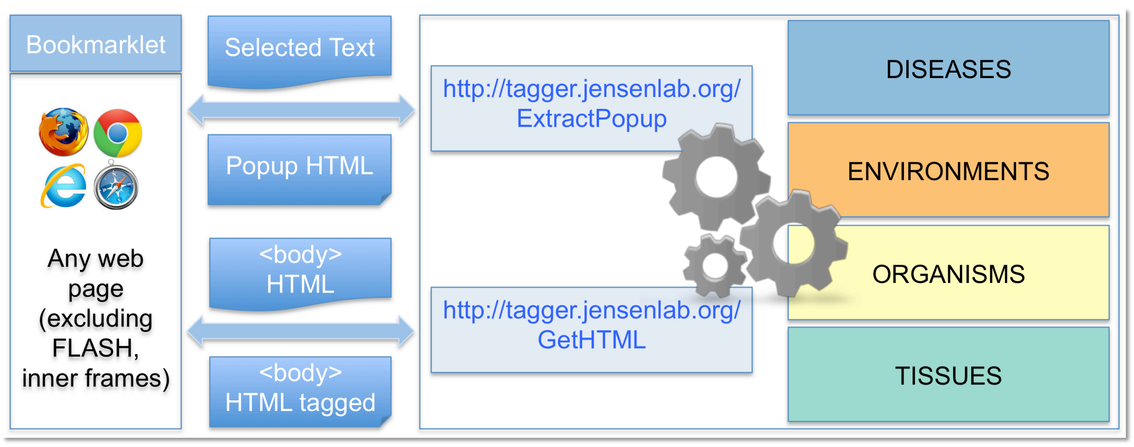


Thorough FAQ-style documentation is also available at the same web page. The open source taggers underlying EXTRACT are available at:

- ENVIRONMENTS: http://environments.hcmr.gr
- SPECIES and ORGANISMS: http://species.hcmr.gr
- TISSUES: http://tissues.jensenlab.org/
- DISEASES: http://diseases.jensenlab.org/

https://extract.hcmr.gr/

Architecture

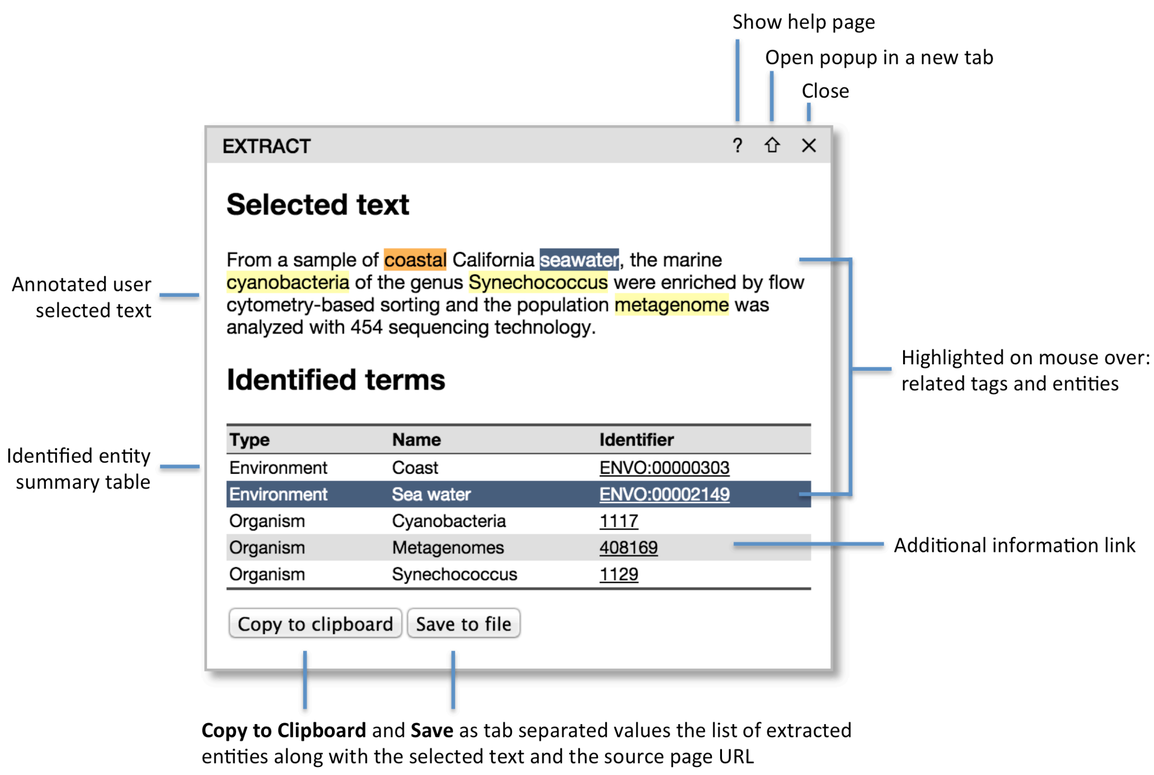


EXTRACT consists of three components: 1) a NER server, 2) the bookmarklet that allows users to submit text from a web page, and 3) a popup for inspecting terms and extracting annotations.

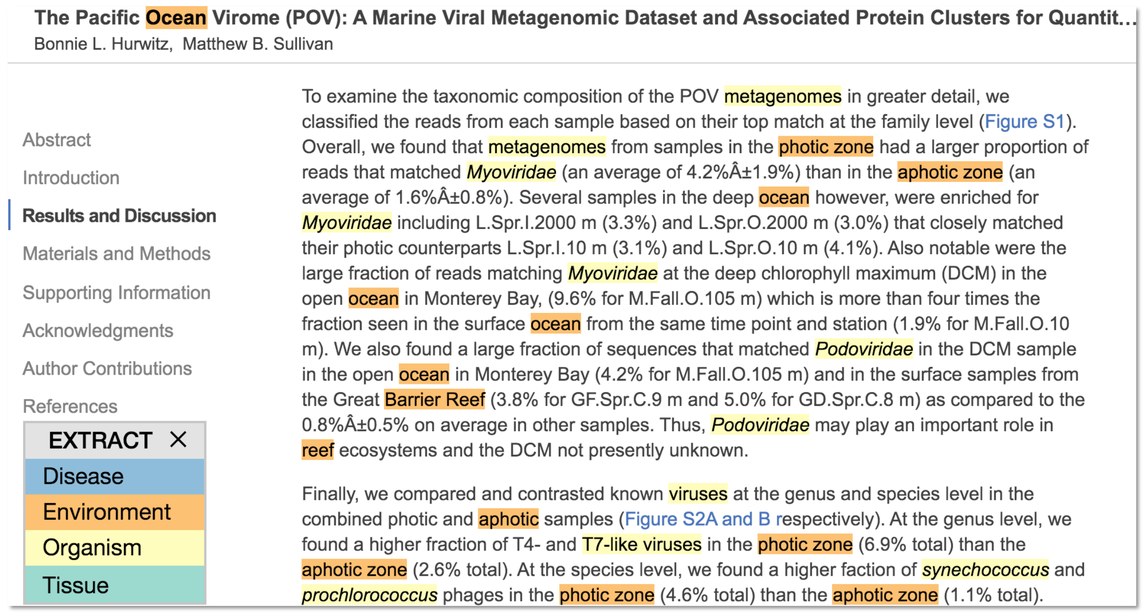
Acknowledgments

The authors thank Sune Pletscher-Frankild for developing the Python web framework that the EXTRACT server builds upon and Cecilia Arighi and Lynette Hirschman for testing EXTRACT, suggestions for improvements, and finding evaluators. This work was supported by the LifeWatchGreece Research Infra-structure [384676-94/GSRT/ NSRF(C&E)], the Micro B3 Project [287589], the Earth System Science and Environmental Management COST Action [ES1103] and the Novo Nordisk Foundation [NNF14CC0001]. EP received travel funding from Department of Energy [DE-SC0010838].

Features



The EXTRACT popup enables users to inspect the terms identified within a text selection and to collect annotations in tabular form. Users can visually inspect which words correspond to which entities. Two buttons allow users to either copy or save the annotations.



To identify relevant sections in a larger document, users can submit a full page for tagging. Identified terms are highlighted in different colors according to the type of entity, which makes it easy to spot relevant text segments. The example shows an excerpt of reference (7).

References

1. Yilmaz,P., Gilbert,J.A., Knight,R., *et al.* (2011) The genomic standards consortium: bringing standards to life for microbial ecology. ISME J., 5, 1565–1567.
2. Buttigieg,P.L., Morrison,N., Smith,B., *et al.* (2013) The environment ontology: contextualising biological and biomedical entities. J. Biomed. Semant., 4, 43.
3. Pafilis,E., Pletscher-Frankild,S., Schnetzer,J., *et al*. (2015) ENVIRONMENTS and EOL: identification of Environment Ontology terms in text and the annotation of the Encyclopedia of Life. *Bioinformatics*, **31**, 1872–1874.
4. Pafilis,E., Pletscher-Frankild,S.P., Fanini,L., *et al*. (2013) The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLOS ONE*, **8**, e65390.
5. Santos,A., Tsafou,K., Stolte,C., *et al*. (2015) Comprehensive comparison of large-scale tissue expression datasets. *PeerJ*, **3**, e1054.
6. Pletscher-Frankild,S., Pallejà,A., Tsafou,K., *et al*. (2015) DISEASES: Text mining and data integration of disease–gene associations. *Methods*, **74**, 83–89.
7. Hurwitz,B.L. and Sullivan,M.B. (2013) The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One*, **8**, e57355