Putting the Big in Big Data: How Re-Thinking the Appraisal of Electronic Records Can Facilitate Data Analytics in Digital Repositories

Bob Spangler, Leslie Johnston
NARA
April 26, 2016

“Big Data” and email as an exemplar

A slightly different perspective here as opposed to data sets or classic data mining in databases

A large corpus of accessioned and preserved email still is a lot of stuff, even if mostly textual, and applied “data analytics” (even though we’re not talking about raw data) is very valuable

For our purposes here, “big data” == “big content”

The implications of “Re-thinking” appraisal

What if we re-think (whether by design or compelled to do so) the appraisal of email (and perhaps by extension, other forms of electronic records or communications). What are the implications for digital preservation and for analysis of, and access to, the resulting data/content?

Why is this an issue for archives?

An archive, formally defined, is a collection of historical documents or records providing information about a place, institution, or group of people.

By the very nature of their jobs, archivists want to *carefully select*, prepare, describe, and provide access to such documents and records.

The mission of the National Archives is to preserve and provide access to the *permanently valuable* records of the United States, which further implies careful selectivity.

The problem with this needle in a haystack approach is:

The

The haystack is very big.

Should we try to make the haystack smaller?

In a world of big data, sophisticated analytical techniques, and cheap storage, is there value in deliberately and consciously keeping more rather than less?

To what degree does this change the fundamental mission or procedures of an archive?

How do we really define what is “historically valuable” information and do the “big data” capabilities change that definition?

Asked another way...

Again, using email only as an example, would there be value in saving more rather than less for purposes of data analytics, and thus re-considering what appraisal techniques are used for email?

Clearly, we are able to ask this question (for email and other record types) due to the fundamentally different nature of e-records and the capabilities they enable

Reality intrudes

We want to cherry-pick

But we may have to (or want to):

Why re-think email appraisal?

Why can't "better appraisal" make the haystack smaller?

At least in the Federal space, records management of email (which could and would lead to more organized, disciplined, and compact bodies of records) is not generally speaking a roaring success. This is due to:

- Technical impediments
- User work habits
- General misapplication of records management regulations and guidance

This has led in the past to uneven archival appraisal and records management practices often resulting in potential transfers of large, chaotic, bodies of email.

The difficulty of curating email

- Selection is difficult: what should stay and what should go?
    - Signal to noise: how much email of is important?
- Email can simply be a logistical mess: the “Here you go” syndrome
- Not enough eyeballs to process in traditional ways
- Reliance on artificial intelligence or automated categorization?

The classic example in automated categorization

From: Carol

To: Bob

Subject: Here you go

Date: April 23, 2016

Hey, Bob, did you say you wanted to go with us when we take the team out for pizza? Oh, and attached is the newly approved project plan you asked about. By the way, did you see what so-and-so was wearing today?

Where this is heading

Email is a very big part of the proverbial firehose of incoming electronic records that must be preserved.

Processing and management of very large bodies of email does not correlate well to traditional appraisal and e-record processing techniques such as careful culling, selection, and curating.

Maybe, ultimately, that could be a good thing.

NARA’s current appraisal methodology for email

In the interest of simplifying email appraisal and records management, NARA has recently revised its email appraisal methodology to the “Capstone” approach:

Rather than attempting to select, manage, and transfer for preservation by subject or content, agencies can now simply select the mailboxes of high-level officials for archiving.

This can have the effect of making the appraisal model more "casual" or less stringent, resulting in the potential for less carefully selected content, and as you might expect more transfer of email.

Could this be a good thing, even for an selective archive? In a corpus of records that can be so inherently chaotic, might we not receive added value by being less selective?

So: is more better?

Let’s consciously add more stuff: let’s put more of the “big” in big data.

If we can mine for content in the aggregate, e.g. word clouds, and in targeted data mining such as e-discovery, isn’t it better to have more raw material?

Would not abundant content always make our results “better”?

Asked another way....

Won’t large amounts of data (content) afford mining techniques much more power?

if you have a billion data points (email messages), as opposed to 100, aren’t outliers easier to classify and wouldn’t the underlying distribution of that content be more precise?

Caution: Potential for spurious conclusions: The more data there is, the more people can (somewhat ironically) cherry pick data points that confirm what they want it to show.

Applying analytics techniques to email

Familiar and accessible examples such as word clouds, tag clouds, etc. applied to textual content

Search engine full-text indexing, or metadata-based indexing, of “natural language” documents

Leveraging mature e-discovery techniques

Rich set of consistent metadata means that e-mail, as a “semi-structured” format, lends itself well to content mining and analytics

Practical concerns

Duplicates - is it ok to have a lot of duplicated material in an archive (which is typical of many email systems)

Volume - what does “cheap” mean really? (Hint: It doesn’t mean free)

Legal and statutory liability in keeping too much.

Lemons to lemonade? (obligatory cliche)

In summary, let’s explore the possibility of utilizing the inherent difficulties in managing, transferring, processing, preserving and providing access to email to deliberately and consciously keep more, utilizing the techniques of big data analytics to supply added archival value.

Next: Further perspective from Leslie Johnston