John Wilkin, Executive Director

University of Michigan Library
818 Hatcher Graduate Library
920 North University
Ann Arbor, Michigan 48109-1205

Beginning in 2013, HathiTrust will use a cost model that reflects benefits that partners can receive from works stored in HathiTrust rather than the cost associated with storing them. We believe that this new cost model, focusing on consumer values rather than storage costs, better reflects the long-term interests of partners and will more fairly distribute costs across the partners.

The current storage-based cost model for HathiTrust attributes to Partner Libraries the full cost of basic infrastructure related to storing content that they deposit. That is, in 2010 costs, each gigabyte of content entails an array of basic services—storage, basic management, basic access, etc.—that costs $3.86 per GB per year. When a Partner Library deposits content, the Partner Library pays for those infrastructure costs, and thus pays $3.86 per GB per year. Other associated costs are borne as contributions of effort from Partner Libraries or may be funded from auxiliary funds (e.g., grants or a collective allocation process).

This new benefits-based cost model attributes the cost of storing a volume to each library that holds (or held)[1] a corresponding print volume. In this model, a library that pays for a share of the cost of storing content is also acknowledged as receiving the benefit of the content. Those benefits for in-copyright volumes are not always tangible, but include producing replacement copies and offering specialized services permitted by contract or law. Such a cost model would, for in-copyright works, attribute a share of costs to all partnering libraries that hold or held the corresponding print volume. Because all member libraries enjoy the benefits of public domain works, every partnering library will be assumed to hold these works.

This new model of cost attribution will, compared to the current storage-based model, have a smoothing effect, reducing the cost borne by an institution that contributes significant content, and recovering cost from each member of a new class of partner libraries ("Sustaining Partner" libraries) that shares in the benefit of these volumes.[2] The current HathiTrust membership is comprised primarily of institutions contributing large amounts of content and, thus, bearing large costs. At this time (late 2009), we are in conversation with research libraries that do not have

[1] Particularly for Section 108 uses, a library may wish to withdraw (and thus no longer hold) a volume. We should store information that shows that this library once ***held*** the volume in question.
[2] The philosophy of collective costs and collective holdings already underpins much of the CIC approach to HathiTrust.

large amounts of content to contribute, but wish to join HathiTrust to participate in its curatorial work. Under the current model, partners with smaller amounts of content pay relatively small amounts and have access to large amounts of content. By applying a model based on shared holdings, we will see some reduction in the cost per *contributing* institution as more libraries join the effort. Clearly, as these new libraries join, it must be with the understanding that a new cost model will work to distribute these costs in an equitable way that reflects the benefits accruing to all partners.

This holdings-based cost model will incorporate a number of precise elements about costs and the "sharedness" of the content. The cost will be re-calibrated each year, as costs for infrastructure will change each year. It is also the case that the current partners would like to be able to use shared funds to develop new services and functionality. Consequently, we will multiply the cost of infrastructure by a variable amount (adjusted periodically) to fund those new services and functionality. In this new holdings-based cost model, the costs to an institution per year will be calculated as follows:

For public domain volumes:

(PD*X*C)/N

where

- PD is the *total* number of public domain volumes in HathiTrust (assumed to be "held" by *all* partner libraries). This number will also include in-copyright works where the rights holder has given the members free use of the content.
- C is the average annual cost to provide basic support for a volume. Note that costs will vary by volume, as each printed volume will vary in number of pages and in average file size. We will use an average cost that will be periodically recalculated.
- X (greater than one) is a value with which we multiply C to generate a surplus.
- N is the total number of partner libraries.

For a given in-copyright volume, IC:

IC = (C*X)/H

where

- C is the average annual cost to provide basic support for a volume (as above)
- X (greater than one) is a value with which we multiply C to generate a surplus.
- H is the number of partner libraries that hold a given print IC volume

Initially, HathiTrust proposes using a value of two (2) for X, i.e., doubling the cost of maintenance in order to build a fund for services. Thus, 50% of the funds collected will go to development and the other 50% will cover the costs of storing content, again lowering cost proportionally for institutions that contribute large numbers of commonly held volumes.

We believe that this new model will be both equitable and sustainable, but acknowledge that it also presents significant challenges. The biggest of these is the lack of information in the library community currently, on a volume-by-volume and institution-by-institution basis, about overlap in our print collections. No organization, including OCLC, stores information about *holdings*, and even where OCLC succeeds in approximating this, it lacks volume-specific information. We also face challenges with regard to reliable data. For example, although each of our institutions individually stores volume-specific information, enumeration and chronology information is represented so variously that manual remediation will be required to make it uniform. Ensuring that this information is up to date is an additional concern. HathiTrust, either by itself or in collaboration with another organization, will attempt to create a system to store partner holdings information in such a way that it can be updated constantly, by partner institutions themselves or by central HathiTrust staff.

We believe that this volume-specific infrastructure will be valuable for a number of purposes, including:

- **de-duplication**: although duplication of contents is not costly, storing duplicates compromises the user experience and it obscures collection development needs;
- **management of corresponding print volumes**: how will we know that we can withdraw a print journal without having volume-specific information?
- **legal uses of in-copyright materials**: For example, Section 108 uses will depend on having a clear sense of which institutions own(ed) which print volumes.

A clear sense of volume-specific information for digital materials and corresponding print volumes will be needed as our collaboration in HathiTrust develops.

HathiTrust plans to launch this new cost model in 2013, during the second phase of our initiative (i.e., subsequent to the first five years of HathiTrust). Prior to that time, we will work to develop the necessary infrastructure to be able to perform these calculations reliably. We will also, effective immediately, entertain membership from other research libraries that wish to share this curatorial role. To calculate costs for these Sustaining Partner libraries, we will use overlap formulas developed in our current explorations of this model with RLG, ReCAP and New York University Library. All funds generated through the participation of Sustaining Partner libraries prior to 2013 will be devoted the development of the new common holdings infrastructure.

jpw, 12 Feb 2010