Deliverable: D7.1 - Community contributed bibliography

Partners: OU, KIT

Compiled by: David Morse (OU), Dauvit King (OU), Guido Sautter (KIT)

December 2011

WP7: Biodiversity literature access and data mining

Notes on D7.1 - Community contributed bibliography

Due

23:59 Wednesday 30 November 2011

Not delivered to schedule. This brief report states why we were not able to formally release the service at the intended time, and the steps we will take towards achieving the deliverable.

Issues hindering progress

An early concern was the changing landscape in which the bibliography was to be delivered. There were two aspects to this concern.

ViBRANT technical architecture

One aspect related to the bibliography's integration into Scratchpads. Initially we expected to work with Scratchpads, which were based on Drupal 6. This expectation was superseded by Workpackage 2 bringing forward the development of the Drupal 7 based Scratchpads 2.0 within the ViBRANT project. The complicating factor here being the design of the Biblio module within Drupal 6, and the variety of solutions proposed within the biodiversity community as well as the Drupal community at large to modifying or even replacing this module. Effort was expended investigating this issue, effort which has subsequently not contributed to delivery of the bibliography as now envisaged. This work included investigating the means of propagating our content to existing services such as CiteBank (http://citebank.org/).

CiteBank

The second aspect of the changing landscape relates to CiteBank. The initial vision of the ViBRANT project hosting our own bibliography, with all the long term issues of supporting and running hardware and software that this would entail, was superseded when further investigating hosting of the bibliography. CiteBank was identified as a strong candidate with its strapline *an open access repository for biodiversity publications*. CiteBank is the bibliographic offshoot of the Biodiversity Heritage Library (BHL, http://www.biodiversitylibrary.org/). Citebank matched the sustainable model for the comprehensive bibliography service provided by the *Digital Bibliography & Library Project* (DBLP, http://www.informatik.uni-trier.de/~ley/db/) in the domain of computer science. Work investigating the alternatives (such as using a commercially provided service such as Connotea http://www.connotea.org/, Zotero http://www.zotero.org/or Mendeley http://www.mendeley.com/) identified flaws of one sort or another and in the Summer of 2011 the choice of CiteBank as the repository of bibliographic information was confirmed. Work began on automatically populating CiteBank with data from Scratchpads and other sources via the OAI-PMH protocol.

There were, however, delays in working with CiteBank. Some were technical and administrative, such as shown in this excerpt from an e-mail received by David King from Trish Rose-Sandler which shows that even the normally straightforward task of getting a login to CiteBank was delayed:

> "My apologies for the long delay in this happening. We just discovered our server has been blocking requests to register with Citebank [sic] so we're going back and manually approving users."

However, this work was undermined by the decision of the Biodiversity Heritage Library to re-purpose CiteBank purely as an index to its own content and not for CiteBank to serve as a repository of bibliographic references for the larger community. The ViBRANT project were informed of this decision on 27th October 2011. This change in the functionality of CiteBank necessitated a change in plan within the Work Package. Unfortunately this change in direction was complicated owing to a staffing issue.

Staffing

Though the Work Package had recovered from earlier staffing issues, specifically the later recruitment to the project of both Guido Sautter and David King, new problems emerged in August when David King's mother fell ill. It was a short illness and she died in October. However, it has meant that David took considerable amounts of leave, both paid and unpaid, at this time.

The primary effect of David King's absence was the delay in resolving issues, such as those related to CiteBank. Specifically, with the re-purposing of CiteBank and its replacement by BHL with another service, it was agreed with CiteBank that Work Package 7 would take over the old service in January 2012. That however, would not meet the due date for D7-1. Therefore, alternative arrangements had to be pursued. One option was to establish a clone of CiteBank using the code that BHL has made publicly available (http://code.google.com/p/bhl-bits/). This was pursued with an implementation within the OU. However, this is purely an internal offering. A secondary effect of David King's absence was that the work became entangled with changes within the Open University's technical support arrangements.

Support arrangements within the OU

The Open University has had a general university-wide technical support service, faculty-based support teams and several dedicated departmental based services for those departments with specialist requirements, including the Department of Computing. However, all these services have now been rationalised into one university-wide IT service. The change came into force on 1st December, although requests for work were being deferred during November to make the transition to the new structure easier. This meant our requests for servers and other infrastructure access were being processed, albeit slowly.

There is a further complication in that any public-facing web site (including research web sites) now have to be approved as part of the University's Digital Communications Strategy by the Communications section of our IT support services. This is another reason for the delay in gaining approval for mounting public-facing websites, in addition to a backlog of work created by reorganisation of IT services at the University. As a result, it currently takes three weeks to obtain a public-facing web address, accessible outside the OU's firewall. The consequence of the changes to the OU's IT support arrangements and Digital Communications Strategy are explained in the next section setting out the current status of our deliverable.

Current status

Given the issues with CiteBank that have arisen during October, we attempted to implement two interim solutions to meet the deliverable. One was to build our own version of CiteBank, which we did, the second was to build our own Drupal service which we also did. However, both solutions sit behind the OU's firewall and hence will remain inaccessible for several weeks pending processing of requests for public facing web addresses. Further, both solutions are based on the unsatisfactory Drupal 6 Biblio module. Therefore, we are implementing a different interim solution.

Guido Sautter, as part of his work for ViBRANT and his previous work on processing bibliographic references has been developing a suitable storage system. The code forms the basis for RefBank, which can be seen at http://plazi2.cs.umb.edu:8080/RefBank/search. At present, this node of RefBank only contains some 4,000 records relating to Hymenoptera. We are adopting this code as our base infrastructure for the community constructed bibliography. Currently, we have one implementation within the OU, sized only for the current 4,000 records. On Tuesday next week (December 6th), we should have a larger-scale version of RefBank, sized for 500,000 records running. Though note, this will not be immediately publicly available. See The way forwardon page 4. In the meantime, we are amending the branding of the code and web front-end to fit in with ViBRANT, and are populating the database with more records.

In the long term, we believe that using the RefBank code base will be beneficial because:

1. we can replicate content across all servers. Replication is built into the architecture of RefBank and the system is already running on two servers (three if you count the one that is internal to the OU). The ability to replicate content builds in recoverability to the service and also exposes the service to more potential users whose data, wherever uploaded, will contribute to the *bibliography of life*, and
2. as a stand-alone web service it is immune to Drupal version changes. Furthermore, while we will code a module that will integrate the services offered by RefBank into Scratchpads (this module will, of course, require maintenance in response to changes in Drupal and Scratchpads) the service can also be accessed outside Scratchpads. Through being available to this larger body of users, this should aid its long term sustainability.

An example of the user interface is shown in Illustration 1 (this screen-dump was taken before the ongoing work on re-branding the interface had begun).

The result of running a search for references with an author of 'Agosti' is shown in Illustration 2. Note that references can be downloaded in BibTeX, Dublin Core, MODS and RIS (EndNote) formats.

It is possible for end-users to populate RefBank by uploading references as shown in Illustration 3, in accordance with the aim of delivering a community contributed bibliography. However, we are not using this method to populate the database with initial data to help it achieve critical mass and thereby be useful for its users. We are bulk-loading the database directly.

The way forward

We expect the OU implementation of RefBank to be available with a substantial number of references in the week beginning 5th December 2011. We will continue to add to the references

from existing publicly available resources. We are fortunate that because of the server architecture required by RefBank, our technical services have agreed to expose the service to the web using a temporary URI. This URI will be superseded by the correct one once the Communications section of our central technical support have completed the formal work in 2-3 weeks time.

Pending formal delivery of the OU-hosted service, however, any contributions made to the RefBank service will be replicated into the ViBRANT D7-1 deliverable. Therefore, while the OU hosted, ViBRANT-branded service is not formally available on its due date, the service does exist not only within the OU but in a form that is accessible and usable by end users.

In the short to medium term, we anticipate making the following technical and functional improvements to the RefBank service.

1. Connect RefBank to the Scratchpad 2.0 architecture so that users of Scratchpads are offered the full functionality of the RefBank system, from within Scratchpads.
2. Develop file upload facilities so that users can bulk-upload references in certain formats, in addition to the interactive parsing facilities shown in Illustration 3.
3. Enhance the facilities with a “parse / re-parse this reference” mechanism, to be displayed in the search result page shown in Illustration 2.
4. Add support for plain text export for those users who prefer not to work with a reference manager.
5. Expose the content to Mendeley.

In the medium to long term, we will

1. Integrate the reference parser that is built in to RefBank with the reference discovery engine that is also being developed by WP7 so that users can extract references from papers hosted by BHL and other online repositories.
2. Port the system to PHP. We will also look to implement a different database engine, possibly MySQL if we decide to retain the relational model or MongoDB if we opt for the NoSQL model. This port will add resilience to the service, with each node using a different technology to deliver its share of the service.
3. Investigate the use of Apache SOLR for indexing if performance becomes and issue.
4. Investigate and implement exposing the content as RDF triples for data linking.

We anticipate that release of some of these developments will form Milestones specified under MS7.15 *Define further milestones in the light of usage and feedback*.

Appendix: illustrations

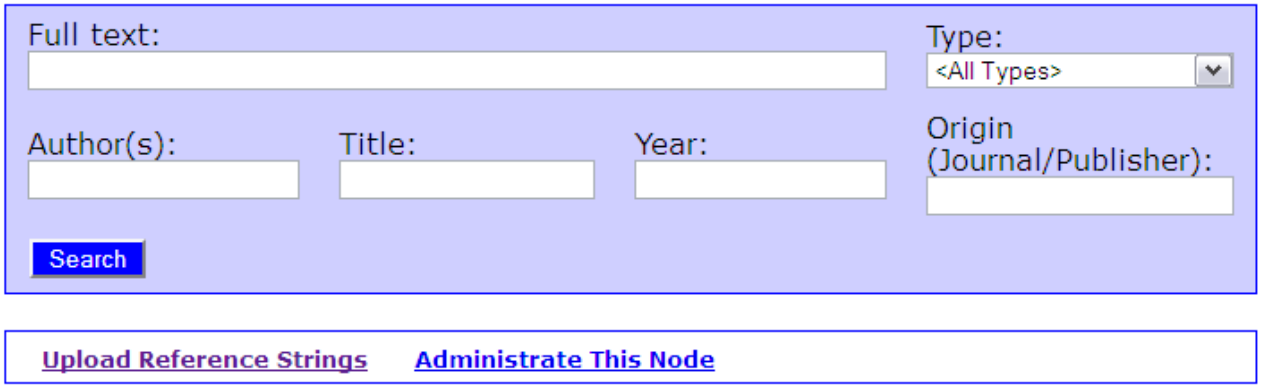

Illustration 1: RefBank - search dialog box

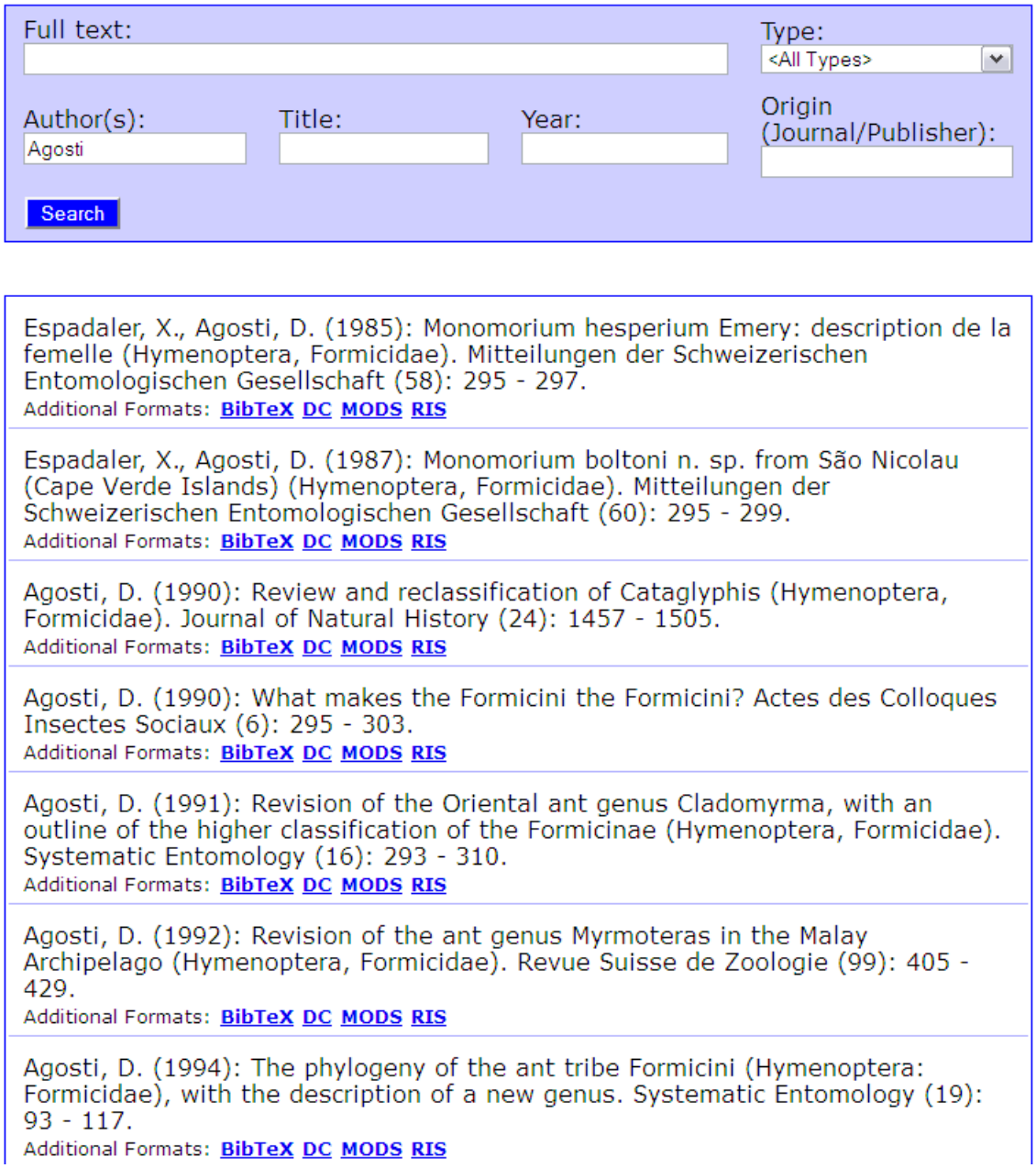

Illustration 2: RefBank - search results box

Illustration 3: RefBank - upload dialog box