BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 11 2007, pages 1434–1436
doi:10.1093/bioinformatics/btm109

Data and text mining

uBioRSS: Tracking taxonomic literature using RSS

Patrick R. Leary[1], David P. Remsen[2], Catherine N. Norton[1], David J. Patterson[1] and Indra Neil Sarkar[1,*]

[1]MBL Informatics, Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543, USA and
[2]Global Biodiversity Information Facility Secretariat, Universitetsparken 15, DK-2100 Copenhagen, Denmark

Received on February 2, 2007; revised and accepted on March 13, 2007

Advance Access publication March 28, 2007

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Summary: Web content syndication through standard formats such as RSS and ATOM has become an increasingly popular mechanism for publishers, news sources and blogs to disseminate regularly updated content. These standardized syndication formats deliver content directly to the subscriber, allowing them to locally aggregate content from a variety of sources instead of having to find the information on multiple websites. The uBioRSS application is a 'taxonomically intelligent' service customized for the biological sciences. It aggregates syndicated content from academic publishers and science news feeds, and then uses a taxonomic Named Entity Recognition algorithm to identify and index taxonomic names within those data streams. The resulting name index is cross-referenced to current global taxonomic datasets to provide context for browsing the publications by taxonomic group. This process, called taxonomic indexing, draws upon services developed specifically for biological sciences, collectively referred to as 'taxonomic intelligence'. Such value-added enhancements can provide biologists with accelerated and improved access to current biological content.

Availability: http://names.ubio.org/rss/

Contact: sarkar@mbl.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Really simple syndication (RSS) refers to a family of 'web feed' formats that are used to serve content that changes on a regular basis such as the latest scientific journal articles, news headlines and blog entries. Traditionally, content provided by web servers is linked to a relatively stable address to provide a consistent identifier for content access. Many publishers, for example, use unique addresses, such as Uniform Resource Locators, Handles or Digital Object Identifiers, to identify published articles. Individual volumes may also be provided with relatively persistent and unique addresses. This strategy works by linking an address to a persistent data object but requires new addresses be assigned to new iterations of the object.

*To whom correspondence should be addressed.

RSS inverts this model by providing a single address that references changing content represented by a standardized set of metadata formats, so they can be collated and consistently parsed. The address is represented as a URL just like any other web address. The primary distinguishing feature of RSS is that it represents a standardized way for content providers to syndicate dynamic content. This enables RSS clients, known as 'feed readers' or 'aggregators' to access one or multiple RSS addresses (or 'feeds') in a standardized way. The actual data that the address resolves to varies slightly across different versions of RSS. RSS 1.0 is based on the resource description framework (RDF) and offers a relatively rich, though more complex mechanism for formatting feed content. RSS 2.0 is a simpler XML format. Both RSS 1.0 and 2.0 are relatively stable and in common use.

The uBio Project, located at the MBLWHOI library in Woods Hole, Massachusetts, is an informatics initiative focused on addressing access impediments within biological information resources that are inherent to organism names. Built on the premise that scientific names are a ubiquitous element in any recorded information that references a species, the uBio project has been developing taxonomic information retrieval systems (Patterson *et al.*, 2006). uBio maintains a set of indexes that catalog taxon names as they appear within scientific literature and authoritative nomenclatural and taxonomic lists. The uBio name index, NameBank, represents a collection of over 9.5 million scientific and vernacular name forms. NameBank is cross-referenced with a secondary index, Classification Bank, which catalogs authoritative classifications and species checklists. Building on these two indexes, we have developed an RSS aggregation application, uBioRSS, to facilitate the monitoring of taxonomic literature.

2 METHODS

The uBioRSS application synergizes content syndication, taxonomic Named Entity Recognition and the uBio indices NameBank and ClassificationBank. This application can be described as a 'taxonomically intelligent' RSS aggregator with biologically relevant organization and navigation structures. Hundreds of RSS feeds representing hundreds of scientific journals and science sources are scanned daily for new content that references scientific taxa. The collective index of taxa is then cross-referenced to taxonomies in ClassificationBank. The set of matching names can thus be organized

and accessed by traversing any of the matching classifications. For example, using the 2006 Annual Checklist of the Catalogue of Life Partnership (http://www.catalogueoflife.org), a user can navigate from all articles (51 000 references) to those that reference animals (25 000 references), through to the insects (8000 references) to the flies (2600 references). With four mouse clicks, the user can filter the entire corpus of 51 000 references from over 500 published sources down to the 2600 resources that reference any of the 100 000 species of flies. This kind of biological filtering can be an enormous time saver, and might be intractable if not through applications such as uBioRSS.

Scientific names are identified in natural language text using a class of Named Entity Recognition (NER) algorithms termed Taxonomic Name Recognition (TNR) (Konig *et al*., 2005). TNR is a variant form of NER that is designed to identify scientific names in text. TaxonFinder is our TNR tool that has the ability to recognize scientific name strings known to NameBank. Based on specialized rule sets and the polynomial nature of taxon names, TaxonFinder can also discover new name combinations. A peer review process examines these new names and eventually adds them to the NameBank index. Subsequent updates of the scientific name lexicons incorporate any new NameBank entries, and the TaxonFinder algorithm thus will be able to positively identify more scientific names.

Users of the uBioRSS application can create personalized profiles to narrow the breadth of sources from which the content is aggregated and the taxonomic coverage of the content. Users may choose to only view articles from specific academic publishers for example. Alternatively, users can choose to display content that refers to organisms from a particular taxonomic group, or from a regional or thematic list, such as the IUCN Red List of Threatened Species (http://www.iucnredlist.org). For example, an ant specialist may choose to only view content that references ants. This sample user could select the entire family of ants, or they may provide a custom list of species they wish to monitor. They may further refine the set of ant literature by selecting to show ant literature from a particular data provider. Each user can receive updates on new content that matches their criteria by subscribing to a custom uBioRSS feed, or they may receive weekly email updates.

Customized uBioRSS feeds can be used in a variety of ways and embedded into external applications. The principal demand for custom uBioRSS feeds comes from scientists wanting to monitor literature on species from their domain of research. The MBLWHOI Library created a custom profile to track published articles citing significant species in marine biology and oceanography research. The library homepage displays recent articles relevant to these fields of research (http://www.mblwhoilibrary.org).

Some websites with a biological focus may wish to receive dynamic updates on literature referencing organisms in their domain. AntWeb (http://www.antweb.org) is a website about ants from the California Academy of Science. AntWeb has created a custom uBioRSS feed to monitor each of the ant species they detail on their website. They now have an 'In the News' section of their homepage dedicated to displaying the latest ant literature, all of which come from their custom uBioRSS feed. Another expert species resource Fishbase (http://www.fishbase.org), a 'global information system on fishes', embeds citation information in its species pages provided by a custom version of the application.

3 CONSIDERATIONS AND EXPANSIONS

Current TNR algorithms can recognize name strings, but without more advanced natural language processing techniques, such as word sense disambiguation algorithms, it will not be possible to reconcile homonymous taxonomic names without human review. For example, the genus name *Aotus* refers to both a monkey and a pea. Without specifying the particular name usage, an article citing the monkey *Aotus* may incorrectly appear among a summary of plant literature. There are plans for future revisions of our TNR algorithm to incorporate word sense disambiguation strategies to help automate identification of intended name usage.

uBioRSS relies entirely on content syndication standards for identifying biological content, but not all publishers currently provide such feeds. For this reason, uBioRSS does not have the breadth of coverage of major bibliographic databases, such as PubMed, which have methods for publishers to submit citations directly to their databases, and sometimes ahead of publication. PubMed has indexed over 30 000 journals and uBioRSS currently indexes about 600 verified biological RSS feeds. Searching PubMed for articles on *Drosophila melanogaster*, a common 'fruit fly', over the last 30 days produced 69 results, whereas uBioRSS identified 54 articles over the same time span. The main reason for this incongruity is that PubMed has a greater breadth of coverage.

We are constantly adding new feeds as they become available, and users can also add feeds that we have not identified. Sometimes the same article appears in more than one feed, and we must identify these duplicates so that they only

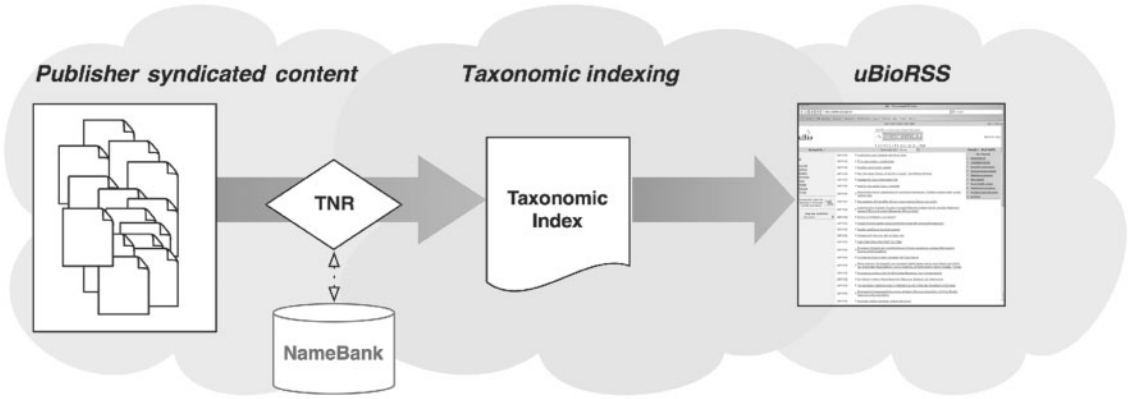

Fig. 1. An overview of the uBioRSS indexing process. A color version of this figure can be found in www.bioinformatics.oxfordjournals.org online.

appear once in the application. To identify duplicates we look for identical DOIs (globally unique and persistent digital object identifiers), or identical article titles with similar article content.

Reference management programs are becoming increasingly popular, and would certainly benefit from a species-centric reference index. Bibliographic reference managers allow people to collate, tag and share citations over the Internet. Users can add descriptive tags to stored citations to organize their citations and find other citations with similar tags. The growing biological index that is being compiled by the uBioRSS application could be used to add scientific name strings as identifying tags to these reference managers (Page, 2006). Users would then be able to search for literature within these services that reference a particular organism name. Taxonomic intelligence could then further improve this process by adding multi-language common names and nomenclatural synonyms as additional tags. The taxonomic indexing process for uBioRSS already associates NameBankIDs, or NameBank LSIDs, with references, or DOIs, so auto-tagging references with taxonomically intelligent tags is a reasonable next step.

4 CONCLUSIONS

uBioRSS is a unique syndicated content aggregator because of its implementation of taxonomic name recognition and taxonomic indexing. The customizable taxonomic hierarchies and custom user settings make uBioRSS appeal to a broad audience of varying interests. As more services take advantage of this index between organisms and literature and display the data providers' content, the data providers themselves will benefit from the increased visibility. The only prerequisite for inclusion into the biological index is providing content via content syndication standard formats. Since these syndication standards can be easy to implement, applications like uBioRSS may provide incentive for more publishers to syndicate content.

ACKNOWLEDGEMENTS

The uBio project was started with funding from the Andrew W. Mellon foundation. I.N.S. is also partially supported by the DAB Lindberg Research Fellowship from the Medical Library Association. Funding to pay the Open Access publication charges was provided by the MBLWHOI Library.

Conflict of Interest: none declared.

REFERENCES

Konig,D. *et al.* (2005) TaxonGrab: extracting taxonomic names from text. *Biodiversity Informatics*, **2**, 79–82.

Page,R.D.M. (2006) SemAnt: automatically growing an ant bibliography. http://semant.blogspot.com/2006/10/automatically-growing-ant-bibliography.html

Patterson,D.J. *et al.* (2006) Taxonomic indexing – extending the role of taxonomy. *Syst. Biol.*, **55**, 367–373.