SPECIES

Organism Name Identification in the Scientific Literature

species@hcmr.gr

Evangelos Pafilis[1*], Sune Frankild[2*], Lucia Fanini[1+], Sarah Faulwetter[1+], Christina Pavloudi[1+]
Katerina Vasileiadou[1+], Christos Arvanitidis[1], Lars Juhl Jensen[2]

*: shared, +: in alphabetic order, correspondence to: pafilis@hcmr.gr, lars.juhl.jensen@cpr.ku.dk

[1] Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine Research (HCMR), 71003 Heraklion, Crete, Greece
[2] Disease Systems Biology, Novo Nordisk Foundation for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

Motivation

The identification of species mentions in scientific documents can have an beneficial effect to a wide range of scientific projects. Improving protein mention recognition and protein-protein association extraction (Gerner *et al.*, 2010) are two such cases.

The ability to classify documents according to the organisms they mention can also support document indexing, retrieval and clustering systems (Gerner *et al.*, 2010).

The association of taxonomic mentions in a cross-scientific-domain collection of documents can support the *synthesis* of biological knowledge across several levels of its organisation (such as the molecular, the organismal and ecosystem one).

Databases, such as the NCBI Taxonomy [1], that act as a nexus among species (and their lineage), genetic sequences, relevant population variation information, and other related pieces of knowledge, are facilitating such efforts.

Functionality

SPECIES is a command line application capable of identifying taxonomic mentions at the level of species in documents and mapping them to corresponding NCBI Taxonomy [1] database entries.

Given a folder with plain text files, SPECIES based on its taxonomic name and synonym dictionary reports the taxonomic mentions (start, end position in each document), the detected term and the corresponding NCBI Taxonomy database record identifier (Fig 1).

Besides binomials following the Linnaean naming convention, recognised taxonomic mentions include acronyms, common names and abbreviations, as well as misspellings and the rest of the naming types supported by the NCBI Taxonomy [1].

Additionally, SPECIES is able to: consider the different ways authors may write an organism, such as “zebrafish, zebra-fish, zebra fish”, and detect abbreviated forms of scientific names (e.g. *C. sativa*) and disambiguate them to their full-length co-mentioned species name (if any).

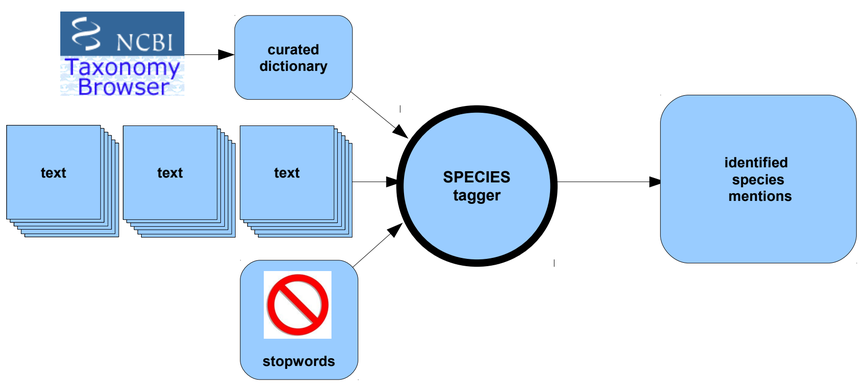

Figure 1: Document annotation process and underlying components. The only parameter required by SPECIES is the path to a folder with plain text documents

Methods

SPECIES follows a dictionary lookup approach to identify taxonomic mentions. A curated dictionary, based on the taxon name information available in the NCBI Taxonomy, orthographically expanded and combined with run-time term matching constitute its core (Fig 1).

A manually curated list of common English words that unfortunately happen to be species names as well ensures that these words are excluded from the analysis (Fig 1).

Availability

Both the SPECIES and the S800 corpus are open source and will be freely available at http://species.hcmr.gr

References – Web Resources

- Gerner M., Nendic, G. and Bergman, C. M. (2010) *LINNAEUS: a species name identification system for biomedical literature*. BMC Bioinformatics 11:85

1. NCBI Taxonomy Database: http://www.ncbi.nlm.nih.gov/taxonomy
2. Linnaeus System: http://linnaeus.sourceforge.net/
3. PubMed Resource: http://www.ncbi.nlm.nih.gov/pubmed/
4. Environment Ontology: http://environmentontology.org
5. Wordle: http://www.wordle.net

Performance

To evaluate SPECIES it has been compared against Linnaeus (Gerner *et al.*, 2010, [2]), a commonly used species name recognition system. The two systems were compared both in terms of tagging accuracy (Table 1) and in terms of tagging speed (Fig 2). These are preliminary results and subject to further improvement upon software release.

A. Linnaeus Corpus (L100)		
Software	LINNAEUS	SPECIES
Precision	75.7%	82.4%
Recall	97.7%	97.3%
F-score	85.3%	89.2%

B. Species Corpus (S800)		
Software	LINNAEUS	SPECIES
Precision	79.7%	78.3%
Recall	90.3%	90.3%
F-score	84.7%	83.8%

Table 1: SPECIES and LINNAEUS accuracy evaluated at the document level against the manually annotated corpora of **A.** Linnaeus (L100, full-text-based) and **B.** the novel Species corpus (S800, abstract-based)
(Precision =TP/ TP+FP, Recall = TP/TP+FN,
F-score = 2*precision*recall/(precision+recall))

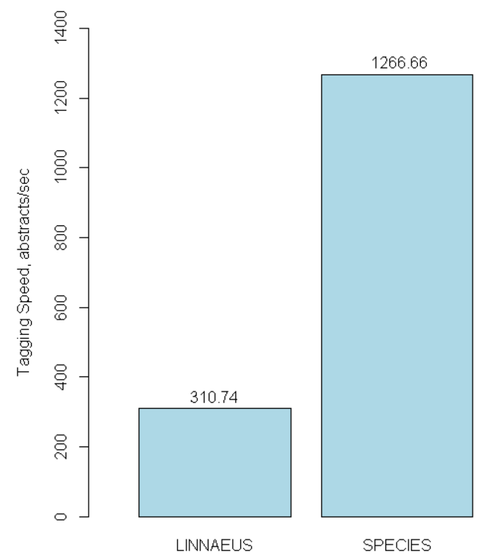

Figure 2: SPECIES and LINNAEUS tagging speed measured against of set of ~536K PubMed [3] abstracts
(experiment conducted using a single-thread run on an Intel 2,27GHz, 24 GB RAM)

S800 Corpus

To thoroughly evaluate the accuracy of SPECIES a novel abstract-based corpus has been manually annotated (Table 1). The annotation comprised the identification of organism mentions and their mapping to the corresponding NCBI Taxonomy identifier in 800 PubMed [3] abstracts.

To better assess the performance of organism name identification in journals from different scientific fields the abstracts were selected from representative journals in: Bacteriology, Botanics, Entomology, Medicine, Mycology, Protistology, Virology and Zoology.

For interested researchers S800 has not only been annotated at the species level; higher taxa mentions (such as genera, families and orders) have also been considered.

Future Directions

In this experiment SPECIES focused on identifying taxonomic mentions at the species level. Its simple architecture (Fig 1) ensures its increased extensibility. SPECIES could be extended e.g. to identify genera and other higher taxa mentions as well as other types of entities.

To better satisfy the aims of biodiversity research we are already experimenting with the identification of environment descriptive terms (Fig 3) based on the Environmental Ontology [4]. Functional traits and life-cycle stages are other entity types of interest.

Conducting co-mentioning analysis among these entity types is a mid- to long-term objective.

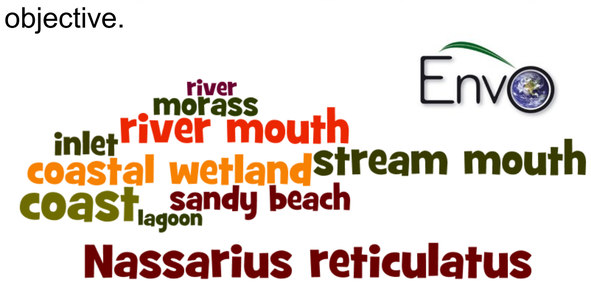

Figure 3: A tag cloud displaying the most frequently co-mentioned environment descriptive terms in PubMed with a small European snail *Nassarius reticulatus*. The term font-size corresponds to the co-mentioning frequency (figure created using [5], experiment conducted on November 2011)

Acknowledgements

EP and LF have been funded by the MARBIGEN EU FP7 REGPOT Project (Reference: 264089). LJJ and SuF by the Novo Nordisk Foundation Center for Protein Research. A visit of EP in NNFCPR has been funded by an EMBO Short Term Fellowship (356-2011).

species@hcmr.gr