Slaying the dragons: What is at risk and how do we rescue it?
Sponsored by the NDSA Content Working Group
NDIIPP/NDSA Partners Meeting Workshop
July 20, 2011
3:15 p.m.

Presenters: Abbie Grotke (Library of Congress); Kristine Hanna (Internet Archive); Cathy Hartman (University of North Texas); Abby Smith Rumsey (NDIIPP/Library of Congress), facilitator

Attendees: 48

Highlights of the topics covered during the meeting.
The group discussed the goals of a proposed "adoption clearinghouse" for at risk data and how best to define at risk content. Four broad content categories were identified for discussion breakout groups: public/government; cultural heritage; news/event journalism; and data sets, directories and software. There was consensus that we need to reach out to communities that have or know about at risk data, including scientists, citizens, government agencies, artists and enthusiasts. This outreach should include raising awareness about digital preservation as well as education in best practices.

The goals of the session were to prepare for the creation of the clearinghouse by reviewing the suggested list of categories of at risk content and getting a sense of collecting priorities in the near- and long-term. The group divided into breakout sessions for each of the content categories in order to identify specific actors and institutions that are likely to collect and explore how the Content Working Group can connect with them. Different types of risk were discussed: economic, technological, legal, and orphaned content.

Librarians generally care about "expensive stuff," such as Elsevier, however publishers that can still make money from publications have a vested interest in preserving it. What's at risk are small humanities journals; if it costs as much to preserve the New England Journal of Medicine as it does to preserve a quarterly poetry journal, who will make it a priority to save that poetry journal? There's a question of economic incentive: the fact that something is expensive means it's not at risk, it has a sustainable business model that will help ensure ongoing interest and attention.

The group discussed the difference between content that faces technological danger (such as obsolete formats) versus content at risk of going away; can this distinction factor into the prioritization of archiving content? When organizations fail, the fate of their records are cast into question (for instance, startup ventures, magazines, websites). We should help organizations prepare for the hand-off, rather than wait for them to fail.

Some organizations involved in similar activities are a CODATA (International Committee on Data for Science and Technology) task group on data rescue for both analog and digital scientific data. It also has a clearinghouse activity underway to match

at-risk data with institutions that can preserve that data, but it is still in the early stages. Data-PASS (Data Preservation Alliance for the Social Sciences) at the University of Michigan is concerned with social science research data sets.

The NDIIPP Blue Ribbon Task Force on Sustainable Digital Preservation and Access identified different categories or risk: risks for technical purposes (due to obsolete, complicated, or proprietary formats); legal risks (who has the right to preserve it?), economic risks (who has the resources to preserve it?), organizational risks (organizations go away, reorganize), risk of ignorance (lack of expertise or best practices), and uncertainty of the long term value. There may be a misalignment between who decides content has value, who can pay for it, and who will be the steward.

Libraries don't collect based on categories of risk – acquisitions need to meet collection policy criteria, so we need to think in those terms.

Kristine Hanna discussed the work that the Content Working Group has done to determine content at risk in their monthly conference calls. She identified four broad content categories: public/government (local, state, and federal); cultural heritage (creative and unique websites); news and events (newspapers, digital citizen journalism, events, and political); and data sets, directories and software. The attendees broke out into discussion groups to focus on each category.

Action Items/Advice for Content Categories

State & local government: Require lifecycle management in state, local, county agencies; educate content creators. Concern about data on hard drives in offices; retention is required but there are no guidelines. Need more appraisal criteria for government information to see what is now missed. Rescue material sitting in offices (on floppies, personal hard drives, or other storage media that can be missed.)

Cultural heritage community: The clearinghouse can be viable and valuable if the tone is right; need to be seen as enabling the capture of at-risk content, not as supervisors or managers. Engage enthusiasts, subject matter experts, consumers, technologists. Define categories of resources: resources related to performance, to images, to creative files (such as literature). Reach out to communities to make sure they're doing a minimum of ground work that will allow us to capture their material someday. The need for community outreach is critical.

News and Events: The New York Times is not at risk. Priorities need to be on local news, newspapers, events (rapidly changing). Add a crowdsourcing component to the CWG clearinghouse to allow the public to nominate content. For archiving blogs, suggest a plug-in that would allow bloggers to self-identify as being willing to be archived. Make a creative commons license the default in Blogger.

Datasets and directories: A cultural change needs to occur: researchers are protective of their data. Journals are beginning to respond to NSF's requirement to have a data management plan, and are increasingly getting from their authors the data files behind the

tables/figures they publish (“raw” data vs. summary/derived data). Need to reach out to domain scientists to define what is at risk, and convince appraisal people that they need more than just summary data, get as close to raw data as they can.