Databases and ontologies

Client-side integration of life science literature resources

Richard Easty1,†,∗ and Nikolay Nikolov2,†,∗

1 Computing Department, InfoLab21, Lancaster University, South Drive, Lancaster LA1 4WA and
2 Department of Applied Mathematics and Theoretical Physics, University of Cambridge Wilberforce Road,
Cambridge, CB3 0WA, UK

Received on April 8, 2009; revised on August 31, 2009; accepted on September 15, 2009
Advance Access publication September 18, 2009

ABSTRACT

Motivation: The online resources in the life sciences are characterized by a great fragmentation and one of the pressing issues of bioinformatics is making the integration of these resources a smoother and more flexible process than it is currently. Here we present i-cite, a browser extension, which implements a client-side model of integration which improves the navigation within the rapidly increasing life science literature and links terms from it to corresponding non-textual data.

Availability: http://i-cite.org.
Contact: nan23@cam.ac.uk

1 INTRODUCTION

Since 2001 more than half a million new entries are added each year to PubMed (http://www.ncbi.nlm.nih.gov/pubmed), the most popular life science literature database, which contains already more than 17-million metadata records. In addition, vast amounts are stored in PubMedCentral (http://www.pubmedcentral.nih.gov), at the websites of publishers and within institutional repositories. This publication deluge is both blessing and blight as the life science community, to use a popular saying, is drowning in data but starving for knowledge. Thus establishing links between related publications and between publications and related pieces of non-textual data has the potential to significantly improve the navigation within the life science-related information and facilitate its utilization. We are addressing this challenge in three different ways:

• by providing bi-directional navigation between related life science publications via their citation network;
• by linking online life science bibliographies to corresponding PubMed or Agricola records;
• by linking terms in the life science literature to corresponding data resources or other related pieces of knowledge.

In Sections 2–4, we describe each of these features. In Section 5, we give a general overview of i-cite’s implementation. In Section 6, we will give a brief account of the related work. In Section 7, we will conclude and consider i-cite’s future.

†To whom correspondence should be addressed.
∗The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

2 BI-DIRECTIONAL CITATION LINKS

Citations link thematically related publications and allow tracing of a development of a line of research. Navigating the literature via the citation links is thus something that researchers often do as part of their everyday work. Unfortunately, these navigation routes are not so conveniently available for life science literature. PubMed lacks any citation linkage because the citation links need to be extracted from the full texts of the publications and these are often not available due to copyright restrictions. Google Scholar (http://scholar.google.com) has found a workaround for this problem by harvesting publications from the web and entering indexing agreements with some publishers. It offers arguably the most complete freely available citation network for scientific literature. However, Google Scholar being a generic literature search engine lacks the resources that specialized life science literature databases like PubMed offer. In addition, it offers only unidirectional navigation through the citation network (only ‘cited by’ links, no ‘cites’ links). CiteXplore (http://www.ebi.ac.uk/citexplore), by contrast, offers both types of linkage although the completeness of its ‘cited by’ link coverage is inferior to that of Google Scholar. Thus, by integrating these mutually complementing literature resources the navigation paths available for traversing the life science literature can be enhanced. i-cite does this (Figs 1 and 2) by cross-linking these resources and embedding complementing citations and other data in the PubMed and Google Scholar web pages.

Fig. 1. i-cite enhances PubMed’s by embedding bi-directional navigation paths through the citation network (by performing fuzzy match for corresponding entries in Google Scholar and CiteXplore) and links for freely available copies of the full text (as harvested by Google Scholar).
3 ENHANCING ONLINE BIBLIOGRAPHIES

The amount of online life-science literature published outside the traditional journal and conference proceedings venues grows rapidly. i-cite enhances the bibliographical information of such online life science resources by providing bi-directional navigation through the citation network and links to corresponding PubMed and Agricola (http://agricola.nal.usda.gov) records (as found through fuzzy search in CiteXplore). It does this by recognizing citation microformats (http://microformats.org/wiki/hresume#Publications) and performing recognition of the elements of a citation (e.g. authors, title) by parsing it through ParseCt, a text-mining tool employing Conditional Random Fields (Council et al., 2008). The extracted citation elements are used to perform detailed searches in Google Scholar and CiteXplore and fetch the corresponding data in real time and embed it in the original web page (Fig. 3).

4 LINKING TERMS TO RELATED DATA

The life sciences are data-rich and it would improve the understanding of life science literature if occurrences of non-trivial domain entities like proteins or organic compounds were cross-linked to corresponding data. i-cite’s approach is to incorporate calls to two online text mining resources—Whatisit (Rebholtz et al., 2008) and ChemAxon’s Chemicalize (http://www.chemicalize.org) service. As result, the PubMed content can be automatically parsed for a number of biomedical entities, most importantly proteins for which links to corresponding entries in UniProt (The UniProt Consortium, 2009) are inserted through Whatizit. In addition, organic compounds are identified and a popup displaying their 2D structure is embedded within the currently viewed page (Fig. 4).

5 METHODS

i-cite is currently available only for Firefox and consists of a core framework that handles low-level processes (e.g. auto-update of the client scripts, parsing a webpage’s code, monitoring of user-driven events) and provides a convenient interface for third parties to associate specific behavior (e.g. calling a web service) with specific user interactions. i-cite modular design is chosen specifically to both facilitate the extension of its functionality and to give the user the option to choose which feature to enable. Each feature is implemented by a separate module (called a widget) which associates a certain user action (e.g. a mouse-over) with certain behavior (e.g. a call of a web service) by using a simplified API.

i-cite is programmed in Javascript. It uses XUL for the content it generates dynamically. More detail about the i-cite architecture and API can be found in the i-cite developer documentation (http://tinyurl.com/i-cite-dev).

6 RELATED WORK

Zotero (http://www.zotero.org) and LibX (Bailey and Back, 2006) have both provided citation recognition functionality in a browser extension but they have been created with a general audience in mind and lack the distinct life science related functionality of i-cite. Biebar (https://addons.mozilla.org/en-US/firefox/addon/169) adds a browser toolbar that enables quick search in several bioinformatics databases. It does not do any content integration. Reflect (Pafilis et al., 2008) parses the content of a web page and enhances it with information about recognized genes, proteins and small molecules which is similar in kind (though not identical) to the i-cite functionality described in section 4. However, it does not have any of the main i-cite features (cross-linking and integration of life science literature resources) thus i-cite and Reflect are more complementary than competing.

7 CONCLUSION AND OUTLOOK

Data integration in the life sciences is usually done in a server-based mode—by setting up a web site combining data from various sources. Instead of creating yet another web site we integrated life science literature resources within the browser client—by enhancing existing resources thus fluently integrating into existing workflows of the life science researchers. The utility of this approach has been validated by the rapidly growing i-cite user community (nearly 2000 downloads during the three months following the submission of this article). Further directions for our work include automatic recognition of life science citations in free form (i.e. not marked up by citation microformats) on any web page and visualization of
citation maps which would further improve the navigation within
the life science literature.

ACKNOWLEDGEMENTS
The work on i-cite was initiated while both authors were members
of the Literature Services group at the European Bioinformatics
Institute. The authors acknowledge the initial support for i-cite’s
development by Peter Stoehr.

Funding: The work of N.N. was in part supported by Cambridge
European Trust Bursary.

Conflict of Interest: none declared.

REFERENCES
Library Hi Tech., 24, 290–304.
In Proceedings of the 6th International Language Resources and Evaluation
Conference (LREC’08), European Language Resources Association, Marrakech,
Morocco.
Rebolledo-Schulmann,D. et al. (2008) Text processing through Web services: calling