Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4509–4513
Marseille, 11–16 May 2020

An Experiment in Annotating Animal Species Names from ISTEX Resources

Sabine Barreaux, Dominique Besagni
INIST – CNRS 2, rue Jean Zay 54519 Vandœuvre-lès-Nancy
{sabine.barreaux, dominique.besagni}@inist.fr

Abstract

To exploit scientific publications from global research for TDM purposes, the ISTEX platform enriched its data with value-added information to ease access to its full-text documents. We built an experiment to explore new enrichment possibilities in documents focussing on scientific named entities recognition which could be integrated into ISTEX resources. This led to testing two detection tools for animal species names in a corpus of 100 documents in zoology. This makes it possible to provide the French scientific community with an annotated reference corpus available for use to measure these tools' performance.

Keywords: named entity recognition, zoology, ISTEX, Text and Data Mining

1. Introduction

The availability of an ever-increasing volume of scientific publications requires a number of means to automatically mine into large sets of documents and extract implicit knowledge and correlations to be used in research work. But before being able to carry out this text search, one needs to address the questions of accessing these documents in full text and of pre-processing them, the quality of the documents being often uneven.

The ISTEX platform (Initiative d'Excellence en Information Scientifique et Technique)[1] offers a solution to the French higher education and research community by providing access to retrospective collections of scientific literature in all disciplines (Colcanap, 2013). This vast multidisciplinary and multilingual collection contains more than 23 million scientific publications and is constantly enriched with information to improve the text quality to optimize machine processing, i.e. categorization of documents by scientific fields, extraction and structuring of bibliographic references, detection of named entities, structuring of the full text in XML-TEI from PDF (Collignon and Cuxac, 2017).

This platform provides massive value-added downloads and represents an unparalleled resource for text mining research (Cuxac and Thouvenin, 2017).

In this article, we will describe an experiment led to continue the work on ISTEX data enriching by testing two tools for the detection of a new type of named entities, animal species names. We will trace the steps of building and annotating a reference corpus in zoology from ISTEX resources and compare the performance of the two tools tested: entity-fishing[2] and IRC3sp. We will conclude by considering the possibilities of using these tools in ISTEX.

2. Use Cases

Users can run queries in the ISTEX platform on traditional bibliographic fields, but also on fields that exploit the value-added information injected into documents. Among the enhancements to ISTEX documents, we will focus on named entities. This enrichment was provided by implementing the UNITEX CasSys tool at INIST[3] in collaboration with LIFAT[4]. At present, this tool has made it possible to detect the named entities contained in nearly 16 million documents by categorizing them using a set of labels based on those used in the MUC exercises. They were supplemented on this occasion by more specific labels in response to specific needs for scientific and technical information (Maurel *et al.*, 2019):

- person's name
- geographical place name
- administrative place name
- date
- organization name
- funding organizations and funded projects
- provider organization of resources
- pointer to bibliographic reference
- bibliographical reference
- URL

However, the task of recognizing named entities has emerged during MUC exercises as a task in its own right in information extraction through the detection of person, place, organisation names, or temporal expressions and numerical expressions in unstructured texts. This recognition task has become much more diversified with the inclusion of entity typologies pertaining to speciality fields. Yadav *et al.* (2018) and Nadeau *et al.* (2009) especially give a complete inventory in their studies. In particular, biological entities (proteins, DNA, cell lines, cell types, etc.) are detected and annotated in the GENIA corpus, a reference corpus collated from Medline records (Ohta *et al.* 2002). Drug names were searched in biomedical texts during the 2013 SemEval campaign (Segura Bedmar *et al.* 2013) and bacterial taxon names were identified in scientific web pages during the BioNLP evaluation tours (Bossy *et al.* 2011, 2013).

To move further in the recognition of named entities in ISTEX, and possibly complement the types of entities offered by ISTEX for information retrieval, we wanted to test a tool to detect scientific named entities independently of language and domain, and using the full texts available in ISTEX in PDF or text form.

[1] Excellence initiative in scientific and technical information: https:www.istex.fr
[2] http://nerd.huma-num.fr/nerd/
[3] Institute for Scientific and Technical Information, Vandœuvre-lès-Nancy, France, in charge of hosting the ISTEX platform
[4] Laboratory of Fundamental and Applied Computer Science of Tours, France

We focused on the entity-fishing tool that performs this task by automatically identifying and disambiguating Wikidata entities in PDF documents and multilingual texts. Entity-fishing is developed by Science Miner[5], a company that has already contributed to enrich ISTEX with the GROBID (GeneRation Of BIbliographic Data) tool for the detection and structuring of bibliographic references contained in full texts (Lopez, 2009). We wanted to resume our collaboration with this company by testing another of their tools. In another project with the French National Museum of Natural History, we focused more particularly on a version of the entity-fishing tool that automatically identifies animal and plant species names in full texts (Lopez, 2017).
This is why we built a corpus of 31,778 ISTEX zoology documents containing animal species names.
From this corpus, we isolated a subset of 100 documents that served as a reference sample to evaluate the performance of entity-fishing and another tool developed at the Inist, IRC3sp.

3. Constitution of the Reference Corpus

After a needs analysis phase with the team in charge of testing the tool, we established the following criteria to define the content of the corpus:

- Each document must contain at least one species name from the kingdom Animalia
- Microorganisms and fungi are to be avoided
- A wide variety of animal species is available
- The name of the species must be in Latin
- Language of the document: English
- Publication dates: from 1950 onwards
- File type: medium to high quality full text PDF (version from 1.2 + quality score from 3.0) and no image PDF
- Must contain abstracts

3.1 Selection of the Complete Corpus

A first step was to find all the documents corresponding to the defined criteria in ISTEX.
We built 11 requests, each one being the transcription of these criteria into the ISTEX API query language, combined with zoological terms from the following major groups: arthropods, amphibians, echinoderms, sponges, insects, mammals, molluscs, birds, fishes, reptiles, and worms.

https://api.istex.fr/document/?q=abstract:((species OR genus) AND (arthropod arachnid* acari* centiped* crustac* /spiders?/ /mites?/ /scorpions?/ /barnacles?/ /crabs?/ /lobsters?/ /shrimps?/)) AND language:"eng" AND qualityIndicators.pdfVersion:[1.2 TO *] AND qualityIndicators.score:[3.0 TO *] AND (publicationDate:[1950 TO *] OR copyrightDate:[1950 TO *]) NOT (/insects?/ entomolog* fungu* bacteria* /microorganisms?/ /viruse?s?/ neuro* botan* protozoa*)*

Figure 1: Example of a requests for the "arthropod" group

The concatenation of the results of these 11 requests resulted in the retrieval of 31,778 documents.

3.2 Selection of a Reference Corpus

To find species names in a corpus, it was necessary to have well-structured XML texts. Although these texts in the ISTEX databases are converted into the same TEI format now, at the time of our experiment we had only the XML files supplied by the different publishers with different DTDs. So we decided to work with only one set of documents from the publisher with more structured documents. As shown in table 1, Wiley has by far the largest number of XML files.

Publisher	Total	Structured
Wiley	16 129	8 401
Elsevier	9 732	1 342
Brill	429	426
Oxford University Press	344	282
Royal Society of Chemistry	115	51
Institute Of Physics	10	9
Emerald	13	6
Nature	5	5
British Medical Journal	10	1
De Gruyter	298	0
Sage	42	0
Springer	4 651	0

Table 1: Number of documents and structured documents per publisher in the original corpus

We randomly selected 100 documents from the Wiley subset of documents, making sure that all zoological groups were represented.

4. Reference Corpus Annotation

4.1 Species Names

The species, or taxon, is the basic level in the classification of living organisms. But common names, also known as vernacular names, are often ambiguous. So, since the publication by Carl Linnaeus of *Systema naturæ* (10th edition) in 1758, species have been given a two-part Latin (or Latinised) name made of the generic name for the genus to which the species belongs and the specific name for the species within the genus. By convention, the generic name is capitalised and the specific name is in lowercase. Also, a species name is in italics when printed and underscored when hand-written. When a species name is used repeatedly in a document, the generic name must be written in full the first time, but after that it may be abbreviated to its initial followed by a period, e.g. "*C. lupus*" for "*Canis lupus*". If other species of the same genus are cited, their generic name may also be abbreviated as long as it appears in full before, e.g. "*Canis lupus*, *C. aureus*, *C. latrans*".

4.2 Annotation methodology

4.2.1 Automatic Annotation

The method we used, called "T+rex" for "**T**ypography + **re**gular **ex**pression", first looks for the XML tag indicating a text in italics, then checks with a regular expression that the embedded text is compatible with a species name, either in its long form or in its abbreviated one. Any acceptable

[5] http://science-miner.com/

term is first compared to a list of Latin expressions (from Wikipedia[6]) like "*in vitro*" or "*ad libitum*" that may be capitalized at the beginning of a sentence and be confused with a species name. Based on the text structure, the programme limits its search for species name to its title, abstract and body, avoiding the bibliography and possible annexes.

4.2.2 Human Validation

The list of species obtained by T+rex is then compared to our resource extracted from different databases: Catalogue of Life[7], The Plant List[8] and AlgaeBase[9]. Their respective and overlapping contributions are 93.7%, 14.56% and 1.56%. The unmatched terms are checked manually by our in-house expert in the field to see if we encountered a valid but yet unreferenced name, a typing error or just an italicised expression without interest.

As one of the tested tools, i.e. entity-fishing, searches for species regardless of the kingdom they belong to, we consider all species names to avoid a bias in the precision measure. As seen in table 2, we obtained 1351 different species names with 1464 occurrences, one occurrence being a species appearing at least once in one document.

Kingdom	Nb. of species	Nb. of occurrences
Animalia	1 250	1 362
Plantae	97	97
Bacteria	2	3
Chromista	1	1
Protozoa	1	1
Total	1 351	1 464

Table 2: Distribution of species by kingdom

5. Test of annotation tools

5.1 Annotation Tools

5.1.1 Entity-fishing[10]

Entity-fishing was designed by the Science Miner company with a contribution from INRIA Paris to perform semantic content enrichment of PDF documents. Based on the document structure identified by GROBID which is a state-of-the-art tool for structuring the body of a scientific paper from a PDF input, it proposes entity recognition and disambiguation using Wikidata as a resource. The structuration "avoid labelling bibliographical callout, running foot and head notes, figure content, and identify the useful areas of the text (header, paragraphs, captions, etc.), handling multiple columns, hyphen, etc."[11] The disambiguation is done by supervised machine learning trained on pages from Wikipedia. It works at document level, for example a PDF with layout positioning and structure-aware annotations. It is also possible to apply filters based on Wikidata properties and values, allowing to create specialised entity identification and extraction as taxon entities.

5.1.2 IRC3sp

IRC3[12] is a recognition tool based on a pattern-matching method. It was first developed to find chemical and enzyme names from an authoritative list in scientific articles (Royauté *et al.*, 2003; Royauté *et al.*, 2004). Since these names have their own syntax that includes punctuation signs, hyphens, brackets and quotation marks (e.g. "*3'(2'),5'-Bisphosphate nucleotidase*"), processing them with usual NLP tools is problematic.

In the case of species names, one additional problem is abbreviated forms, because different species may have the same abbreviation. For example, the common carp "*Cyprinus carpio*" and the river carpsucker "*Carpiodes carpio*" have the same abbreviation "*C. carpio*". So a simple list of abbreviations with the corresponding long form is just not possible. We developed a variant of our programme, named IRC3sp, to solve the problem. In a first step, for each document, the tool searches for the long form of species names. Then, from the list of species names, the tool extracts the list of generic names, as "*Canis*" from "*Canis lupus*", and generates a list of possible abbreviations: "*C. aureaus*", "*C. latrans*", "*C. lupus*", etc. Next, the document is processed again using the list of names initially found and the generated list of abbreviations. This is how we obtain a list of names appearing in the same order than in the document. This is very important to remove the remaining ambiguities because if several names match an abbreviation, we select the species belonging to the most recent genus cited in full. As IRC3sp works on full text files, we extracted the text from the XML files used by T+rex keeping only the title, the abstract and the body of the text.

5.2 Evaluation

5.2.1 Evaluation Procedure

For IRC3sp, the evaluation is pretty straightforward. It is a simple comparison with the list of species names from T+rex.

With entity-fishing, we obtain not only the species names but also all taxa from species to kingdom, including family, order, class and even sublevels. For each input file, we have 3 output files: a JSON file (very verbose with a lot of information from Wikidata), a CSV file and a TEI file (made for the ISTEX database). The CSV file contains for each entry the observed term in the text, the preferred term in Wikidata, the taxon and its rank, the number of occurrences and the Wikidata identifier. After filtering the results to keep the rank "Species", the original terms from the documents are categorised into scientific names (full, abbreviated or partial) or vernacular names (single-word or multi-word). Then, we check if the species name inferred from that original term is correct. Finally, we compare with the species names obtained by T+rex knowing that entity-fishing gives the name currently used as a preferred term in Wikidata while T+rex gives the name as it appears in the document. In some older documents, we may find some obsolete names. For example, the polar bear "*Ursus maritimus*" used to be called "*Thalarctos maritimus*". And

[6] https://en.wikipedia.org/wiki/List_of_Latin_phrases_(full)
[7] http://www.catalogueoflife.org/ (May 2019)
[8] http://www.theplantlist.org/ (July 2017)
[9] https://www.algaebase.org/ (October 2017)
[10] https://github.com/kermitt2/entity-fishing
[11] https://nerd.readthedocs.io/en/latest/
[12] https://git.istex.fr/scodex/IRC3

for some species, there is still disagreement amongst experts on what the correct name should be.

5.2.2 Results

We applied the classic measures of precision and recall, as well as the F-measure on both tools as seen in table 2.

	IRC3sp	entity-fishing
Expected	1 464	1 464
Found	1 384	994
Correct	1 383	824
Precision (%)	99.9	82.9
Recall (%)	94.5	56.3
F-measure (%)	97.1	67.1

Table 2: Precision, recall and F-measure of both tools

IRC3sp has great precision and good recall while entity-fishing scores lower values. As previously mentioned for entity-fishing, we categorised the original terms found in the documents and we determined if the corresponding species were correct. Table 3 shows the effectiveness of each category.

Original terms	Total	Correct	%
Full scientific name	804	803	99.9
Abbreviated scientific name	90	72	80.0
Partial scientific name	47	17	36.2
Higher taxon (e.g. family)	2	2	100.0
Single word vernacular name	188	112	59.6
Multi-word vernacular name	448	409	91.3

Table 3: Rate of success for each category of terms with entity-fishing

As we can see, full scientific names have a very high rate of success, albeit imperfect because of an error in Wikidata. The involved species "*Melolontha hippocastani*" has the common name "Cockchafer" as preferred name and that name has "*Melolontha melolontha*" as scientific name. The vernacular names should have been respectively "European forest cockchafer" and "common European cockchafer". Abbreviated scientific names have a lower score because the inferred species belongs sometimes to a genus not even cited in the document. Multi-word vernacular names achieve a much better rate and the mistakes usually happen when just part of the name is found by entity-fishing, for example "Guinean devil ray" instead of "lesser Guinean devil ray". Single-word vernacular names are more of a problem because many are ambiguous and partial scientific names, generic or specific names, are more often wrong than not.

So we decided to keep only species inferred from full or abbreviated scientific names and multi-word vernacular names and recalculate precision and recall on that "clean" dataset. As we can see in table 4, the number of correct names is about the same while the precision has notably increased.

	raw data	"clean" data
Expected	1 464	1 464
Found	994	894
Correct	824	820
Precision (%)	82.9	91.7
Recall (%)	56.3	56.0
F-measure (%)	67.1	69.6

Table 4: Precision, recall and F-measure for entity-fishing

5.2.3 Comparison

IRC3sp shows a great success rate with few errors. Actually, the recall is limited by the completeness of its resources. For example, Catalogue of Life boasts of having 92 % of all species names as of May 2019. Plus, some of its contributing databases do not record every ancient name now obsolete and that is a problem when processing old documents from an archive like ISTEX.

Entity-fishing has a lower success rate with more errors, but it is a work in progress and there is scope for improvement. For example, species names were extracted from bibliographic references in 13 documents, so the text segmentation worked fine on most documents but can still progress. Also, guessing the species name from just a generic or specific name is to be avoided. Likewise, single-word vernacular names can sometimes be very discriminating, but with an overall success rate of only 59.6%, they should not be used either. Nonetheless, it is interesting to note that among the 820 correct species names found by entity-fishing, in 12 instances entity-fishing found the correct name while IRC3sp could not. In some cases, it was because the name was not in IRC3sp resources and in the other cases because it used the vernacular name while the scientific name was misspelled.

6. Results

The evaluation carried out in this experiment highlighted the good performance of IRC3sp with an F-measure of 97.1% and a recall of 94.5%. This is the result of the constant improvements we made to deal with the specific problem of abbreviated forms in species name detection and to build a resource from reference databases as comprehensive as possible. The results are lower for entity-fishing with an F-measure of 69.6% and a recall of 56% but seem promising because some improvements could be made during the step of full-text structuring and the step of entity resolution from textual mention. For the moment, we keep in mind the good precision with a score of 91.7% which will be relevant for a future step of our enrichment process. It may be noted that the comparison is not perfect because entity-fishing works on PDF files while IRC3sp works with text files we extracted from well-structured XML files. We should test entity-fishing with the same text files used for IRC3sp for a better comparison.

This experimentation also led to build a reference corpus that can be reused for other evaluations. This corpus is named 'Animalia 100' and is available at the following address: https://systematique-animallrec.corpus.istex.fr/. This website makes it possible to explore the content of the corpus using charts representing different views on documents, including a tree graph of the systematic classification of the different animal species names found by the three tools. Users can also download the corresponding documents using the associated ISTEX-DL service[13]. The list of detected species names is also

[13] As access to ISTEX documents is subject to authentication, they cannot be distributed as such to the entire international scientific community.

available on this website as a JSON file containing the extracted species names with the metadata of the texts in which they were found.

7. Future Work and Prospects

This experiment led us to present a reference corpus to the French scientific community on a dedicated website and to evaluate the performance of two scientific named entity detection tools against a reference list found by a third detection tool.

The next step of this work will consist in integrating the species names detected by these tools as an enrichment into the ISTEX platform. The detection results of scientific named entities will be transformed into the TEI-Standoff XML pivot format used for the storage and distribution of enrichments in ISTEX.

Meanwhile, work is underway on the treatment of the complete zoological corpus. The recognition of animal taxa has already been carried out on the entire corpus of 31,778 documents and has made it possible to detect more than 60,000 names of animal species. A similar work is being carried out for the recognition of plant taxa in a corpus of 51,480 botanical documents from ISTEX.

These names of animal and plant species detected using entity-fishing and IRC3sp in these larger corpora will be injected into the platform in XML TEI-Standoff format and will be accessible for use in queries by any user in the long run.

Once implemented into ISTEX, these new enhancements can be linked to ontologies, for example via the Wikidata identifiers provided by entity-fishing, and fed into the ISTEX Triplestore[14] to take advantage of the new information search capabilities offered by the Semantic Web (Ee, 2019).

8. Bibliographical References

Bossy R., Jourde J. Bessières P., van de Guchte M., and Nédellec C. (2011). BioNLP shared task 2011 – Bacteria Biotope. In Proceedings of the BioNLP Shared Task 2011 Workshop, pages 56–64, Portland, Oregon, USA, June. Association for Computational Linguistics (ACL).

Bossy R., Golik W., Ratkovic Z., Bessières P., and Nédellec C. (2013). BioNLP shared task 2013–an overview of the bacteria biotope task. In Proceedings of the BioNLP Shared Task 2013 Workshop, pages 161–169, Sofia, Bulgaria, August. Association for Computational Linguistics (ACL).

Colcanap. G. (2013). Istex : un gisement documentaire producteur de connaissances: De l'idée de licences nationales à la construction d'un projet. *Bulletin des bibliothèques de France*, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB), 58 (1): 66-71. hal-01827544

Collignon, A. and Cuxac, P. (2017). ISTEX : des enrichissements au Web de données. *I2D– Information, données & documents*, 4(54): 8-15.

Cuxac, P. and Thouvenin, N. (2017). Archives numériques et fouille de textes : le projet ISTEX. In Actes de la conférence Extraction et Gestion des connaissances (EGC), Atelier Fouille de textes, pages 43-51, Grenoble, France, janvier.

Ee, M. H. (2019) Mining Text, Linking Entities – NLB's Journey. Paper presented at: IFLA WLIC 2019 - Athens, Greece - Libraries: dialogue for change in Session 114 - Knowledge Management with Information Technology and Big Data.

Ehrmann, M. (2008). Les Entités Nommées, de la linguistique au TAL: Statut théorique et méthodes de désambiguïsation. PhD Thesis. Paris Diderot University, 2008. Français. tel-01639190

Lopez, P. (2009). Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In: Agosti M., Borbinha J., Kapidakis S., Papatheodorou C., Tsakonas G. (eds) Research and Advanced Technology for Digital Libraries. ECDL 2009. Lecture Notes in Computer Science, vol 5714. Springer, Berlin, Heidelberg

Lopez P. (2017). Entity-fishing: An open source tool for fishing Wikidata entities in text and PDF documents. Proceedings of the first conference dedicated to the wikidata community (WikiDataCon 2017), Berlin, Germany, October.

Maurel D., Morale E., Thouvenin N., Ringot P., and Turri A. (2019). ISTEX: A database of twenty million scientific papers with a mining tool which uses named entities. *Information*, 10 (5): 178.

Ohta, T., Tateisi, Y., Kim, J.D. (2002). The GENIA Corpus: An Annotated Research Abstract Corpus in Molecular Biology Domain. In Proceedings of the second international conference on Human Language Technology Research, pages 82-86, San Diego, USA, March. Association for Computational Linguistics (ACL).

Nadeau D. and Sekine S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1):3–26.

Royauté J., François C., Zasadzinski A., Besagni D., Dessen P., Le Minor S. and Maunoury M.-T. (2003). Mining corpora of texts on genes involved in thyroid cancers: a bioinformatic text mining and clustering process. In Proceedings of the European Conference on Computational Biology (ECCB), pp. 77-78.

Royauté J., François C., Zasadzinski A., Besagni D., Dessen P., Le Minor S. and Maunoury M.-T. (2004). Approche terminologique et infométrique de fouille de données textuelles dans un corpus sur la génétique des cancers de la thyroïde. In EGC 2004, vol. RNTI-E-2, pp. 465-476.

Segura Bedmar I., Martinez P., and Herrero Zazo M. (2013). Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (DDI Extraction 2013). Association for Computational Linguistics (ACL).

Yadav, V. and Bethard, S.A. (2018). Survey on recent advances in named entity recognition from deep learning models. In Proceedings of the 27th International Conference on Computational Linguistics, pp. 2145–2158, Santa Fe, NM, USA, August.

[14] https://data.istex.fr/triplestore/sparql/