Peter Suber
psuber@cyber.law.harvard.edu
bit.ly/suber
April 25, 2011

**Enhanced social tagging for classification and current awareness**

**Proposal for Library Lab funding**

I launched the Open Access Tracking Project (OATP) in April 2009. OATP uses social tagging to classify online resources about open access (OA) and to provide a real-time alert service for new OA developments.

This proposal is initially about enhancing OATP. But its chief purpose is to generalize the tools on which OATP is built so that they can support real-time classification and alert services for any topic whatsoever.

Currently OATP uses Connotea, the tagging platform from the Nature Publishing Group. I picked Connotea because it's open-source and because it's optimized for tagging academic resources. (When you tag an academic article, it automatically extracts the key metadata, uses them in the tag record, and makes them available for export in a variety of formats.) While Connotea works well enough for classification and real-time alerts, it needs new features to fulfill all the purposes of OATP. The same is true of every other tagging platform I've had time to investigate.

When I started at the Berkman Center in 2009, I worked with Dan Collis-Puro from the Berkman development team on adding features to Connotea. For a variety of reasons, we abandoned that project. (I'd be happy to say more about this.)

My current proposal takes a new approach to adding the features that OATP still needs. Instead of modifying Connotea or any other tagging software, I'd like to build a new piece of middleware to stand between taggers and readers. Taggers could use any tagging platform they liked, for example Connotea, CiteULike, Mendeley, Delicious, and so on. Each tag from these platforms generates an RSS feed of tagged items. The new piece of middleware would collect these feeds, braid them together, improve them in various ways, and then produce output feeds to which readers could subscribe. For the purpose of this proposal, I'll call this new piece of middleware the hub.

The hub would enhance the input feeds in at least three specific ways: remove duplicates, remove spam, and implement a standard vocabulary or ontology by replacing deprecated tags with preferred or consensus tags.

I deliberately launched OA without a standard vocabulary or ontology. There was only one predefined tag ("oa.new" for new items about OA). All other tags were user-defined (for example,

"oa.journals", "oa.policies", "oa.biology", "oa.france"). I wanted to encourage participation by allowing user-defined tags, and knew that enforcing a standard vocabulary would discourage participation. But because a standard vocabulary is much more useful than a folksonomy, my plan was to cultivate user uptake and then nudge the user community from its spontaneous folksonomy toward a useful ontology. The plan is on track, and over the past two years, the ontology has reached a high state of completeness. One role of the hub is to help implement it.

The hub would produce output feeds matching the ontology, even if the input feeds from taggers continue to use deprecated tags. For example, if the ontology deprecates "oa.journal" (singular) and prefers "oa.journals" (plural), then the hub would make the substitution as it generated the output feed. Some enhancements of this kind could be coded into the hub and applied automatically. Others would have to be made manually, by project managers. For example, the ontology uses "oa.history" for developments on OA in the field of history, and "oa.history_of" for OA developments on the history of OA itself. If I see that a tagger used one tag when he/she should have used the other, then I could add the proper tag to the item, and subtract the improper tag from the item. The changes would appear in the hub's output feeds as soon as they were made and would not affect the original tag records, for example, in Connotea or Delicious.

The incentive for users to subscribe to the hub's feeds, rather than to the feeds directly generated by taggers in Connotea, Delicious, or other platforms, is that the hub's feeds would be more comprehensive, contain no duplicates, contain no spam, and use the project ontology.

OATP already exists, and is already the most comprehensive source of OA-related news anywhere. But to move the next level, it must overcome three problems:

- It must accept tags from any tagging platform, not just from Connotea. Project participants want choice; Connotea itself has an uncertain future; some tagging systems are better than others and better ones will continue to appear; and interoperable tagging will be useful for many purposes beyond the current project.
- It must clean its output feeds of duplicates and spam.
- It must combine the virtues of a folksonomy (maximizing user friendliness and inviting input) and an ontology (maximizing the usefulness of the output). It must support automated and interactive enhancement of the input feeds before generating the output feeds.

By solving all these problems, the hub would support a powerful generalization of OATP. The OA community will use the hub to track and classify developments on OA, but other research communities could use it to track and classify developments on any topics they choose. The hub will offer generalized support for crowdsourced, ontology-based tracking projects. It will provide the software and research communities will only have to provide the crowd and the ontology.

The generalized form of OATP would help librarians and library patrons track and classify online resources on any topic. If subject-matter librarians did this for their topics, or researchers did it for

theirs, they would create a tag-based classification of work on their topics for searching, sorting, organizing, and current awareness.

OATP home page
http://oad.simmons.edu/oadwiki/OA_tracking_project

My plans for the hub are ambitious and over time I'd like to implement them all. But the plans could be implemented in stages. Here are 10 major stages in rough priority order. I hope to finish at least stages 1-4 in the current funding period. That would give us a tool we could actually use while we looked for ways to build the remaining stages.

1. Build the basic hub. It should be able to subscribe to RSS feeds from the major tagging platforms, produce output RSS feeds, combine any number of input feeds into one output feed (without changing the URL of the output feed every time it incorporates new input feeds), and allow both automated and interactive modifications of the output feeds. It should store all tag records for its output feeds in order to support searching, duplicate-detection, backup, export, modification, and so on. The basic hub could be a new piece of software or a modified version of an existing piece of open-source software.

2. Enable the hub to remove duplicate items from output feeds. When the input feeds contains the same item more than once, the output feed should contain the item only once, but that item should include every (approved) tag applied to it in the input feeds.

3. Enable the hub to delete spam from output feeds. As we identify project spammers, we could add their usernames to a file and the hub could omit their tag records from the output feeds. Other spam could be removed manually by project managers.

4. Enable the hub to support tag convergence and the project ontology. We could write substitution commands to a file for the hub to implement automatically (e.g. "oa.journal → oa.journals"). Project managers must also be able to add new tags to given records (e.g. if a record about a new French OA project arrives without the "oa.france" tag, a human participant should be able to add that tag from within the hub). They must also be able to replace an erroneous tag (e.g. "oa.history") with a correct tag (e.g. "oa.history_of").

5. Enable the hub to import all existing OATP tag records from Connotea. These should be stored in the hub, along with tag records created after the hub's launch. (Since OATP has only used Connotea until now, the hub would not have to import records from any other platform.) The hub should treat the old OATP records from Connotea just like the new OATP records for the purposes of searching, duplicate-detection, export, and so on.

6. Support retroactive tag revision. If the user community approves a certain tag today but deprecates it tomorrow, we should be able to replace the deprecated tag with the preferred tag in all the tag records stored in the hub.

7. Enable the hub to run boolean searches across all its stored tag records. At the user's choice, the searches should cover tags, text within tagged records, or both.

8. Enable the hub to keep statistics on how often a given tag has been used, how often a given tag has been used by a given tagger, how many project tags are in use, how many taggers are contributing to the project, and so on.

9. Enable the hub to search the pages tagged by project tags, not just the tag records themselves.

10. Support some kind of ratings so that users can subscribe to output feeds limited to items with a certain rating or higher.

**Budget**

I'd like use all the time allowed under the LibLib guidelines, right up to November 30, 2010. I'd like a full-time programmer for that period (amount unknown), and $15k for my time as the principal investigator. If there are costs in hosting the hub (e.g. at Harvard Library or Berkman Center), the budget should include those costs as well.

Because the web form for proposals required a number, I entered c. $35k. But this should be adjusted to reflect the actual cost of the developer and hosting.