W16–W19 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh453

PubCrawler: keeping up comfortably with PubMed and GenBank

Karsten Hokamp* and Kenneth H. Wolfe[1]

Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada and [1]Department of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland

Received February 15, 2004; Revised and Accepted April 21, 2004

ABSTRACT

The free PubCrawler web service (http://www.pubcrawler.ie) has been operating for five years and so far has brought literature and sequence updates to over 22 000 users. It provides information on a personalized web page whenever new articles appear in PubMed or when new sequences are found in GenBank that are specific to customized queries. The server also acts as an automatic alerting system by sending out short notifications or emails with the latest updates as soon as they become available. A new output format and more flexibility for the email formatting help PubCrawler cope with increasing challenges arising from browser incompatibilities and mail filters, therefore making it suitable for a wide range of users.

INTRODUCTION

Sequence and literature databases are growing at a phenomenal pace. Keeping up to date with the latest developments requires frequent searches through web portals, such as the NCBI's Entrez (1). Even in specialized areas of research, tens or hundreds of hits are often found and users need to sift through them. PubCrawler started its existence as a Perl script that automatically kept track of new results for predefined queries to PubMed and GenBank through the NCBI's Entrez search system. Its usefulness inspired the attempt to publish and share it with the research community. A more user-friendly interface was required, which led to the development of a web service as a wrapper around the program. The site went online in March 1999 (2) providing an update alerting system somewhat similar to the ISI's Current Contents™, but completely free. Upon registration, users can set up their queries for PubMed and GenBank through the PubCrawler Configurator. These are stored and executed at customizable intervals such as daily or weekly. Once hits for a query are retrieved, they are compared with previous reports found for that user, leaving only the new items to be compiled into a web page that closely resembles the look and feel of the familiar Entrez pages. Alerting occurs by email through short notifications or delivery of the complete results.

Table 1. A list of free literature update alerting services

Service	URL
Amedeo	http://www.amedeo.com
BioMail	http://biomail.sourceforge.net/biomail
JADE	http://www.biodigital.org/jade
PubCrawler	http://www.pubcrawler.ie
PubMed Cubby	http://www.pubmed.gov
ScienceDirect	http://www.sciencedirect.com

A number of other selective dissemination of information (SDI) services exist, both commercial and free to the public. Some examples include PubMed Cubby, BioMail, JADE, OVID and ScienceDirect (Table 1). Together with PubCrawler, these have all been recently reviewed (3,4). In this paper, we present information about the usage of the PubCrawler web service and report on changes and developments that have occurred during the past five years.

USAGE

The following sections provide a quick overview of how to use the PubCrawler web service. More details are available through an online tutorial at http://pubcrawler.gen.tcd.ie/tutorial.

Registration

Upon registration, users choose their own account name and password, which protects their personal results page from others. Additionally, a contact email address is required for the alerts when relevant new database entries appear. Specification of a schedule allows queries to be triggered at certain time points. In addition to the daily and weekly range, we have, upon user request, also added a monthly

*To whom correspondence should be addressed. Tel: +1 604 291 5414; Fax: +1 604 291 5583; Email: pubcrawler@tcd.ie

Figure 1. A sample output from a PubCrawler results page. An index at the top provides a quick overview of how many new results were retrieved for each query. Aliases can be used to provide descriptive handles for them. Numbers are reported for the total amount of hits and for the previously unseen items that are filtered out for presentation. Older hits are accessible through links up to an adjustable age limit. Multiple queries can be combined under the same alias. Hyperlinks help navigating through the page to get easily from one section to the next. For each article or sequence a checkbox is provided. Using these allows narrowing down the list of interesting items, which can then be retrieved with the click of a button in a format of choice from the NCBI site.

option. All submitted data are treated with the strictest confidence. Nevertheless, the transparency of the Internet should caution users not to provide sensitive passwords or addresses.

Query configuration

One of the strong points of PubCrawler is the user-friendly configuration of even complex queries through the Configurator,

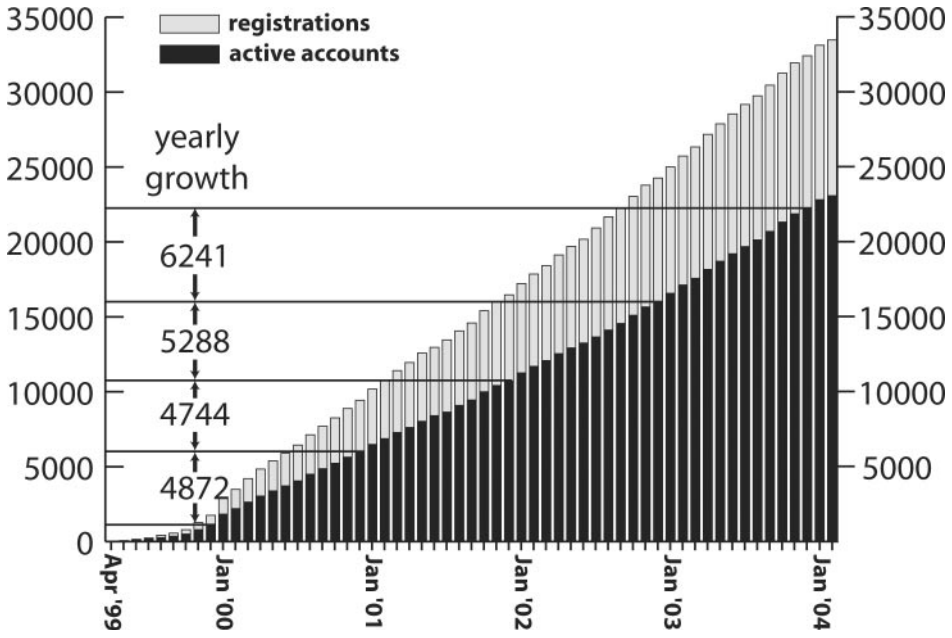

Figure 2. Overview of PubCrawler accounts. The graph presents the total number of new registrations (sum of grey and black bars) as well as the total number of active accounts (black bars), which are those with one or more queries configured and a working email address. For each year's end a horizontal line indicates the number of active accounts and the differences per year are expressed in numbers.

which provides pull-down menus for search fields and Boolean operators. For both PubMed and GenBank queries, there is no restriction on the size or number of queries that can be constructed, which allows very comprehensive searches to be carried out. An interesting feature provided by Entrez lies in the ability to carry out neighbourhood searches within their databases. This is also integrated into PubCrawler and provides notification of new entries related to one's favourite articles or sequences.

Output

Users can access their results any time on a personal web page on the PubCrawler server, or they can have them sent by email. The latter method aids in keeping copies of results from different time points, since the web pages are overwritten every time the queries are carried out. Another option consists of a short notification that is sent to users, alerting them of new updates on their results page.

Originally, the goal of the output was to stay as close as possible to the format provided by the NCBI. Incompatibility with reference managers, and with firewalls, particularly for full results sent by email, moved us towards a change in this policy. The results on the PubCrawler web pages still resemble the NCBI output, but the underlying data structure as well as extra functionality has been revised to avoid browser incompatibility problems (Figure 1). For the emails sent to users, PubCrawler now offers the inclusion or removal of features such as JavaScript, hyperlinks and style sheets, as well as flexibility in the format of the reported hits, i.e. brief, summary and XML. Users can strip any elements from the emails, getting down to the bare results, to avoid their blockage by local mail filter systems. This should provide a sufficient degree of flexibility to satisfy a wide range of technical requirements and personal preferences.

MAINTENANCE

The PubCrawler web service runs with relatively little supervision, and administration consists mostly of referring users to the list of Frequently Asked Questions. One issue that rose in importance is dealing with closed accounts. Without intervention, the number of returned emails that result from changed or deleted addresses would quickly rise into the hundreds in a matter of weeks. This is now handled semi-automatically. Subtracting deleted and inactive PubCrawler accounts from the total number of registrations still shows an increasing growth figure, which resulted in over 6200 additional active accounts in 2003 alone (Figure 2). An account is considered active if one or more queries have been set up and the email address seems to be working. Some users set up multiple accounts, but the discrepancy between active accounts and the associated email addresses is only 3.2% (23 082 accounts versus 22 357 addresses as of February 2004).

Several times modifications of the scripts have been necessary to adjust to new formats and interfaces chosen by the NCBI, but the latest change to the E-Utilities (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html) will hopefully provide a long-term solution.

Queries need to be carried out at off-peak times of the NCBI's Entrez system and at intervals of at least 3 s. To meet these requirements, the load has been spread across multiple computers, which are handled from the main server through a bioinformatics wrapper script (5) originally developed for parallelizing BLAST searches on a UNIX cluster. Further nodes can be easily integrated to meet rising demands.

CONCLUSIONS

Even though occasional published references to PubCrawler boost its popularity, the vast majority of users report that they

found out about it through word of mouth (>70%). Together with the steady increase of user numbers, this is an encouraging indication of PubCrawler's usefulness. The recently added features will further improve its functionality and ensure that PubCrawler continues to be an important tool for biomedical researchers.

ACKNOWLEDGEMENTS

We thank the US National Library of Medicine (NLM) and PubMed for providing access to their databases. Many thanks to Kevin Byrne for help with the system administration and to Indra Konnerth for the new logo and design. K.H. holds a postdoctoral award from the Michael Smith Foundation for Health Research. K.H.W.'s laboratory is supported by Science Foundation Ireland. The European Molecular Biomedical Network (EMBNet) supported the early stage of the PubCrawler project with funding.

REFERENCES

1. Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. *et al*. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res*., **32**, D35–D40.
2. Hokamp,K. and Wolfe,K. (1999) What's new in the library? What's new in GenBank? Let PubCrawler tell you. *Trends Genet*., **15**, 471–472.
3. Shultz,M. and De Groote,S.L. (2003) MEDLINE SDI services: how do they compare? *J. Med. Libr. Assoc*., **91**, 460–467.
4. Carnall,D. (2002) Website of the week: email alerting services. *BMJ*, **324**, 56.
5. Hokamp,K., Shields,D.C., Wolfe,K.H. and Caffrey,D.R. (2003) Wrapping up BLAST and other applications for use on Unix clusters. *Bioinformatics*, **19**, 441–442.