It’s a pleasure to be a part of this because one of my favorite conversations to have with people in the field is about the relationship (and decoupling) between price, cost and value. The previous talks have looked at the impact of data and personal or social marketing, and I look forward to the panel discussion. My presentation has a unique aspect on this in that it is, to a degree, a mix of all of these, plus one that hasn’t been specifically addressed yet – trying to determine impact from the point of view of the Academic Institution. I want also to thank my co-author on this presentation, George Porter, who was a driving force behind the creation and expansion of our repository.
First a little background about Caltech and the Library. The past few years have seen between 150 and 200 PhD dissertations granted per year by the Institution. We currently have 6 subject librarians plus our Repository Librarian. We are a small operation, and thus strategically choose our initiatives.
My last ACS talk in San Francisco focused on “internal” library initiatives serving primarily Caltech faculty and students. At the very end, I briefly mentioned our Institutional Repository, and its role in serving the Greater Scholarly Community. I’m pleased to be able to talk now more about how this expands to the world outside of Caltech in terms of the world accessing our materials, and how we assist our researchers with communicating with “the world” in terms of information promotion and legal compliance. I’m going to describe our repository and our services, and then towards the end, look at this service and show how it ties in with our institutional mission.
The Caltech Institutional Repository is formally known as Caltech CODA – the Caltech Collection of Open Digital Archives. Started in 2001, it is comprised of several different repositories. I will focus on two of them today – CaltechAUTHORS and CaltechTHESIS. Other repositories are CampusPubs which contains our student newspaper and yearbooks, OH which has alumni oral histories, LabNotes which has notebooks and research papers, ES which contains the Engineering & Science magazine, and CONF which houses conference proceedings. More information can be found on our CODA LibGuide.
CaltechAUTHORS was initially launched in 2001 and populated two ways. Motivated individuals (a few faculty and post docs) deposited their own work, and library staff harvested papers based on the ability to make the Version of Record freely available on behalf of the author. In 2009, CaltechAUTHORS was designated by the Faculty Board as the official repository for Caltech scholarly output. Today, staff are trained to perform weekly address/affiliation searches of Web of Science, SciFinder, PubMed, and MathSciNet to search for citations and enter the relevant metadata and files into the repository. Librarians review submissions before posting. Historical “backfile” projects to complete scholarly records of faculty are done as requested and as time permits otherwise.
CaltechAUTHORS allows faculty to comply with the University’s open access policy, and allows for manuscripts and/or final PDF copies of papers to be made publicly available in compliance with copyright. Waivers are granted when needed for articles appearing in non-cooperating journals (such as Nature).
The deposit process involves filling out a form and attaching articles and supplementary information. These are made publicly available when possible (for example when an article has a CC license or an author manuscript version is allowed to be posted). Metadata is added, including author names and identifiers, links to articles & supplementary information, and funding information. Administrative staff often request training on how to deposit for their faculty.
Caltech introduced a new design and layout for the campus websites, which included an option to link directly to CaltechAUTHORS to display an up-to-date list of a faculty member’s publications. There are also advanced options for faculty to create and display custom lists. From my time at the University of Florida, I know this can be an important “marketing” tool (insomuch as Caltech needs marketing) – no one like to see a website that is outdated and a publication list that stopped in 2007. CaltechAUTHORS makes this extremely easy.
Recently the Library has partnered with the Caltech Media Relations department to provide ongoing support to enhance Caltech Press Releases with direct links to papers, including full text where possible. A recent paper on graphene is an example of this. The work is highlighted on the Caltech homepage, and the link in the press release to the scientific article takes a user to the citation in CaltechAUTHORS. From here the user can freely view the abstract, grant and author information, and other metadata. If the article is freely available in CaltechAUTHORS, a direct link to it is provided. Links to publisher pages and other repositories are also provided, and a user who has a subscription to the journal can access it directly through them.
As of now AUTHORS has just over 47,000 records, and almost 6 million downloads since July 2008. There has also been a significant increase in referrals from the Caltech homepage due to the Media Relations effort. This is an access ratio of about 127:1 (6M downloads / 47K articles). So, in terms of impact, we can see that people are looking at our content (but like Tony said, views don’t necessarily translate to impact). Eprints currently allows heatmap generation for individual articles or authors, but not by a topic. We’re curious about the global reach of papers on the specific topics listed here.
4 papers, ~2700 accesses
To get an idea of the global reach of CaltechAUTHORS we used the following procedure to analyze access data. The keywords indicated were searched against records in Eprints, matching on Title, Abstract, or Keyword, and the relevant articles were identified. The access records for those articles were then downloaded, and the IP’s were uploaded to OpenHeatMap.com to generate the maps.
8 papers, ~19,000 accesses
34 papers, ~39,700 accesses
50 papers, ~41,300 accesses
46 papers, ~37,700 accesses
George and Kathy teach classes on CaltechAUTHORS. We have two flavors – one for the general community to learn about what’s in it and how they can contribute, and one specifically for faculty support staff. The staff one focuses on depositing, and is usually of interest to those faculty who want to have their complete publication record in CaltechAUTHORS to use on their website, and also going forward for things like the new NIH Biosketch Bibliography section. This was a very recent development in which I worked with the NIH to have institutional repository bibliographies allowed as acceptable links in the Biosketch, and I would be happy to discuss that in the Q&A or later on during the conference.
Download ratio of 157:1. Alumni who graduated before the ETD requirement occasionally request to have their dissertations scanned and made available, but the majority of requests are from scholars from around the world. When a request comes in, library staff attempt to contact the author and have them fill out a permissions form, after which the dissertation is scanned and made freely available. People want to read our work, as evidenced by the requests.
Kathy Johnson, our Institutional Repository Librarian, teaches several classes each term on the overall process for thesis depositing and related graduation requirements. She also maintains an online LibGuide to assist students and answer questions. (Nice to see the recognition by Tony of ETDs.)
So we’ve talked about what CODA is, and what it provides, and why. So the next questions is, “so what?” What is value / impact? Like Robert mentioned, are these easy-to-measure metrics useful beyond justifying a political agenda? And why does it matter so much in academia? Everything needs to have its cost justified, from subscriptions to salaries. The cost for CODA is in staff time, so is it worth it to provide this resource for the world? Reminiscent of the CCDC, influencing future research, here are Caltech’s Mission and Vision statements. Advancing knowledge for it own sake is what academia is really about, not treating faculty as factories. The demand for our materials demonstrates their value – things won’t get used if they aren’t out there. And like the desktop CSD, there may be applications we may never know about. These tools also serve faculty needs as well in terms of promotional material and citation list management, so internal value should not be overlooked.
Future work – repository software still may have a long way to go to get granular access data.

Engagement – Tony’s example of someone taking his work further and really engaging with it (versus something that could be considered “passive” like citations.) A huge debt of thanks goes to Betsy Coles for wrangling the data out of our Eprints server, which allowed me to generate the heat maps. Kathy Johnson, along with my co-author George, teach several classes each term on CODA. And finally the Library for support. And of course...
...everyone who works so hard on CODA.