What are we to Keep?
James R. Jacobs
Documents to the People, Spring 2015, p 13-19.

(part of the collaboratively written feature, "Thoughts on the National Collection" by James R. Jacobs, Shari Laster, Aimee C. Quinn, and Barbie Selby. I'm posting my segment as it was written under a Creative Commons Attribution-NonCommercial-ShareAlike CC BY-NC-SA license)

The question of "how many copies" of print documents the FDLP should collectively keep is the wrong question asked for the wrong reasons and trying to answer it will only lead to the wrong answers and irreparable loss of information. For me, even thinking about answering it raises more questions. How can we know how many copies to keep unless we specify the purposes for which we wish to keep them? What are those purposes? How will we know if we are meeting our goals? How will discarding paper benefit users? How can we be sure that we are not losing information when we discard paper copies if we do not have an inventory of the paper copies that exist? How can we implement a policy that is so vague that it doesn't define things like "a requisite number of copies," and how decisions will be made, and which apparently treats a born-digital XML document created by GPO and an indifferent digitization without OCR text and missing its maps and foldouts as of equal value?

Let's be clear. We are talking about the records of our democracy. Loss of even a single page could damage the ability of historians, journalists, economists, and citizens to understand our history and hold our government accountable for it successes and its failures. We have those documents now in our libraries; there are not hundreds or even dozens of copies of these documents floating around in used bookstores or elsewhere. They are in our charge.

I know many librarians that I respect think we have no choice, that libraries have to discard paper copies.[1] They believe that we can do so safely (that is, without risking loss of content) once we have digital surrogates of those paper copies. They think that if we save one or two copies, we can always go back and correct digitization mistakes or omissions. I am here to tell you that those opinions are wrong and we know they are wrong from existing data, experience, research, and common sense. We risk significant, irreplaceable loss of information and damage to our libraries' missions if we discard without taking sufficient care to minimize the risks; and picking a number out of a hat (without research, data and measurable goals) is the definition of "without taking sufficient care."

I have very little space, so, instead of trying to convince you that any particular number I might pull out of my hat is better than any number anyone else pulls out of theirs, I want to provide you with some other facts and numbers with which you might not be familiar.

HathiTrust (HT) is often used as buzzword-defense by advocates of discarding paper documents. So it is reasonable to ask how much we can (or should) rely on HT for replacing the documents we discard. Paul Conway's research on the quality of HT digitizations includes these findings:[2]

- Although a "small proportion" of pages are unreadable, that small proportion adds up. In just 12% of the volumes in HT he examined he estimates that there are about 1.4 million pages with "indecipherable" text.
- Fully one-quarter of 1000 volumes examined contain at least one page image whose content is "unreadable."
- 1% of the volumes he examined had missing pages and an additional 2% were severely dog-eared or had portions of the pages missing.
- In a 1000 volume sample, only 64.9% were considered accurate and complete enough to be considered "reliably intelligible surrogates."

HathiTrust itself reported in 2012 that 84.9 percent of the volumes it examined had one or more OCR errors, 11% of the pages had one or more errors, and the average number of errors per volume was 156. When the Center for Research Libraries certified HT as a trusted digital repository, it specifically noted that the quality assurance measures for their digital content *do not yet support the goal of withdrawing print volumes*.[3]

We have existing research that examines the number of copies of journals we should keep. This research presents examples that vary from a low of 5 to a high of 96 copies. Should we choose one of those numbers? *No.* Even with this wide range of examples, we cannot use this research to determine how many copies of documents we should retain. Why? Because even this research does not recommend numbers; instead it recommends methods a library can use to determine numbers. Most of this research deals with text only, not illustrations, tables of numbers, maps, and so forth. Most of this research estimates the number of copies to keep for a very limited goal: that we can, with high confidence and low risk, have one copy after 100 years.

If we want or need access copies or re-digitization copies – which I emphatically believe we do – or want to have more than one preservation copy in 100 years (yes, we must), we will have to do better than wave our hands and say "two copies is enough" in 2015.

Let me close with a positive suggestion. Those who truly believe that it is time to discard paper copies of our historic documents collections should at least contemplate the damage that could be inadvertently caused if careful measures are not taken to avoid damage and loss of information. Specifically, before librarians even consider a policy of mass discarding of documents we should:

- Know that we have complete, accurate digitizations of those documents.
- Know that they are safely held in a publicly accessible, trusted digital repository.
- Specify our goals for keeping paper copies (including access, Interlibrary borrowing, preservation, and re-digitization).
- Have accurate and complete bibliographic holdings data so that we can make informed decisions.
- Have reliable research that recommends methods of applying the data to achieve our goals.

FDLP libraries have a commitment to our country's history. No one else has this same commitment and no one else has the same valuable resources that we are thinking about discarding as if they no longer had value. Your management may want to renege on that commitment, but before you acquiesce, please consider the damage this might do. If you want to know more, please see the additional information and bibliographies we are putting on FGI. http://freegovinfo.info/dttp/.

James R. Jacobs
Stanford University Libraries

[1] I disagree with that premise profoundly, but I do not have room here to explain why. I will post more about this on FGI.

[2] Conway says that "large-scale digitization has established a new ethical norm" for digitizations. Should this "new norm" apply to the records of democracy? Paul Conway, "Measuring Content Quality in a Preservation Repository: HathiTrust and Large-Scale Book Digitization." Proceedings of 7th International Conference on Preservation of Digital Objects, iPres 2010, 19-24 Sept. 2010, Vienna, Austria, pp. 95-102. http://hdl.handle.net/2027.42/85227.

[3] Certification Report on the HathiTrust Digital Repository. March 2011. Center for Research Libraries. http://bit.ly/crl-hathitrust-report.