Computer-Supported Elicitation of Curatorial Intent

Dagstuhl Seminar Proceedings 10291
Automation in Digital Preservation
Christopher A. Lee, University of North Carolina
September 20, 2010

> *"Too many miss how different architectures embed different values, and that only by selecting these different architectures—these different codes—can we establish and promote our values."*
>
> - Lawrence Lessig, 2006

Individuals across society are generating digital traces of their lives; and these traces can be stored, processed, combined and reused in numerous ways. Many cultural institutions – including libraries, archives and museums (LAMs) – have collecting missions that include personal papers, manuscripts and other non-institutional materials. There are unprecedented opportunities for a more inclusive social memory through the curation of digital materials of individuals from all segments of society.

Despite a massive increase in the volume and complexity of personal digital collections, research and literature designed specifically to guide this activity has been relatively limited. Information professionals will need to have tools and approaches to better acquire digital collections from individuals. Just as importantly, they must care for collections in ways that reflect the values – i.e. "characteristics or principles by which [they] consider something desirable or worthwhile" [9] – of relevant stakeholders. There has been considerable progress in recent years in the development of digital repository software but relatively little progress on methods for articulating or applying the requirements of creators, donors, and other individuals who are associated with collections.

In this document, I propose a program of research to investigate and address the above issues. I present a case for the development and testing of computer-supported mechanisms to elicit the curatorial intent of individuals in relation to digital objects being transferred to repositories.

When acquiring, managing and providing access to materials, professionals in collecting institutions must consider various norms, laws, codes of ethics, policies, procedures and personal values. As they address curation of digital collections, they are increasingly discovering "policy vacuums" [8], in which there is no existing guidance on new issues, and "latent ambiguities" [6] in established guidance. Digital resources are composed of interacting components that can be considered and accessed at different levels of representation (e.g. bitstream, through a filesystem; files as rendered through specific applications; records composed of multiple files; abstract "works"; larger aggregations such as web sites). To ensure integrity and future use, digital curation professionals must make decisions regarding treatment at multiple levels of representation. Figure 1 provides a graphic representation of these factors and considerations.

Dagstuhl Seminar Proceedings 10291
Automation in Digital Preservation
http://drops.dagstuhl.de/opus/volltexte/2010/2761

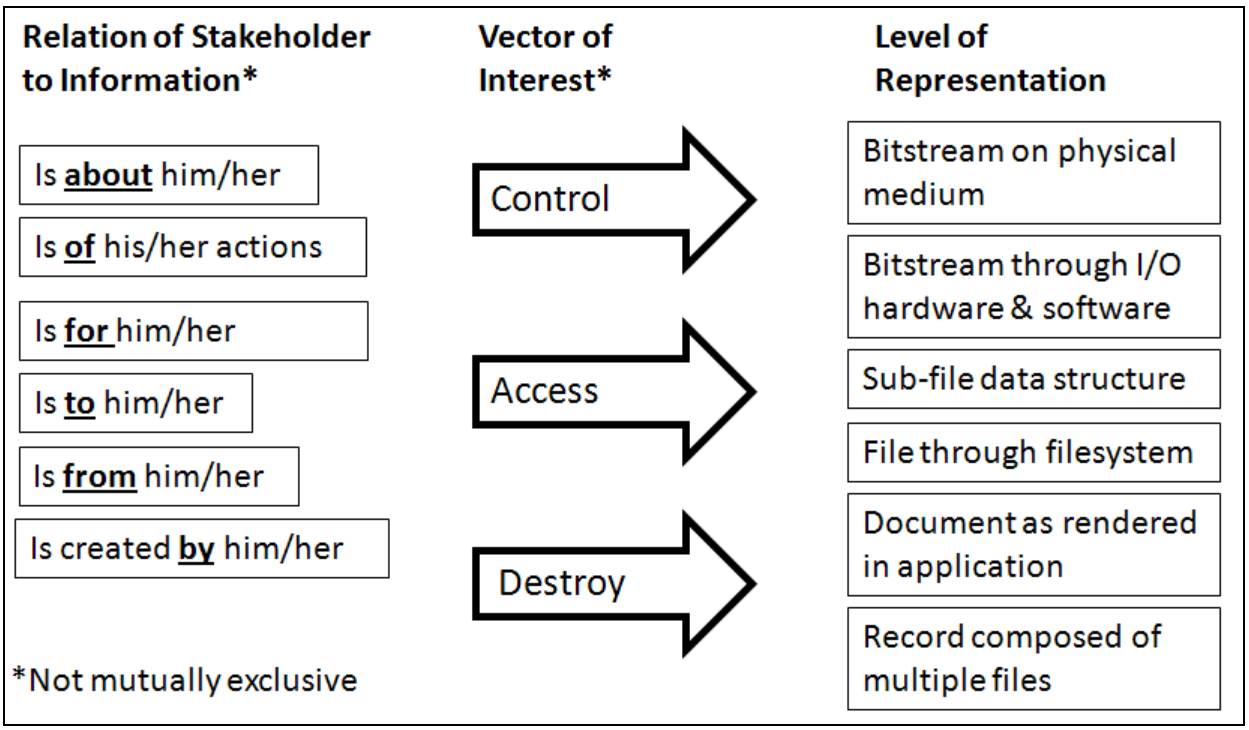

Figure 1 - Motivating Framework for the Ethics of Digital Curation

A stakeholder is a party who has a vested interest in how something is done. Professional decisions often hinge on identifying relevant stakeholders and determining what courses of action will best promote the stakeholders' interests. There are often many stakeholders with a legitimate interest in the ways that particular digital objects are managed, preserved and disseminated (see Figure 1).

The literature on digital archives tends to place a great emphasis on the "virtual" (i.e. intangible) nature of electronic resources. Computer systems have "an illusion of immateriality by detecting error and correcting it" [5], but it is essential to recognize that digital objects are created and perpetuated through physical things (e.g. charged magnetic particles, pulses of light, holes in disks). This materiality brings challenges, because data must be read from specific artifacts, which can become damaged or obsolete. However, the materiality of digital objects also brings unprecedented opportunities for description, interpretation and use. There is a substantial body of information within the underlying data structures of computer systems that can often be discovered or recovered. Recovery of data from physical media has been a topic of discussion in the professional library and archives literature for several years. There is also a large and quickly expanding industry associated with digital forensics, which focuses on the discovery, recovery, and validation of information from computer systems that is often not immediately visible to common users.

Digital curation professionals are faced with many new decisions, which require an understanding of underlying digital representations, in order to appropriately enact professional values and appropriately reflect the values of individuals. Consider the following exemplary cases:

- When acquiring a disk as part of a collection, should the curator create a bit-level image of the disk, in order to ensure the potential to recreate not only all the data but also system and program files?
- Should she retain "hidden" data in a Word document or only retain what she assumes to be the text that the author intended?
- If the disk includes a Microsoft Outlook .pst file (including saved and sent messages, calendar items, draft and deleted messages, address book, and possibly viruses), should she retain the whole .pst file,

or simply extract messages and attachments that were sent and received?

- If a collection documents the life of an individual, what should be the scope for collecting information associated with that person's online presence (e.g. postings, affiliations, profiles, micro-contributions)?
- If a repository routinely "normalizes" submitted files into designated file formats, under what circumstances will the normalization violate the intentions of creator or other interested stakeholders?

I propose the concept of "curatorial intent"[1] as a way of operationalizing the values of stakeholders. In order to enable specific digital curation practices, curatorial intent must ultimately be expressed in terms of policies, rules and services for the (1) production of significant properties [4,13] of digital objects and (2) control over access to particular components or levels of representation.

There is currently very little guidance available for answering the above questions. Collecting institutions traditionally rely on written donor agreements – when available – for determining the curatorial intent of donors. However, the language used in donor agreements is rarely specific enough to resolve specific technical questions about curatorial intent. What exactly did the donor intend to transfer; what types or level of representation are particularly sensitive to the parties represented in the materials; and how might the encoding or rendering decisions of a repository promote or violate the interests of relevant stakeholders? Digital repositories have developed numerous submission processes and procedures, with considerable attention to issues of data integrity and validation. Currently missing from these activities are: (1) a clear articulation of actionable values (i.e. values stated in ways that can guide specific digital curation actions) and (2) mechanisms for eliciting from individuals whether, how and to what extent those values are important to them in relation to the curation of their personal materials within collecting institutions. In short, professionals responsible for the care of personal digital materials, need to significantly expand and enhance the notion of a "donor agreement" in order to ensure that they are doing their jobs effectively, responsibly and appropriately.

A framework for addressing the ethics of digital curation should reflect both well-established professional values and the representational complexity of digital collections. In order to incorporate ethical considerations into information systems, it will also be important to create explicit policies, procedures and interfaces for eliciting, recording, implementing and testing the curatorial intent of stakeholders. For digital objects that are to be acquired and managed by collecting institutions, it is often appropriate to elicit curatorial intent when donations or submissions are being negotiated, defined and administered.

When taking custody of digital collections, it is important for LAMs to perform curatorial actions that are consistent, scalable and verifiable. Manual ad-hoc actions can be very expensive, error-prone, and it may not be possible to determine whether actions were carried out in specified ways. Rule-oriented data management is designed to address such issues through middleware that can free digital collections from many dependencies on specific hardware and software, in order to support preservation over time. It can also allow LAM professionals to translate human requirements into operations carried out by computers. They can first develop policies and then write sets of rules that will support those policies. The rules can be executed through the use of micro-services [1,11], which are simple programs designed to carry out specific computer operations. The rules can evolve, and they can be executed in different hardware and software environments over time. Implementing such a system requires detailed articulation of policies

[1] This concept draws significant inspiration from the writing of del Pazo et al [2] about "preservation intent," which they define as "a clear articulation of a commitment to preserve an object, the specific elements of the object [that] should be preserved, and a clear time line for the duration of preservation."

and rules, which could be prohibitively difficult for individuals who are submitting materials to a repository to carry out themselves.

A vital missing piece of the current suite of tools is the interface between individual creators and repositories. In order to apply policies that reflect the values and preferences of individuals, there must be interfaces for submission that elicit those values and preferences in ways that are approachable to end users. This process can then drive much more detailed policies and rules on the back end.

Before professionals can develop and implement systems that are attentive to the values of various stakeholders, they must clarify what those values might be and consider how the values could be enacted through policies, practices and systems. This will require further development of principles, policies and prototype user interfaces for the acquisition of born-digital personal archives into repositories.

An initial articulation of value statements can draw from several sources, including established taxonomies of values of individuals [12], codes of ethics from professional associations, values elicited from information policy investigations [9], and value types from the literature on value-sensitive design [3].

Some examples of such values are contextual integrity, autonomy, and authenticity. An example of policy could be that whenever a designated stakeholder (e.g. the donor of materials to an archive) is represented in an image, the repository should elicit at the time of acquisition the stakeholder's approval of any significant changes to the image that will be disclosed to the public. A set of rules would then enact this policy in specific ways, by eliciting, recording and executing the curatorial intent of the stakeholder. Note that there is no single or definitive translation of values into policies or policies into rules. A defining feature of policy development is that it attempts to reflect numerous (sometimes conflicting and sometimes complementary) values in ways that advance the interests of relevant stakeholders.

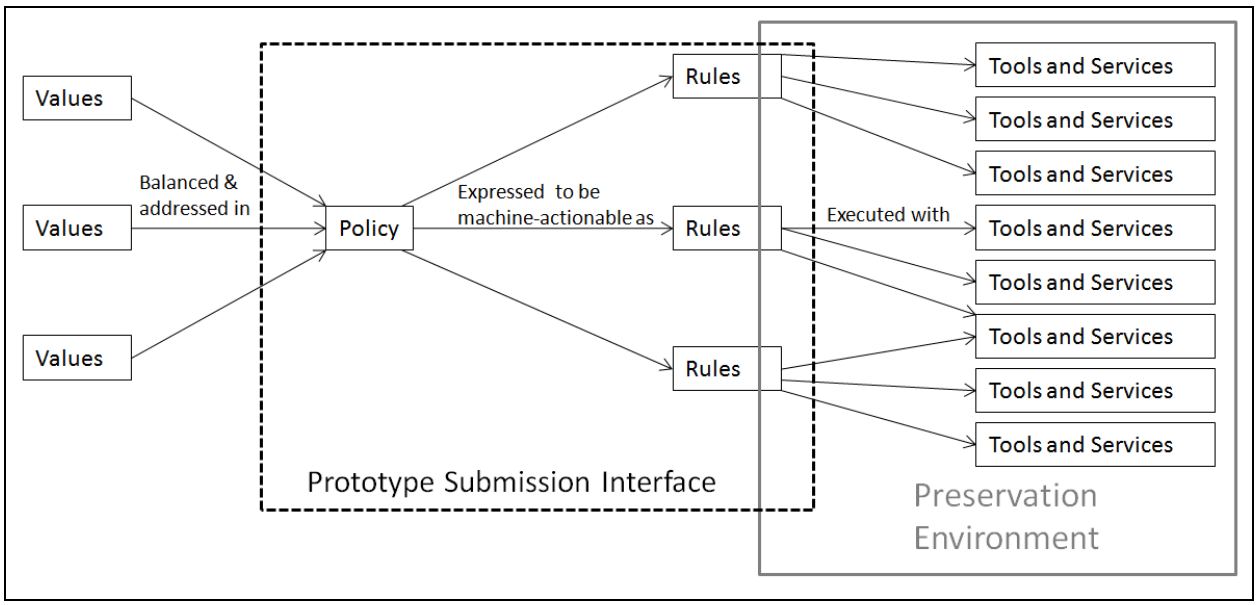

Figure 2 - Role of Submission Interface in Value-Enabled Curation

One objective of this line of research and development is an interface that allows individuals to express their preferences (curatorial intent) in ways that are approachable and meaningful to them, but which are translated by the interface into instructions that are machine-actionable on the back-end (see figure 2). Specifically, the interface could attempt to elicit curatorial intent from users in several different ways:

(1) by allowing them to express their intent using "controlled natural language" [7,10], which is a form of expression designed to be approachable to non-programmers but also easily translated into software instructions than everyday language,
(2) presenting questions to answer through standard user interface elements, including radio buttons and check boxes, and
(3) collecting reactions to examples presented to them, e.g. a Word document with change tracking revealed; view of a file through a Hex editor to reveal underlying information; Macintosh resource fork and Windows Registry information associated with files; a document normalized into PDF-A (a format designed for preservation, which often removes various features of the original document).

Preservation of digital collections requires numerous forms of automation, in order for actions to be scalable, accurate and consistent. I recommend an active program of research and development to ensure that digital preservation efforts can systematically advance individual and professional values.

References

[1] Abrams, Stephen, John Kunze, and David Loy. "An Emergent Micro-Services Approach to Digital Curation Infrastructure." *International Journal of Digital Curation* 5, no. 1 (2010): 172-86.
[2] del Pozo, Nick, Andrew Stawowczyk Long, and David Pearson. "'Land of the Lost': A Discussion of What Can Be Preserved through Digital Preservation." *Library Hi Tech* 28, no. 2 (2010): 290-300.
[3] Friedman, Batya, Peter H. Kahn, and Alan Borning. "Value Sensitive Design and Information Systems." In *Human-Computer Interaction and Management Information Systems: Foundations*, edited by Ping Zhang Dennis Galletta, 348-72. New York, NY: M.E. Sharpe, 2006.
[4] Hedstrom, Margaret, and Christopher A. Lee. "Significant Properties of Digital Objects: Definitions, Applications, Implications." In *Proceedings of the DLM-Forum 2002, Barcelona, 6-8 May 2002: @Ccess and Preservation of Electronic Information: Best Practices and Solutions*, 218-27. Luxembourg: Office for Official Publications of the European Communities, 2002.
[5] Kirschenbaum, Matthew G. *Mechanisms: New Media and the Forensic Imagination*. Cambridge, MA: MIT Press, 2008.
[6] Lessig, Lawrence. *Code: Version 2.0*. New York, NY: Basic Books, 2006.
[7] Macias, Benjamin, and Stephen G. Pulman. "A Method for Controlling the Production of Specifications in Natural Language." *The Computer Journal* 38, no. 4 (1995): 310-318.
[8] Moor, James H. "What Is Computer Ethics?" *Metaphilosophy* 16, no. 4 (1985): 266-75.
[9] Overman, E. Sam, and Anthony G. Cahill. "Information Policy: A Study of Values in the Policy Process." *Policy Studies Review* 9, no. 4 (1990): 803–18.
[10] Power, Richard, Donia Scott, and Roger Evans. "What You See Is What You Meant: Direct Knowledge Editing with Natural Language Feedback." In *Proceedings of the 13th Biennial European Conference on Artificial Intelligence*, edited by Henri Prade, 677-681. New York, NY: John Wiley & Sons, 1998.
[11] Rajasekar, Arcot, Michael Wan, Reagan Moore, Wayne Schroeder, Sheau-Yen Chen, Lucas Gilbert, Chien-Yi Hou, Christopher A. Lee, Richard Marciano, Paul Tooby, Antoine de Torcy, and Bing Zhu. *iRODS Primer: Integrated Rule-Oriented Data System*. San Rafael, CA: Morgan & Claypool, 2010.
[12] Schwartz, Shalom H. "Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries." *Advances in Experimental Social Psychology* 25 (1992): 1-65.
[13] Yeo, Geoffrey. "'Nothing Is the Same as Something Else': Significant Properties and Notions of Identity and Originality." *Archival Science* 10 (2010): 85-116.