Capturing Relevant Information for Digital Curation

Chirag Shah
School of Information and Library Science
University of North Carolina
Chapel Hill NC 27599, USA
chirag@unc.edu

Gary Marchionini
School of Information and Library Science
University of North Carolina
Chapel Hill NC 27599, USA
march@ils.unc.edu

Categories and Subject Descriptors: H.3.6 [Information Storage and Retrieval]: Library Automation

General Terms: Design, Human Factors, Management

Keywords: Digital curation, Digital Preservation, Contextual information

Digital curation primarily involves selecting, preserving, and insuring access to a repository of digital information. We argue that metadata and rich contextual information are crucial to long-term access to digital assets. The Vidarch Project[1] aims to develop policies and tools that help video curators discover and add contextual elements that will help future generations not only find but also make sense of video content.

A key question is - what context to include? We do not want to collect everything (*e.g.*, everything in the *London Times* the day the video was first released) and at the same time, we want to make sure that we do not miss the most pertinent contexualizing information. Instead of leaving this problem to the effectiveness of some automated process or as a burden to the curator alone, we suggest a hybrid approach. We propose to use four different aspects of relevance to help a curator determine what contextual information to include. These aspects along with the corresponding functional parts of the system are described below. An outline of the system is given in Figure 1.

1. *Algorithmic relevance:* computer programs based on some typical IR algorithms search the Web for contextualizing documents, news, images, or videos.
2. *Cognitive relevance:* the curator, uses his/her his background and knowledge to identify specialized databases and/or websites and the system searches these sources and returns candidate items.
3. *Situational relevance:* users tag information based on their own contexts. Not everything tagged is reliable, but the tags give the curator a good idea about how people relate information of various kinds.
4. These explicit user tags can be augmented by a fourth kind of relevance that combines machine processes and user activity. A monitoring component is embedded in the system that can monitor defined certain sources for events and news and user community activities (*e.g.*, search terms, click streams, links) and alert the curator based on the preferences given.

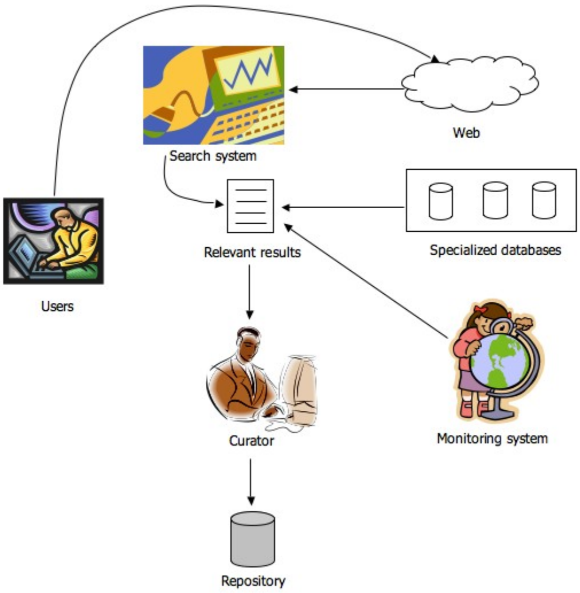

Figure 1: Digital Curation Architecture

The above proposed scheme may seem very ambitious, but we believe it is feasible. We have developed a prototype system that can search through specialized databases for metadata or go out to the Web in search for documents, news, images, or videos. The search results are presented to the curator and then left to his/her discretion on whether to add to the repository. At present we are focusing on digital video metadata. We have OpenVideo[2] and Prelinger[3] collections to our disposal. We will expand the present system with the following ongoing efforts:

- Provide the curator with a way to give preferences about specialized databases and sites.
- Determine how to assess the site metadata structure and harvest the best items.
- Determine how to incorporate user produced additional metadata such as tagging.
- Conceptualize the monitoring system within the context of implicit recommender system techniques.

[1] http://www.ils.unc.edu/vidarch/

[2] http://www.open-video.org/

[3] http://www.archive.org/details/prelinger

JCDL'07, June 17–22, 2007, Vancouver, British Columbia, Canada.
ACM 978-1-59593-644-8/07/0006.