BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–2

Application Note BibPubMed: An alternative interface to PubMed providing results in BibTeX format

Andrew C. R. Martin[a]*

[a]Department of Biochemistry and Molecular Biology, University College London, Gower Street,
London WC1E 6BT

ABSTRACT

BibPubMed is a web interface to PubMed which returns the results in BibTeX format. The interface allows more specific queries of PubMed to be constructed without having to resort to the full PubMed query language and results can be saved or cut and pasted directly into a file for use with LaTeX and BibTeX

AVAILABILITY

BibPubMed may be accessed at: `http://www.bioinf.org.uk/pubmed/`

CONTACT

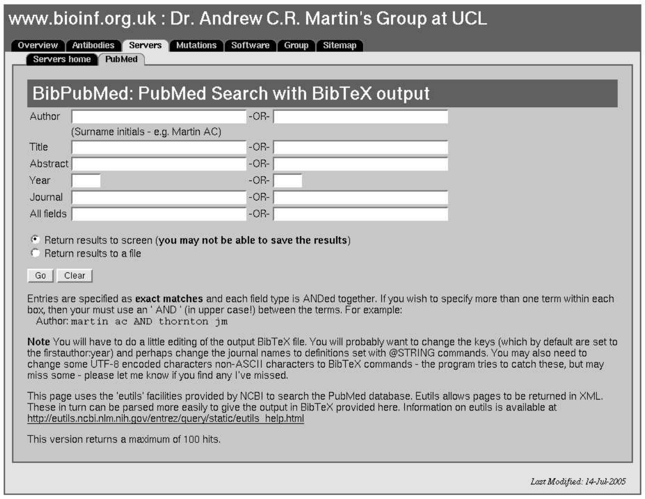
andrew@bioinf.org.uk –or– a.martin@biochem.ucl.ac.uk

1 INTRODUCTION

PubMed is the premiere publicly accessible literature database for the biomedical community. It is normally accessed through a simple text search using the NCBI website. More careful control of queries requires the use of a query language with hard-to-remember field specifications. These are fully documented at `http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html`. For example, to search specifically the author field for ‘Martin’, one must type: `martin[au]`. To search for aricles by Martin in 2005 with the word ‘SNPs’ in the title, one must type: `martin[au] AND snps[ti] AND 2005[dp]` This returns just one hit compared with 20 hits if one searches simply for `martin snps 2005`.

LaTeX (Lamport, 1994), a macro package built on top of Donald Knuth’s TeX system (Knuth, 1985), is a page markup language which allows the production of publication-quality documents with automatic handling of kerning, hyphenation and ligatures. The LaTeX document processing system is widely used in the physical and mathematical sciences and is gaining in popularity for other sciences, particularly Bioinformatics. Several journals (including *Bioinformatics*, *Journal of Molecular Biology* and *Science*) now accept submissions in LaTeX format.

The accompanying program, BibTeX, handles citations and references. Data on references are stored in a simple text file format (a BibTeX .bib file) with each reference including an arbitrary unique key. A citation in the text is referenced via this unique key. The BibTeX program, in association with a selected bibliography style file (a .bst file) creates the bibliography in a format specified by the style file together with inserting the citations in the text in numeric or author/date style. The BibTeX style file can also look after sorting of the bibliography in citation or alphabetical order. While PubMed supports output in HTML, plain text, ASN.1, and MEDLINE formats, it does not support BibTeX. Some online and command-line tools (e.g. `med2bib`, `http://ilab.usc.edu/bibTOhtml/`; `medlinebib`, `http://www.lecb.ncifcrf.gov/˜toms/delila/medlinebib.html`) are available to perform conversions from MEDLINE to BibTeX format. However, there appears to be no online tool that integrates simplified access to specific PubMed search fields with conversion of the output to BibTeX format.



BibPubMed: PubMed Search with BibTeX output

Author -OR-
(Surname initials - e.g. Martin AC)
Title -OR-
Abstract -OR-
Year -OR-
Journal -OR-
All fields -OR-

● Return results to screen (**you may not be able to save the results**)
○ Return results to a file

Entries are specified as **exact matches** and each field type is ANDed together. If you wish to specify more than one term within each box, then your must use an ' AND ' (in upper case!) between the terms. For example:
Author: `martin ac AND thornton jm`

Note You will have to do a little editing of the output BibTeX file. You will probably want to change the keys (which by default are set to the firstauthor:year) and perhaps change the journal names to definitions set with @STRING commands. You may also need to change some UTF-8 encoded characters non-ASCII characters to BibTeX commands - the program tries to catch these, but may miss some - please let me know if you find any I've missed.

This page uses the 'eutils' facilities provided by NCBI to search the PubMed database. Eutils allows pages to be returned in XML. These in turn can be parsed more easily to give the output in BibTeX provided here. Information on eutils is available at http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

This version returns a maximum of 100 hits.

Last Modified: 14-Jul-2005

Fig. 1. The BibPubMed form which allows specific queries to be entered without needing to use the PubMed query language.

We have therefore implemented a web interface which allows specific searches of the most commonly used PubMed fields and which returns results in BibTeX format using the NCBI eUtils system.

2 IMPLEMENTATION

The NCBI has provided a CGI-based remote procedure call interface to the Entrez system (including PubMed) via the ‘Entrez Programming Utilities’ (eUtils). eUtils allows queries to be placed in URLs using the HTTP GET method and returns the results in XML. We have used this system to implement a CGI backend to a form which allows more specific queries of PubMed to be created. The XML results are parsed and the relevant reference information is extracted and converted to BibTeX format.

*To whom correspondence should be addressed

Search PubMed and obtain an XML file containing Entrez primary IDs for the hits:

`http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=martin+ac[AU]&retmax=100`

Obtain an XML file containing the full PubMed data for a given primary ID:

`http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=14764547&retmode=xml`

Fig. 2. Example GET-URL-based queries to search PubMed and return the details for a given hit.

The web form accessible at `http://www.bioinf.org.uk/pubmed/` (Figure 1) allows simple, but specific queries to be constructed much more easily than using the PubMed query language. While the form does not provide access to all the options of the query language, the most commonly used fields are available. Each box on the form allows multiple terms to be AND'ed together and the form allows alternative search terms to be provided for each of the supported fields. Each field type is AND'ed together after alternative search terms have been OR'ed.

The eUtils system separates the search tool from the retrieval of matching entries. Thus the 'esearch' facility is used to execute a query and returns a list of Entrez primary IDs for those papers which match. The 'efetch' facility is used to obtain the complete data in XML for a given primary ID. Figure 2 shows example queries using esearch and efetch.

The CGI script formats the query using the supplied search terms and constructs a URL to access the esearch facility. It uses the Perl LWP module to build the query and package it as an HTTP request, obtaining the results from the NCBI server. The result of the query will be a list of Entrez primary IDs packaged in an XML file. These are then extracted into a simple array and each item in turn is then used to construct a URL to access the efetch facility to obtain the full data for a reference in XML.

Conversion of the XML to BibTeX format is handled by a second Perl script. The XML is parsed using an *ad hoc* parser which was found to be some ten times faster than using XML::DOM. During testing we noted some inconsistencies in the NCBI's XML format (for example some tag names appear in both upper and lower case) and this custom parser is able to handle these whereas a full DOM parser would fail. In addition to extracting the required information from the XML file and formatting it in BibTeX format, this script takes the following actions (1) the author list is manipulated such that each surname is followed by a comma and the forenames (with full stops after initials), each author being separated with an 'and'; (2) the page range is converted into a complete number range — for example '1357-62' is converted to '1357-1362'; (3) the title is modified such that any word which contains capital letters is enclosed in curly brackets, the first letter of every remaining word longer than 4 letters is then up-cased; (4) other special characters (such as percent signs) are escaped with a back-slash; (5) XML entity references (such as &lt; for <) are converted to their LaTeX equivalents; (6) smart quotes are implemented to distinguish opening and closing inverted commas.

The XML data returned by eUtils contains non-ASCII characters (such as accented letters) in UTF-8 format. Support for UTF-8 in LaTeX is limited, so these characters are converted to standard LaTeX constructions using a short sed script based on the script `utf82tex-sed-master` by Richard Mahoney (`http://itec.indica-et-buddhica.org/homepages/scripts/`). For example, the accented character ê is encoded in UTF-8 as <C3><AA> and this is converted to the LaTeX construct `\^{e}`.

Finally the CGI script returns the BibTeX entries to the web server. Depending on the option selected by the user, the MIME header is set to either `text/plain` or `text/x-bibtex`. The former will be displayed on the browser, but it will generally not be possible to save the page to a file (although cut-and-paste will be possible). The latter is a non-standard MIME type and the browser will therefore give the user the option to save the results to a file or open them with an application.

A key is generated for each reference by combining the first author's surname wit the publication year separated by a colon. However, this key is not guaranteed to be unique, so some manual editing of the resulting BibTeX file may be needed to ensure unique keys.

In conclusion, BibPubMed is not designed to provide a complete replacement for the NCBI web interface since it does not provide links out to other resources such as full manuscripts. By definition, the results are not in an attractive screen-readable form. However, the system has been in regular use in our lab for 6 months and has proved invaluable, saving a lot of time when writing documents in LaTeX.

REFERENCES

Knuth, D., (1985). *The TeX Book.* Addison Wesley.

Lamport, L., (1994). *LaTeX: A document preparation system: User's guide and reference manual.* Addison Wesley.