Digital Collections as Big Data

Leslie Johnston, Library of Congress

Digital Preservation 2012

Data is not just generated by satellites, identified during experiments, or collected during surveys.

Datasets are not just scientific and business tables and spreadsheets.

I do not need to convince this audience that we have Big Data in our Libraries, Archives and Museums.

More and more researchers want to use collections as a whole, mining and organizing the information in novel ways.

Researchers use algorithms to mine the rich information and tools to create pictures that translate that information into knowledge.

Researchers may want to interact with a collection of artifacts, or they may want to work with a data corpus.

We still have collections. But what we also have is Big Data, which requires us to rethink the infrastructure that is needed to support Big Data services. Our community used to expect researchers to come to us, ask us questions about our collections, and use our digital collections in our environment.

Now our collections are, more often than not, self-serve.

Case Study: Web Archives

- Web Archives, such as the one at the Library of Congress, may be comprised of billions of files.
- When we began archiving election web sites, we imagined users browsing through the web pages, studying the graphics or use of phrases or links. But when our first researchers came to the Library, they wanted to know about all those topics, but they used scripts to query for them and sort them into categories. They were not very much interested in reading web pages.
- The Library is testing tools for full-text indexing of the entire archive and collection subsets

http://www.loc.gov/webarchiving/

Case Study: Historic Newspapers

DAILY PRESS, NEWPORT NEWS, VA., FRIDAY, OCTOBER 21, 1910. 7

AN EXPLANATION
TO CLEAR UP A FALSE IMPRESSION

IT HAS BEEN WIDELY KNOWN that the Bank of Hampton has held a mortgage of $40 000 on the property we have advertised so extensively for sale for the past 10 days. This mortgage was assumed by us when we purchased the property. It did not fall due until May 25th, 1911, but WE HAVE PAID IT IN FULL WITH INTEREST TO THAT DATE. We understand that the existence of the mortage has deterred many from purchasing and for this reason we deem it wise and proper to advertise the fact of its payment. Any person desiring to do so may examine in the Clerk's Office of the Corporation Court of this city the deed releasing this mortgage. All persons who have hesitated to buy because of this mortgage and those who have purchased may come in at their convenience and obtain a clear deed to the property. If any person has any doubt about the sufficiency of our title to the property, we desire them to make a full examination. Recently it has been passed upon by Messrs. O. D. Batchelor, R. T. Lett, A. C. Garrett, and S. O. Bland, attorneys of this city. Many of the recent purchasers are attorneys, among whom are Messrs. Maryus Jones, W. B. Colonna and Floyd A. Hudgins.

We Sold Nineteen Lots During the First Ten Days for Over Thirty Thousand Dollars, ($30,000.00)

Some of the "calamity howlers" around Newport News and Hampton have nice homes for themselves, but think no others are needed.

Reports have reached us that we could not sell for various reasons, and that if we could, we could not give a clear title to the deed—that we were bluffing. We believe these reports have deterred many purchasers. Many did not take this advice and many who did must now pay more for this property; we can sell and we are selling—we can and we are giving them clear deeds. The developments of the past few days show that we are not "bluffing." Since the advice of these would-be advisers is worthless let us advise you.

The Prices for Another Ten Days Will Be As Follows:

Avenue Corners,	$3,000 each
Avenue Inside Lots	$2,000 each
North Side Thirty-second Street,	$1,200 each
South Side Thirty-second Street,	$1,000 each
South Side Thirty-third Street,	$1,000 each

THESE PRICES ARE LOW!

We advise you not to delay if you want the best home site in the city for a very little money; you can afford to pay more for a home than you would pay for an investment; you have both at these prices. After 10 days more the prices will be raised again.

TERMS NOT "CASH, OR SPECIAL," BUT, JUST EASY TERMS. However, you must live up to restrictions as heretofore advertised.

You will see that we are going to make this the most beautiful resident spot in Newport News. We will grade ahead of the other work for all who want to build at once. We do not promise what we will not perform; trust us to improve the property as we have advertised.

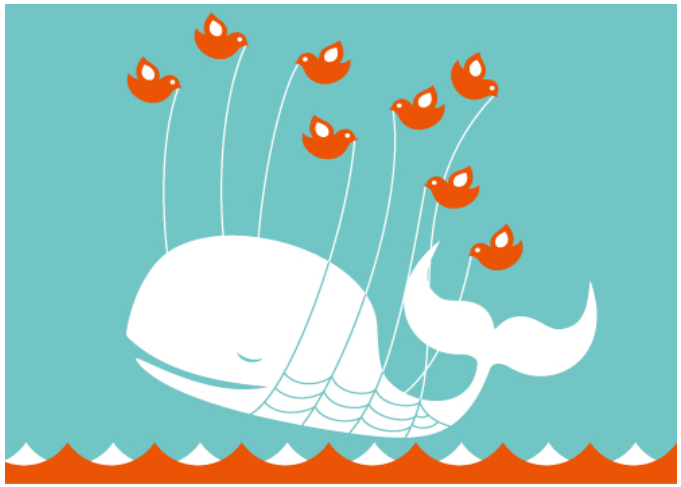
Why do we not hold the property if we think it is going to advance? Why is it that Farmer White, who owns timber land, does not convert his timber into lumber, or the lumber man his lumber into the various finished products and thus make all that there is to be made? Why does not the banker go into the grocery business and make 50 per cent. instead of the 10 per cent. sometimes charged on the money loaned the merchant? We divide our profits and you are to be the judge as to why we do it. Unless you wish to do so, after due consideration, do not take our advice, or that of anyone, given selfishly. If you need a home and expect to remain in this city, and if you are as able to own it as you are to pay rent in the best section, then take your 4 per cent. money out of bank and save the other 4 per cent. you give the landlord; use your own judgment.

Powell Trust Company, Inc.
2612 Washington Avenue, Newport News

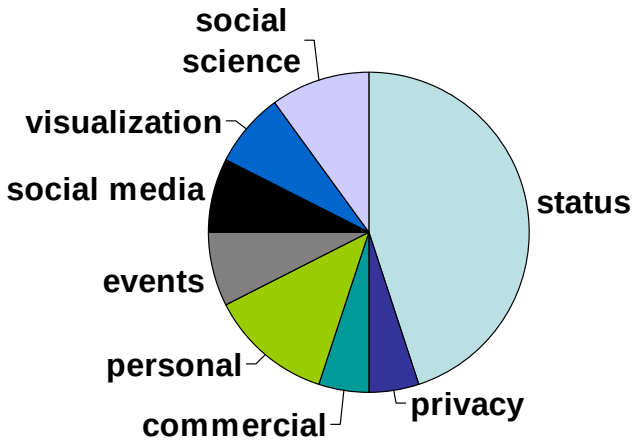
- The Chronicling America collection has 5 million page images from historic newspapers with OCR from organizations in 25 states.
- The site gets approximately 4 million hits per day.
- Some researchers want to search for stories in historic newspapers.
- Some researchers want to mine newspaper OCR for trends across time periods and geographic areas.
- Requests have come in to analyze all 5 million pages.

http://chroniclingamerica.loc.gov/

Case Study: Twitter



- The Twitter archive has 10s of billions of tweets in it.
- Research requests have included users looking for their own Twitter history, the study of the geographic spread of news, the study of the spread of epidemics, and the study of the transmission of new uses of language.



Are our institutions ready?

We are building large digital collections and must consider new ways in which they should be managed and used.

The Library of Congress is proceeding on multiple fronts

The development of a variety of repository services that will be used to ingest and inventory Big Data collections.

The ingest and inventory of such collections, **other than scale**, is basically understood.

How much ingest processing should be done with data collections, or collections that can be treated as data?

Do we process collections to create a variety of derivatives that might be used in various forms of analysis before ingesting them?

Do we have sufficient infrastructure to create full-test indexes for billions of files to support full discovery?

Do we load collections into analytical tools? These products are still in early days for the scale of billions of files.

LC will benchmark ingest and indexing processes in multiple hardware environments.

And what are the service models?

If we decide that we will simply provide access to data, do we limit it to the native format or provide pre-processed or on-the-fly format transformation services for downloads?

Can we handle the download traffic?

Can our staff develop the expertise to provide guidance to researchers in using analytical tools?

Or do we leave researchers to fend for themselves?

The Library is increasingly looking towards self-service – researchers need not ask to download or tell us that they have. We may never know.

BUT, we do have collections that are limited to on-site only access due to licenses or gift agreements. In that case, we may have to provide high-powered workstations with analytical tools for researchers to work with these collections and take analysis outputs away with them.

Both have policy implications and implications for public service staffing.

And now we will discuss…

Leslie Johnston
lesliej@loc.gov