PROCEEDINGS OF T D W G

Biodiversity Information Science and Standards 1: e20152
doi: 10.3897/tdwgproceedings.1.20152

OPEN ACCESS

Conference Abstract

EXTRACT 2.0: interactive identification of biological entities mentioned in text to assist database curation and knowledge extraction

Evangelos Pafilis[‡], Rūdolfs Bērziņš[§], Christos Arvanitidis[‡], Lars Juhl Jensen[§]

‡ Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Crete, Greece
§ Cellular Network Biology Group, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

Corresponding author: Evangelos Pafilis (pafilis@hcmr.gr)

Received: 09 Aug 2017 | Published: 10 Aug 2017

Citation: Pafilis E, Bērziņš R, Arvanitidis C, Jensen L (2017) EXTRACT 2.0: interactive identification of biological entities mentioned in text to assist database curation and knowledge extraction. Proceedings of TDWG 1: e20152. https://doi.org/10.3897/tdwgproceedings.1.20152

Abstract

Data curation is a process occurring in many aspects of biodiversity research, e.g. in digitization of specimen collections and extraction of species occurrences from the legacy literature. Data curation is always characterized by being time demanding and tedious. Gathering information on species and exposing it via search interfaces could be facilitated once phrases of interest have been recognized and the mentioned entities have been linked to community resources.

A curator can benefit from interactive systems that highlight biological entities in a document, indicating sections of interest, and map entities to corresponding database records/ontology terms, and offering an easy mechanism for extracting annotations in a structured form.

EXTRACT (https://extract.hcmr.gr, Pafilis et al. 2016) is a system that aims to address the above challenges. Its web User Interface is a *bookmarklet* that identifies genes/proteins, chemical compounds, organisms, environments, tissues, diseases, phenotypes and Gene Ontology terms mentioned in a web page and maps them to their corresponding database, ontology, and taxonomy entries. Two modes of operation are supported: a. extraction of

biological entities mentioned in user-selected piece of text, and b. full-page tagging. To easily collect extracted annotations, e.g. for use in an Excel spreadsheet, direct *Copy to clipboard* and *Save to file* (tab-delimited) are supported.

EXTRACT was originally developed specifically to facilitate metagenomic sample record annotation (Pafilis et al. 2016). As such it participated in the BioCreative V interactive annotation task. EXTRACT achieved one of the top scores in terms of usability and was evaluated to accelerate curation by 15–25% (Wang et al. 2016).

The latest version of EXTRACT (2.0, Pafilis et al. 2017) serves a much broader audience involving both biomedicine and biodiversity researchers and thus recognizes a wide range of entity types from many community resources:

- Organisms (NCBI Taxonomy, https://www.ncbi.nlm.nih.gov/taxonomy)
- Environments (Environment Ontology, Buttigieg et al. 2016)
- Diseases and phenotypes (Disease Ontology, Kibbe et al. 2014, and Mammalian Phenotype Ontology, Smith and Eppig 2012)
- Tissues and cell lines (Brenda Tissue Ontology, Placzek et al. 2016)
- Biological processes, molecular functions, and cellular components (Gene Ontology, Gene Ontology Consortium 2014)
- Protein-coding and non-coding RNA (ncRNA) genes from more than 2000 organisms (STRING (Szklarczyk et al. 2017) and RAIN (Junge et al. 2017))
- Small molecule compounds (STITCH (Szklarczyk et al. 2015))

In addition to curators benefitting from such a tool, knowledge-base developers can easily integrate the EXTRACT functionality into their own systems. To this end, we provide a robust and thoroughly documented Application Programming Interface (https://extract.hcmr.gr, FAQ section). EXTRACT can thus serve as a building block in large knowledge management pipelines, which also perform downstream tasks such as statistical entity association and association extraction, knowledge graph generation presenting the extracted associations, document indexing and information retrieval.

Such tasks lie at the core of the workshop this abstract has been submitted to and are pertinent to the TDWG 2017 theme, which is dedicated to the integration of species occurrence, gene, phenotype, and environment associations.

Keywords

text mining, named entity recogntion, interactive curation, metadata, genes proteins, organisms, environments

Presenting author

Evangelos Pafilis

Funding program

- The Novo Nordisk Foundation (NNF14CC0001)
- The LifeWatchGreece infrastructure (MIS 384676) (funded by the Greek Government under the General Secretariat of Research and Technology (GSRT), ESFRI Projects, National Strategic Reference Framework (NSRF)
- The EU BON project (http://www.eubon.eu), funded by the EU Framework Programme (FP7/2007-2013) under grant agreement No 308454

References

- Buttigieg PL, Pafilis E, Lewis SE, Schildhauer MP, Walls RL, Mungall CJ (2016) The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. Journal of biomedical semantics 7 (1): 57. https://doi.org/10.1186/s13326-016-0097-6
- Gene Ontology Consortium (2014) Gene Ontology Consortium: going forward. Nucleic acids research 43 (Database issue): 1049-1056. https://doi.org/10.1093/nar/gku1179
- Junge A, Refsgaard JC, Garde C, Pan X, Santos A, Alkan F, Anthon C, von Mering C, Workman CT, Jensen LJ, Gorodkin J (2017) RAIN: RNA-protein Association and Interaction Networks. Database: the journal of biological databases and curation 2017: baw167. https://doi.org/10.1093/database/baw167
- Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D, Parkinson H, Schriml LM (2014) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic acids research 43 (Database issue): 1071-1078. https://doi.org/10.1093/nar/gku1011
- Pafilis E, Bērziņš R, Jensen LJ (2017) EXTRACT 2.0: text-mining-assisted interactive annotation of biomedical named entities and ontology terms. biorxiv.org preprin https://doi.org/10.1101/111088
- Pafilis E, Buttigieg PL, Ferrell B, Pereira E, Schnetzer J, Arvanitidis C, Jensen LJ (2016) EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation. Database 2016: baw005. https://doi.org/10.1093/database/baw005
- Placzek S, Schomburg I, Chang A, Jeske L, Ulbrich M, Tillack J, Schomburg D (2016) BRENDA in 2017: new perspectives and new tools in BRENDA. Nucleic Acids Research 45: D380-D388. https://doi.org/10.1093/nar/gkw952
- Smith C, Eppig J (2012) The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. Mammalian Genome 23: 653-668. https://doi.org/10.1007/s00335-012-9421-3
- Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M (2015) STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic acids research 44 (D1): 380-384. https://doi.org/10.1093/nar/gkv1277

- Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen L, Mering Cv (2017) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Research 45: 362-368. https://doi.org/10.1093/nar/gkw937
- Wang Q, Abdul S, Almeida L, Ananiadou S, Balderas-Martínez Y, Batista-Navarro R, Campos D, Chilton L, Chou H, Contreras G, Cooper L, Dai H, Ferrell B, Fluck J, Gama-Castro S, George N, Gkoutos G, Irin A, Jensen L, Jimenez S, Jue T, Keseler I, Madan S, Matos S, McQuilton P, Milacic M, Mort M, Natarajan J, Pafilis E, Pereira E, Rao S, Rinaldi F, Rothfels K, Salgado D, Silva R, Singh O, Stefancsik R, Su C, Subramani S, Tadepally H, Tsaprouni L, Vasilevsky N, Wang X, Chatr-Aryamontri A, Laulederkind SF, Matis-Mitchell S, McEntyre J, Orchard S, Pundir S, Rodriguez-Esteban R, Auken KV, Lu Z, Schaeffer M, Wu C, Hirschman L, Arighi C (2016) Overview of the interactive task in BioCreative V. Database 2016: baw119. https://doi.org/10.1093/database/baw119