MAGIC, April 12, 2011
Session 2: Technical Hurdles, Research Solutions

What Will It Be?

Ken Thibodeau
Senior Guest Scientist
Information Technology Laboratory
National Institute of Standards and Technology

There are three basic questions we need to ask to judge whether information technology (IT) will create more hurdles or more solutions for journalists' access to government information: will IT make more government info available; will IT improve or facilitate access to that information; and will IT make it easier for journalists to use the records to develop their products?

✦ Will IT make more government information available?

- In one sense it will clearly do so, because the technology has, is, and will make it possible for government at all levels to create, acquire, and retain more and more information. No one has solid information on the quantity of government data, but 20 years ago the Congressional Research Service estimated that more than 90% of all information in the federal government was born digital. Several independent studies over the last decade have shown that digital information is growing very rapidly. There is no apparent reason to doubt that this has also occurred in the government sector.

- Another aspect of availability of information is its survival. A great deal has been said over the last few decades about the potential loss of digital information and the difficulty of preserving it in a way that supports an assertion that it has not been corrupted, but remains authentic. I have held for many years, and continue to hold that there are no proven methods for preserving over the long term the majority of types of digital data that are being created. But there is reason to be optimistic. Over the last 15 years, there has been a substantial increase in recognition of the importance of digital preservation accompanied by large increases in resources devoted to finding ways to accomplish it.

  But IT does improve the probability of survival of information at least in the short term. Following common practices in IT management -- not even best practices, just common ones -- IT reduces and can eliminate individual discretion in the destruction of information. That seems to fly in the face of frequent assertions of how easy it is to alter or delete digital data. But remember: Oliver North was able to shred his paper notes, but he was not able to touch the back up tapes that contained multiple copies of his notes in their original, digital form. Those digital records are in the National Archives.

White House email provides an illustrative case. As a result of a lawsuit brought by Scott Armstrong and other to force the government to preserve electronic records related to the Iran-Contra Affair, the National Archives received the email that was extant at the end of the Administrations of Presidents Reagan and George H. W. Bush, amounting to a few hundred thousand messages. As a continuing consequence of court decisions in that case, NARA received and preserves over 30,000,000 emails from the Clinton White House and over 200,000,000 messages from the last Administration. That's a thousandfold increase in two decades.

The George W. Bush email also illustrates the situation with regard to the second question: if the electronic records survive, are the accessible? In this case, the 82 Terabytes of electronic records of all sorts that NARA received from the last Administration have been indexed and are searchable on any term in their contents. Moreover, the overwhelming majority have bee associated with metadata and contextual information that supports faceted search.

✦ Will IT improve or facilitate access to government information?

There are both positive and negative responses to this question.

On the negative side:

- Legal barriers to access, such as privacy and national security, are implemented in computer systems;
- There are also practical reasons why computer systems are run in a way that excludes access by journalists or others outside the government, primarily because most government systems support the conduct of government business and outsider access would compete or conflict with that purpose.
- Finally, I've heard from some journalists that information that government is making available to them on the Internet, such as in Data.gov, is simply not relevant to their needs.

On the positive side:

- Given a right of access, IT offers increasing and increasingly sophisticated capabilities for finding responsive information
  - *Joan Hof: Nixon Correspondence*
- Current capabilities are from perfect:
  - String seacrches, viz. Google, Bing, et al, are not good either at returning only truly relevant items or at surfacing resources that could tell the whole story.
  - Computers can't yet deal with context; that's why, for example, IBM's Watson identified Toronto an American city.
- But things improving and we can expect search technology to progress substantially.

✦ Will IT make it easier for journalists to use the records to develop their products?

- In addition to the speed of computer processing, there is relevant research on the automated analysis & characterization of digital documents that should be useful to journalists. Some examples:
  - NARA and the Army Research Lab are supporting work by Dr. William Underwood at the Georgia Tech Research Institute towards automated characterization and description of records. Tools his team are developing can distinguish different types of documents, e.g., policy directives from correspondence, and different subtypes; e.g., letters and

memoranda within correspondence. Applying AI techniques to document content, the tools can tell what the records are about; e.g., this set of letters relate to personnel appointments, while those respond to enquiries from citizens, and a third set comprises correspondence with major stakeholders on an important legislative issue.

- Work done by Jorge Roman and Shelly Spearing at Los Alamos National Lab, automatically synthesizes knowledge about entire collections of electronic records; for example, summarizing how prominent themes in the content of a collection evolve over time.

- The National Science Foundation and NARA are supporting research by Dr. Maria Esteva and colleagues at the Texas Advanced Computing Center that is developing relatively simple, but interactive, visual display of information about what’s in collections of millions of electronic records.

- NARA is also looking backwards and supporting research at the National Center for Supercomputer Applications that aims at identifying reliable and economical means for converting old handwritten records into character encoded digital form. Once they’re converted, it will be possible to use the best available automated tools to explore and exploit them.