Zotero/Internet Archive Collaboration Interim Report
For the Andrew W. Mellon Foundation
March 31, 2009

by Dan Cohen, Sean Takats, and Linda Frueh

Executive Summary

The George Mason University Zotero team and Internet Archive Zotero development team have used 2008 effectively to plan and lay the foundation for the roll-out of the Zotero Commons, Zotero-Wayback Machine interactions, and critical infrastructure related to the project. Much of the underlying technical support structure for the project was completed by the two teams in 2008, working in close coordination.

The communication lines between the two organizations are in place--including bi-weekly meetings--and are functioning well. The first-year focus has been on designing and developing the technical base to move files and information between Zotero clients, the Zotero server, and the Internet Archive Data Repository, as well as establishing the critical communications and data layers necessary to execute complex transactions between Zotero and the Archive. Most of the work that has been done is back-end rather than front-end work, and has laid the foundation for a successful 2009 that will enable Zotero users to permanently store and share their materials with other scholars, and leverage the Internet Archive's web archive for citation and annotation purposes, its audio, video and still image collections for primary research materials, and its Open Text Archive for journal literature and books that can be used and shared by scholars.

We believe that the work done by the Zotero and IA teams in 2008 will be highly valuable and reusable by other Mellon RIT projects (such as DuraSpace) as well as other humanities cyberinfrastructure projects. We outline progress on the key deliverables here, and we would be happy to expand upon this interim report over the phone or by other means.

APIs and Server-to-Server Communications

The Internet Archive has had a long history of enabling people to upload data to the archive. However it has never had a formal API. During the course of this project the Archive has developed and made significant progress towards implementing a formal public API for file exchange. Based on the Amazon S3 protocol, this API forms the basis for file transfer to the Archive from the Zotero client. The file-uploading interface is well underway: the Zotero team can currently create directories and upload files to a test IA server via the Zotero client.

Access control has also largely been implemented across the projects. OpenID is now in production use at the Archive and Zotero for account authentication. Implementation of basic edit control and ownership is next. Through our conversations and testing we have ratified the original project plan that access control lists will reside on the Zotero server. Zotero has implemented an OpenID provider to automate the process of creating accounts for Zotero server users at IA.

On the Zotero side, 2008 saw the development of a full-fledged communications and data layer for transactions with the Archive as well as other facilities. The server can now handle synchronous and asynchronous communications (the latter necessary for features like the delayed shipment of OCRed text), as well as file handling using WebDAV. (Zotero users may already take advantage of WebDAV for personal storage.) The Zotero recommendation system is poised to take advantage of the Zotero Commons once files begin to flow into that repository at the Archive.

We are testing a prototype Zotero client with the ability to upload items to the Zotero Commons, and already have a initial screencast of that prototype for beta testers at http://www.zotero.org/static/videos/zotero_commons.mov. We anticipate a Q2 2009 release of an initial Zotero with Commons XPI and contribute/retrieve functionality. For scholarly acquisition of documents in the other direction (from the Archive to a personal research collection), the Zotero team wrote a new translator (the code that make a site Zotero-compatible) so that it better recognizes items on the archive.org site. One-click search of the Internet Archive's Wayback collection for desired web page content was recently implemented in Zotero 1.5b3, an important step toward collaborative scholarly annotation of web pages since scholars now have the option of using a centralized cached/timestamped page rather than different local versions. (The actual annotation technology is waiting for the first phase of the just-funded Open Annotation Collaboration.)

The Archive has already created a system of persistent URIs for Zotero items to enable stable references. Right now, we are uploading Zotero RDF, with the Archive parsing the metadata. We are working out what the long-term data exchange model should be—will we be using something other than Zotero RDF to make it easier to support the proper display of items at IA while also supporting round-tripping of metadata.

Scholarly access to larger collections

The Internet Archive implemented full-text search on Archive text collections through OpenLibrary.org site, a feature that will be applied to Zotero Commons text items in 2009. In addition, over the past year the Archive team has implemented the first phase of OCR conversion which will be applied to uploaded documents. Specifically, the Archive team has established an integrated, automatic workflow whereby scanned text-based pdf files are run through a deriver and OCR software,

and a searchable pdf is created and stored alongside the original document in the Zotero Commons.

In line with the grant proposal to do targeted outreach to scholarly listservs, newsletters, and annual meetings to find scholars with large troves of documents that could be added to the Commons, the Zotero team issued a call for participation for a pilot group of scholars to coordinate the creation of a test Zotero Commons collection. This test will focus on the New Orleans scholars group and primary research materials about the New Orleans area. The goal of the group is to preserve valuable source materials, prove access to them and to build a community of scholarly collaboration.

As part of the pilot project, participants will contribute to the formation of policies and protocols regarding scholarly collaboration more broadly, such as issues related to releasing research data in advance of publication. See http://disc.library.emory.edu/archives/74 for more information (this initial Zotero Commons test group is being managed by former Zotero co-director Connie Sehat, who is now at Emory).

Informal outreach and awareness activities have also been underway, with the Internet Archive team presenting the Zotero Commons at meetings of regional library consortia, the annual Open Content Alliance/Using Digital Collections conference and other scholarly meetings.

Policy Work

The project has been continually monitoring and addressing copyright concerns as the experience with the Zotero server grows and new uses (and possible abuses) are observed. The Zotero team has recently upgraded its terms of service to reflect better a common sense approach to the many functions enabled by the Zotero server and especially the Zotero/IA collaboration. Over the course of the year in the testing of group management and storage requirements for private groups, the Zotero team came to the conclusion that private group file storage would have to take place on a group's own requisitioned space. In other words, the Zotero team has taken action to ensure that infringement does not occur on its server through more stringent terms of service and by offloading private group storage (where files could be exchanged illegally among group members).

GMU Budget Summary

Given some staff transitions and subsequent hiring gaps after the loss of two junior programmers, as well as some conservatism about the "burn rate" on the budget given the difficult economic situation, the Zotero project has spent about two-thirds of the first year allocation, or $151,661. (This total includes a first-phase payment to our subcontractor, CampusMatch LLC, that should clear shortly.) All the expenses to date have been personnel costs. The Zotero team is now once again fully staffed.

Internet Archive Budget Summary

The Internet Archive met its target for the first year project budget. Actual expenses totaled $336,371.62, or 50% of the total Internet Archive grant award. The largest single line item of spending was on salaries and benefits for the IA development team. The project is fully staffed at the Archive. Spending for 2008 also included budgeted expenses for storage servers to support the Zotero Commons. A spreadsheet with detailed expenses is available if needed.