Harvard Scholarship Indexing Project (HSIP): Planning Phase

Library Lab Proposal (Revised)

Amy Brand

October 21, 2011

Project concept

HSIP will build a lightweight metadata-only repository to catalogue and link to scholarly artifacts created at, or in affiliation with, Harvard.

As an initial planning phase, we propose that Library Lab fund a study to gather requirements and assess options for creating within the Library a University-wide bibliographic metadata repository for data management, discovery, and distributed access to works produced by Harvard scholars.

Before-and-after summary

Before HSIP

- No institution-wide record of Harvard scholarship
- Harvard research information in School data silos
- Absence of standard data models
- Researchers and staff required to rekey the same bibliographic metadata into multiple local and central systems

After HSIP

- Panoramic institutional record of Harvard scholarship
- One-stop metadata entry and claiming of auto-populated records
- Standard data formats to ensure reusability and system interoperability
- Bilateral data exchange among the HSIP repository, School-based systems, and outside data sources
- A foundation for data-driven decision processes in Library strategy and other University-wide academic planning endeavors

Why HSIP is more Library Lab than business-as-usual

HSIP is game-changing for Harvard because:

- It proposes distributed cataloguing of objects that the Library may never collect or otherwise steward.
- It situates an important component of the University's active academic information management technology and strategy within the Library.
- It proactively broadens Harvard's scholarly record to include any artifact or ephemera that Harvard researchers choose to record as a scholarly communication.

Rationale

Strategic planning efforts underway within the new Harvard Library organization emphasize, in general terms, a growing role for the Library in scholarly communications and digital scholarship.

The communications of scholars encompass not only book and journal publications, but also “informal” communications and a variety of other work products, some of them made possible by recent advances in information technology. It is not uncommon, for example, for an academic CV or faculty activity report (the productivity report that faculty members provide each year to their Deans) to list news articles, blog postings, artworks, patents, datasets, recordings of government testimony, computer code, and other artifacts.

Efforts already underway within the Library to capture and (potentially) collect publications and other artifacts of scholarship that Harvard researchers deposit, such as DASH and Zone 1, would benefit significantly from a parallel effort to catalogue the scholarly communications of Harvard researchers, whether or not the Library ever holds the artifact itself. Such an index would provide a more comprehensive record of the University’s scholarly process and product, including works to target for deposit in these repositories. At the same time, HSIP records would streamline the work of the Library to create metadata when any of these items are actually deposited in a repository or archive for Library stewardship.

Other tangible HSIP benefits for the Library include:

- A foundation for a data-driven, unified collection development strategy, based on documented areas of scholarly activity and research interest across the University’s Faculties.

- A resource for reference librarians who are called upon to identify Harvard experts outside of their own Schools or areas of expertise.

- A Library-run system to support Schools, researchers, librarians, and administrators by streamlining data entry and management, such that a new citation only has to be keyed in or imported once and it will then propagate to researcher websites, activity report sites, institutional and subject repositories, etc.

- A Library-controlled data source to underpin the development of faculty profiles, search aids, and CV generation tools, as well as open access to this metadata, consistent with other open data efforts.

- An alternative to buying “Harvard data” from third parties, which we are likely to have to do in the near term to support a variety of institutional research and networking initiatives. Efforts to address the commercial publisher stranglehold over academic libraries via institutional repositories and new business models should cover not only full-text items but also bibliographic metadata. In the absence of HSIP, we risk increased reliance on externally provided data and solutions for tracking information about our own scholarly output, information that HSIP could help capture where it originates at the University.

- A means to help legitimize a broader definition of academic productivity – and, by extension, alternative reputational heuristics – by creating a Harvard-wide scholarly record that includes the full range of outputs that researchers currently consider to be scholarly/academic activity and would like to have reliably attributed.

Envisioned user experience

Imagine a flexible interface that allows the user to key in a "fielded" citation or just a work identifier, or to cut and paste a reference, and to designate target sites for the record from a list of export options (profile, activity report, repository, etc).

Imagine a general-purpose artifact schema to support descriptive metadata used in discovery and identification (author/contributor, title, abstract, keywords, date), as well as administrative elements such as artifact type, address, and timestamp, and work and contributor identifiers that also serve as persistent links (e.g. DOIs, ORCIDs).

The ideal solution would employ data standards to support bilateral data exchange among the HSIP tool, School-based systems, and outside data sources. It would acquire structured data where the incentives to contribute are highest – for example, when faculty members are asked to update their annual activity report or CV.

Since most School reporting tools do not function as publication or activity data repositories, these systems could optionally direct users to update their information in the HSIP repository in order to fulfill their reporting requirements (and for export to other systems as appropriate). An auto-update feature based on harvesting data from outside sources would serve as an added incentive to use the service. Alternatively, the School reporting tools themselves could be enhanced to support structured/fielded data entry for publications and other works, and submitted data would then populate both the local report and the HSIP repository.

Planning phase, budget, and related projects

The proposed planning phase is for 95% outsourced effort by a consultant with appropriate expertise, and oversight by the project lead (5% effort). The intended output of this initial phase is a report that addresses:

- Assessment of overall need and feasibility
- School, Library, and IT requirements
- Data models (schema) and standards

- Implementation options and associated costs
- High-level project plan

The proposed budget of $15,000-$18,000 is intended to cover 100-120 consultant hours, for a project to be completed within a six-week timeframe. For example, a private consultant with relevant expertise has a quoted rate of $150 per hour.

The consultant would be directed to obtain input from other stakeholders and experts, including HUIT staff and representatives from each of the Schools. Several internal stakeholders have already expressed an interest in serving as a resource for HSIP:

- Michelle Pearse, Research Librarian for Open Access Initiatives and Scholarly Communication at Harvard Law School, would like to coordinate with the Library Lab funded "deposit@harvard" project and related efforts at HLS.

- Joy Sircar, Associate Dean for Research and Planning at SEAS, is developing a local activity tracking system and has offered to make SEAS faculty data available for proof-of-concept and piloting solution options that may emerge from this project.

- Reinhard Engels, Digital Library Software Engineer working on the DASH repository, has indicated interest in this project as a way to help define requirements for future enhancements to DASH.

- Wendy Gogel, Manager of Digital Content and Projects, is interested in exploring synergies between HSIP and Zone 1, which collects no descriptive metadata.

- David Weinberger, Co-Director, Harvard Law Library Digital Lab, views HSIP as an additional source of metadata for Library Cloud and would like to be consulted on data formats.

As an example of an external expert we would consult in planning HSIP, there is a leading expert on bibliographic metadata and schema development based in the Boston area. He was consulted in the development of DASH and has worked previously with the project lead on other large-scale projects, including the CrossRef XSD schema.

Success?

The proposed study will create a report that can be considered sufficiently rigorous if it provides a high-level project plan, and enables relevant leaders to decide whether to endorse HSIP implementation and an associated request for further funding.