

Thoughts on the Present and Future of DL Research and Funding

J. Stephen Downie
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
1-217-351-5037
jdownie@uiuc.edu

Premise #1: There is a current problem with “creeping incrementalism” in Digital Library (DL) research.

A cursory review of the DL literature and attendance at JCDLs 2001, 2002, and 2003 leaves one generally unexcited. What one usually finds are project-based papers, reports and presentations outlining small, incremental improvements in system performance or suggestions for future work. There is a marked absence of the “big idea” or “revolutionary high-impact success.” Much of science is based on little steps forward so this state of affairs is not unique to the DL world. However, if DL research is going to continue in a post-DLI environment, we as a community will need to generate and demonstrate “excitement” both among ourselves and to the world at large. Another symptom of “creeping incrementalism” is the lack of any spirited philosophical debate in the literature or at conferences. There are no warring schools of thought battling for supremacy. There are no intellectual fisticuffs because there really is very little of import to get excited about. Whatever shape DL research takes in the near future one mark of success or failure will be whether or not identifiable intellectual factions emerge that fundamentally disagree, not on the details, but on the very premises of the others’ scientific and philosophical foundations.

“Creeping incrementalism” is caused by Premises #2 and #3 below.

Premise #2: Current DL research is too project-specific.

There are two principal components to almost every current DL project:

1. research to develop tool set x ; for use with,
2. collection y .

The big problem with this paradigm is that x and y are generally inextricably bound together. Even if a given project is “successful” in that the tools it develops are useful for the collection under consideration, there is little evidence that these tools can, or ever will be, exported for successful use to other collections. This lack of cross-collection exportation only adds to the general DL community sentiment of “so what?”. Unless these tool sets can be shown to work—or fail—in novel contexts, there will be no overarching network of techniques to bind the DL research community together. A lack of intellectual cohesion is what gives rise to paucity of common ground upon which to begin meaningful and exciting scientific discourse.

Premise #3: Current DL research is having minimal impact because the tools and techniques developed are not formally evaluated outside the scope of their generating projects.

Most evaluations of DL techniques currently occur “in-house.” That is, each team, in preparation for publication or presentation, conducts a series of “tests” or “evaluations” of its system(s) using some set of evaluative techniques chosen by the researchers themselves. Closely related to Premise #2, the problem here is that this situation precludes a mechanism to answer these four key questions that would give DL research the significance and impact it requires to be “healthy”:

1. How do we determine, and then appropriately classify, the tasks that should make up the legitimate purviews of the various DL domains?

2. What do we mean by "success"? What do we mean by "failure"?
3. How will we decide when one DL approach works better than another?
4. How do we best decide which DL approach is best suited for a particular task?

Notwithstanding the promising technological advancements being made by the various research teams, DL research has been plagued by one overarching difficulty: There has been no way for research teams to properly and scientifically compare and contrast their various approaches because:

- a) there exists no standard collection(s) against which each team could test its techniques;
- b) there are no standardized sets of performance tasks; and,
- c) there are no standardized evaluation metrics.

Over a decade ago, the National Institute of Standards and Technology developed a testing and evaluation paradigm for the text retrieval community, called TREC (Text REtrieval Conference; <http://trec.nist.org/overview.html>). Under this paradigm, each text retrieval team is given access to:

- a) a standardized, large-scale test collection;
- b) a standardized set of test queries; and,
- c) a standardized evaluation of the results each team generates.

Through informal conversations with fellow researchers at various DL conference venues, I have found *strong* antithesis to the notion of adopting TREC-like evaluation methods as part of DL research and evaluation. Arguments against range from the managerial (i.e., How would we ever organize it/pay for it/etc.?) to the quasi-philosophical (i.e., DLs are too heterogeneous and the TREC paradigm too restrictive in scope to make comparisons meaningful). However, I have yet to hear a cogent argument that TREC has *not* contributed immensely to the advancement of IR research both scientifically and as a research community (i.e., giving IR researchers a common ground of discussion and comparison). Without such a common grounding DL research is condemned to be fragmentary collection of idiosyncratic research endeavors of little lasting impact.

Two Future Funding Models for Consideration

Collection-First Model

Under this model the NSF, in conjunction with such large-scale content holders as the Library of Congress, national libraries, publishers, etc., would *a priori* create a large-scale (peta-byte range?), heterogeneous test collection. The test collection must include all possible forms of digital content: music (both audio and symbolic); text (multilingual, simple and marked up, etc.); video, sound, and scientific data in various forms, and so on. In short it must in some sense replicate the contents of a real library. This collection should be created/selected based upon the idea that at the end of research period, the collection *itself* would find a large and eager audience that includes persons of all ages, education levels and nationalities. This broad range of potential users will also help politically in that supra-funding bodies such as the US Congress will be able to comprehend the potential "big impact" of the research dollars spent.

Once this collection is created, the NSF should structure its funding call so that each applicant *must* conduct its research against this collection. Not all teams will necessarily tackle all aspects of the collection (but preference should be given to those with broadest range of techniques). The NSF should also look toward ensuring that each media type has *several* teams investigating techniques. In this way, a common focus is assured and thus a common discourse can be established. This model also affords meaningful comparisons between approaches on a potential head-to-head basis. For example, given the overlap in research on, say, music, it should be possible to determine which of the teams has actually afforded better access to the materials. Also, failing a clear cut "winner" it would finally be possible, perhaps, to

amalgamate approaches (i.e., interface from one, search engine from another, indexing scheme from a third, etc.) because the teams have the underlying data in common.

Forced Third-Party Evaluation

This model applies to both the “collection first” projects or the traditional technique/collection-bound projects. It involves the idea that all funded projects *must* submit to and actively engage in third-party evaluation of their work. “Third-party” is used here because it will not be the projects themselves (First Party) nor the NSF (Second Party) but rather independently-run evaluation research teams (Third Party) that develop, conduct and report upon the evaluation of the various systems. These Third Party “evaluation projects” will be funded by NSF on a competitive basis. If NSF is funding DL research on a four-year funding cycle, then the call for evaluation projects should go out in Year 2 of the cycle. The evaluation projects themselves should also be on at least a four-year cycle but lagging the “production” projects by one year. This lag time gives two advantages: 1) evaluation project proposers will have an overview of the funded “production” projects so meaningful evaluation paradigms can be constructed; and, 2) post-production evaluation of systems can take place after the production projects have completed their work. Some of these evaluation projects might, or might not, be TREC-like in their orientation. The key here is that the Third Party evaluation projects will help to create a common ground upon which to situate the production projects. Without this common ground, no foundation for future DL research can be laid.