Community Requirements Gathering Chat, Week 2, 27 August 2008

We had some new voices this time! Good to see.

LINKS AND DEMOS

- ePrints item report sample: http://eprints.rclis.org/stat/6766.html
- Minho report sample: http://repositorium.sdum.uminho.pt/stats?level=item&type=access&page=downviews-series&object-id=1822/6177
- IRStats sample: http://irstats.eprints.org/irstats-cadair
- AWStats over an entire DSpace repository: http://researchspace.csir.co.za/awstats/awstats.pl?config=researchspace.csir.co.za&configdir=/etc/awstats/ (thanks to Ina Smith of the University of Pretoria, who could not be present on the chat, for emailing this link)
- "Top downloads" and all-repository statistics on the home page: http://www.ideals.uiuc.edu/

APPLAUSE

The chat took a moment to applaud the Repository Support Project's new DSpace Course (http://hdl.handle.net/2160/615). Contributors Stuart Lewis, Chris Yates, and Claudia Jurgen were all present on the chat. The Course is looking for new contributors, particularly with regard to Manakin/XMLUI; if you can help, please contact Stuart Lewis.

AN IMPORTANT DISTINCTION

Mark Wood pointed out (as have several emails to the list during this week's discussion) that two sharply differing concepts lurk behind the word "statistics": the capture of repository events as they occur, and the distillation of raw event data into useful reports. "Statistics pull patterns out of collections of individual cases," said Mark.
Moreover, not all reports are statistical in nature; some (such as "what's been deposited recently" lists) just regurgitate part of the event stream.

Given accessible event-stream data, many statistical analyses can be done wholly outside of DSpace, and it is unrealistic to expect DSpace to create analyses for every imaginable use-case. Some common use-cases, however, may need to become part of DSpace proper; the trouble is defining them.

COMMON REPORTING NEEDS

All access-related reports (accesses/downloads) should filter out as many crawlers as feasible.

- item accesses, total as well as by month and year
- bitstream downloads, as above
- accesses and downloads by author, as above; authors also want to know what their most popular items are
- incoming links from other websites (via referrers; note that referrer spam may become a problem)

Other possibilities mentioned included:

- alerts for download "spikes" over a short period of time
- on item pages, time of last download
- "popular items in this repository" (recent, total, and monthly, though it was noted that displaying this information to end-users tends to feed unjust power-law distribution of downloads)

Geolocating accesses was not perceived as vital.

PRIVACY ISSUES

Claudia Jurgen noted that the EU has very strict privacy laws that may prevent collecting or retention of information that may identify individual persons. DSpace may therefore not be able to track individuals' site behavior (to put toward "more like this" links or the like).

OTHER DESIDERATA

Technical issues: The widely-praised Minho stats engine does not yet work with XMLUI, and no one on the chat knew of plans to adapt it. Mark Diggory noted that event-capture should be separated from log4j's error capturing.

Shane Beers pointed out that DSpace does not currently offer repository managers much information about the contents of their repositories, which is a significant worry vis-a-vis bitstream preservation. A list of bitstreams by MIME type would be a start.

DSpace also does not help managers investigate deposit patterns and growth. A readily-accessible list of recent deposits as well as a list of deposits per time period (separable by community/collection, so that different communities can be usefully compared) would be useful to repository administrators, and should be relatively easy to build via dc.date.available (or for research-tracking use-cases, dc.date.published) metadata.