

**Exposing and Sharing the Hidden Archive:
Advancing Scholarship and Academic Collaboration**

Final Report on the Zotero Commons Project
To the Andrew W. Mellon Foundation

Daniel J. Cohen
Center for History and New Media
George Mason University

30 September 2010

The Zotero project at the Center for History and New Media at George Mason University is delighted that 2010 saw the full implementation of the Zotero Commons code and associated CHNM deliverables from the grant proposal. At the same time, we learned some important lessons about the unanticipated complexities of trying to marry two major information systems, Zotero and the Internet Archive. As implied by the no-cost extension we requested with IA in December 2009, this project ended up being somewhat more lengthy to execute than we had anticipated during its conception. Ultimately, however, we're excited about the implementation and think it has tremendous promise as a storehouse of scholarly documents that can be put into the public domain and shared among scholarly communities. All the major CHNM deliverables are now functional and seamlessly integrated into the Zotero platform. We look forward to continuing to fine-tune the project over the coming years as a major feature and differentiator of Zotero, and anticipate (as we have discussed in our end-of-project conference calls recently with the Internet Archive) working with IA to tweak various elements to maximize the utility of this service on an ongoing basis.

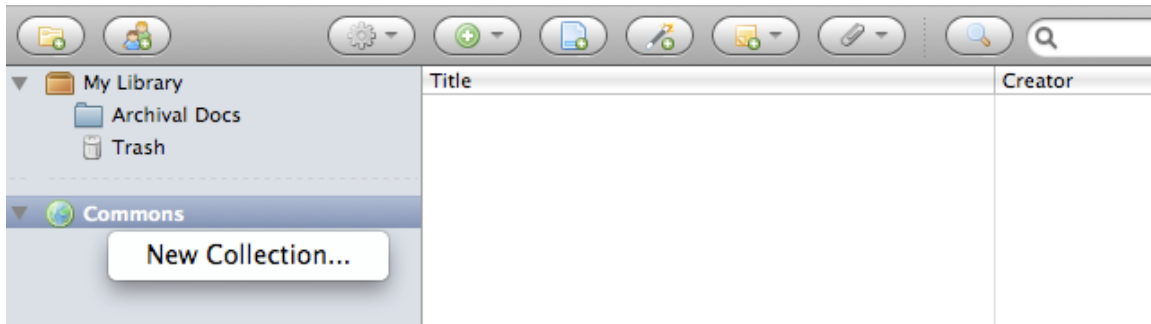
First, an honest appraisal of some of the challenges. This collaborative project ran into two issues that hindered our timely progress and led to substantially more work than anticipated on both sides of the collaboration. The first problem is not uncommon in digital projects that aim to move large or numerous files across the web: the difficulties of ensuring an uncorrupted transfer and associated communication between endpoints. This issue is particularly pronounced given that, e.g., an archeologist can upload a thousand photos of a dig with a simple drag-and-drop from within the Zotero client. After working through various potential solutions using existing technologies and standards, we ended up connecting the Zotero client and server to IA through their clone of the S3 storage protocols that Amazon.com uses for its bespoke server environment, and created an extensive communication and notification system on the Zotero server to handle asynchronous transfers and messages (such as the availability of OCR'd text; see below). The creation of a robust S3-like transfer-and-storage system was a tremendous new addition, and thankfully it handled errors (e.g., when a client disconnected from a network) far better and has produced good long-term results following extensive testing and iterations. It also enhanced IA's API in a way that allowed the Zotero team to build more easily off of it.

Once we had established a robust connection between Zotero and IA, one that allowed not only single-file transfers but the establishment of collections similar to those in Zotero environments (the folders in Zotero become “buckets” in S3 terminology), the second issue had to do with fitting transferred files into IA’s existing information architecture. Here the differing ways IA and Zotero record and display objects came into modest friction; for instance, because Zotero is a reference manager it tends to focus on organization and the display of complete Dublin Core metadata for all objects in a visible pane, whereas IA’s environment, focused more on preservation than active use for things such as bibliographies, sometimes hides metadata in a separate file. We feel this friction has been adequately resolved for the launch, but could probably be updated more (e.g., making it easier to find Commons objects by author on IA’s website) as the Commons receives more use.

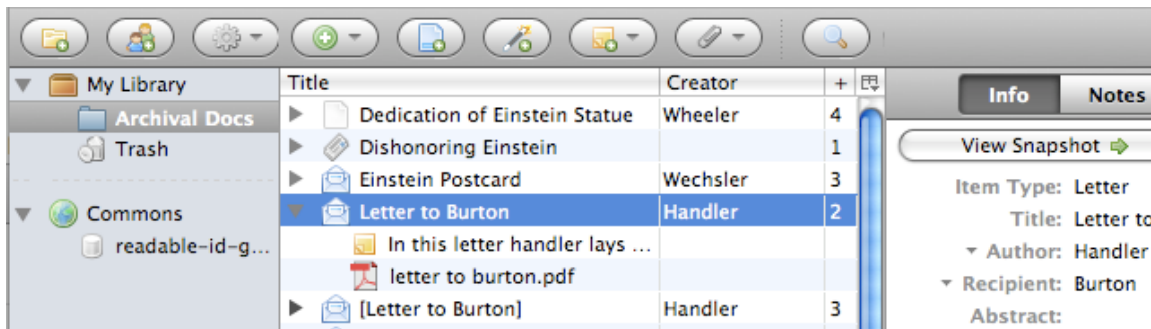
Despite these hurdles, all of the main deliverables have been completed on the CHNM side. Version 2.1 of Zotero pushed out the Zotero Commons code from its initial scholarly test base to all Zotero users who wish to make use of the Commons functionality. Once installed, and as promised in the grant proposal, both the Zotero client and server change their appearance to provide additional options. On zotero.org, the logged-in user sees a new settings tab that stores a key for connecting securely to IA through their API. In the Zotero client a new section and set of icons appears in the lower half of the Firefox extension (see screenshots, below). Notably, all of this code will also work in the Mellon-funded standalone version of Zotero that will connect to non-Firefox browsers, so the Commons work is future-proofed.

Setting up a new collection to upload files to IA is identical to the process for adding a new collection in one’s personal library in the top half of the interface—one click and type in of the name of the collection. In short, the Commons component seems just like a local collection, but with the files replicated in the cloud. Dragging and dropping items onto the collection icon seamlessly transfers the Zotero files and their metadata to IA, where they show up soon thereafter. Items can be added to a Commons collection at any time, from any collection in the upper half of the Zotero interface.

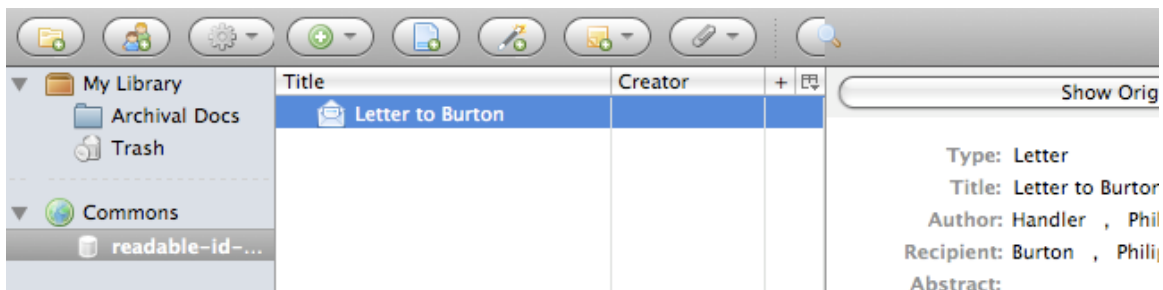
Textual documents are queued to go through the OCR process at IA, and handled on our side by the notification scheme established by the Zotero server. During our testing phase, it took only about 10-15 minutes on average for IA to return enhanced PDFs with the OCR’d text attached to them. The Zotero server then handles the synchronization of the OCR’d text back to the client, where it is further indexed for instant searchability in the Zotero client interface. In some cases there can be a delay in the confirmation of a file arriving initially at IA; in those cases the IA API and the Zotero server delay the appearance of icons in the client interface until the transfer is complete. In short, there is a delicate asynchronous communications dance between the Zotero client, the Zotero server, and the IA servers, and we’re proud of the solution we have achieved given all of the moving pieces and the universally recognized problems of “sync.”



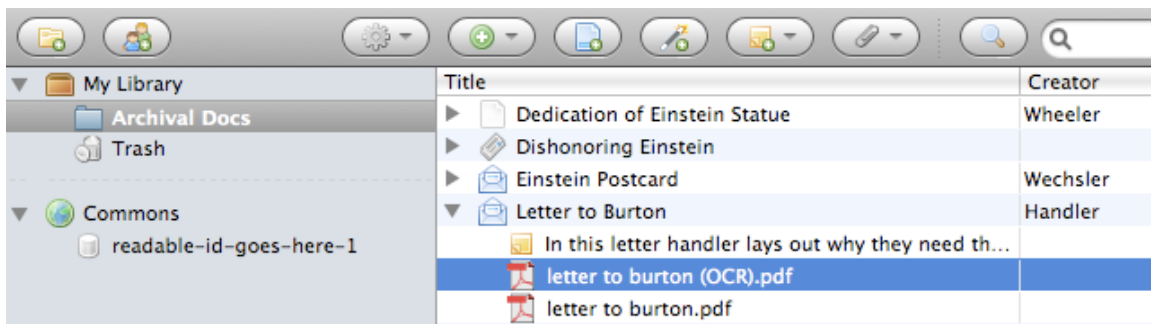
Creation of a new Zotero Commons collection, both locally and on IA at the same time.



Highlighting and dragging an item from a scholar's personal collection to the Zotero Commons.



Appearance of the item in the Zotero Commons following initial processing through the S3-like IA API.



Following OCR of the uploaded document, the enhanced (OCRed) PDF shows up as an additional attachment to the item (we preserve the original as a distinct attachment) in the scholar's personal collection. She can now search the full text of the document in the Zotero interface.

The project also is now poised to benefit from the social and group-oriented functionality of Zotero 2.0. Public and private groups can collaborate on the creation of a curated collection of scholarly materials on the Zotero server, which can then be ported over by the manager of the group to IA for long-term storage and availability. In addition, as promised in the grant proposal, the Zotero client added a method for accessing a canonical version of a web page item in IA's Wayback Machine, which makes it much easier for groups to comment upon the exact same item (saved not only by URL, as had previously happened in Zotero, but also by date and time of access).

This project began with the goal of making available permanent, persistent archival, offsite storage for long-term management and use of digital content; the ability to share resources publicly for easy access by other scholars; and the simplified discovery of new, related resources and potential collaboration opportunities. At the basis of all of this was a deep connection between a user-oriented tool with a large user base and a preservation partner with a substantial archival infrastructure. Despite some hurdles along the way, this connection has been forged and files can now flow from the decentralized universe of scholars to the centralized resource at archive.org—and from there back out to scholars again. Far from considering this project over, we consider it the launching pad for large-scale experimentation on what scholars can assemble together into virtual libraries in their fields.

We thank the Andrew W. Mellon Foundation for their generous support of the Zotero project in general, and this Zotero Commons project with IA in particular.