2010 Chicago Colloquium on Digital Humanities & Computer Science: Poster Abstract

# Data Curation Education for the Humanities: Principles & Challenges

Allen H. Renear, Trevor Muñoz, Kevin Trainor
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

September 2010

## Background

Data curation has been described as the active and ongoing management of data throughout its entire lifecycle of interest and usefulness to scholarship. Curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time [1]. Originally conceptualized as an e-Science problem precipitated by large amounts of data in digital formats, data curation is an emerging problem for the humanities as well, as both data and analytical practices become increasingly digital. Research groups working with cultural content, as well as libraries, museums, archives and other institutions, are all in need of new expertise.

## The Program

To prepare information professionals for the unique challenges of humanities data in digital formats, the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign received a grant from the Institute of Museum and Library Services (IMLS) to extend our existing Data Curation Education Program (DCEP) [6] to include humanities data. Among the activities underway are needs analysis studies, curriculum design, case study development, a fellowship program, an internship program, a summer institute for in-service training, and the development of a framework of cross-disciplinary data curation concepts.

## Principles

**The curation of cultural information has much in common with scientific data curation—and it has distinctive features as well.** Humanities data curation can take advantage of the theories, tools and practices being developed within the larger scientific data curation community. However we must also accommodate the distinctive features of humanities research, and recognize the unique characteristics of humanistic data that derive from the ineluctable intentionality of cultural artifacts.

**The humanities have already evolved sophisticated curation practices; this rich tradition, analyzed and understood, must inform the development of the data curation curriculum.** Data curation has always been at the heart of humanistic research, in fact if not in name. Developing a data curation curriculum involves, in part, identifying and understanding long-standing principles and practices. At the same time, the new digital context challenges the methods and techniques with which we pursue traditional curatorial objectives such as authenticity, provenance, authoritative reference, and annotation—and even challenges our understanding of those objectives. The need for improved foundations is urgent.

**Many professions and disciplines contribute to the development of data curation—however library and information science provides an overarching framework.** Library and Information Science (LIS) is a large, well-established discipline with extensive research programs in the areas of data representation and retrieval, user communities and their information behavior, and collection and service development and management. Its focus on research-based support for use and users of information, and its long-standing involvement with information organization, storage, and retrieval, data formats, metadata, and access provides the needed context for the development of data curation education [4].

## Organization

Data Curation is a specialization within the GSLIS Masters program, coordinated from within the GSLIS Center for Informatics Research in Science and Scholarship (CIRSS). The project to extend the specialization to include humanities data has an advisory board: Lorcan Dempsey, Vice President and Chief Strategist, OCLC (Online Computer Library Center); Gregory Crane, Editor-in-Chief, The Perseus Digital Library, Tufts University; Julia Flanders, Director, The Brown University Women Writers Project; Harold Short, Director, Centre for Computing in the Humanities, King's College London; Daniel Pitti, Associate Director, Institute for Advanced Technology in the Humanities, University of Virginia; Christian-Emil Ore, Director of Research, Unit for Digital Documentation, University of Oslo.

# Activities

The curriculum and our summer institute for in-service education have been described elsewhere [5]. Here we present two activities that may be of particular interest to DHCS participants.

## Needs Assessment

To ensure our curriculum includes relevant skills we are carrying out two needs analysis projects. Preliminary results will be available soon. *Survey/Interviews:* We are conducting a survey and structured interviews with management and professional staff at selected humanities computing centers to identify problems faced in data management, as well as current best practices and future needs for data expertise. Humanities computing centers were chosen over libraries, institutional repositories or campus IT groups because, at the current time, the greatest level of relevant experience still seems to be concentrated in these settings. *Position Descriptions Analysis:* With another DCEP project (led by Melissa Cragin) we are assessing job postings in the sciences, social sciences, and humanities for skills and education deemed relevant to data curation by potential employers [2].

## Best Practices Guide

All of these experiences are being reflected in the development of an online guide to data curation research and best practice, serving humanities scholars, project managers, and information professionals. Exploiting contemporary social media, the guide will incorporate user-contributed commentary and will, we hope, provide sustainable, synoptic access for anyone looking for authoritative current information on humanities data curation best practice.

# Challenges

Some critical challenges remain—topics that have been overlooked by other treatments of humanities data curation but which we feel are exceptionally important.

**Levels of Abstraction:** Every humanist is familiar with distinctions such as work vs. text, text vs. edition, and edition vs. copy, even if it is debate-able what these distinctions mean and how many are needed—or even if they are possible. Curators of humanities data cannot avoid decisions about how to understand and represent such notions and how indicate these relationships amongst datasets (e.g., a more precise version of: *these two files contain the same text in different formats*). This will include accommodating such things as variation in XML languages and encoding conventions. We are collaborating with an NSF-funded project also doing work in this area and expect our results to make connections with the influential *Functional Requirements for Bibliographic Records* (FRBR) model of the documentary universe [3].

**Levels of Interpretation:** One important notion that appears to exist in both the sciences and the humanities is the distinction between data that is relatively raw or in some sense accepted as given, and data that is "processed" or the result of interpretation and analysis. We have had some success finding intuitive alignments between the widely used NASA data level categories and traditional notions of levels of editorial intervention found in textual criticism. This suggests that the development of common frameworks of curatorial concepts and terminology across disciplines is possible. However, neither NASA nor the literature of textual philology provides an adequate conceptual (as opposed to merely operational) account of what data levels are, focusing instead on what features should be at what level. We suspect this will be a problem for other alignment projects as well.

**Unique Features of Cultural Data:** Although there are similarities in nature between cultural and natural objects, there are also fundamental differences. These differences account for differences in scholarly analysis, and these differences in turn have significance for data curation. Cultural objects, such as texts, are fundamentally constituted by social practices, including the attitudes, intentions, and affective responses, of the persons and societies that create and consume these objects. At the same time, since cultural objects are available to us "from the inside", so to speak, we can reason about them in a different way than we can reason about natural objects. Indeed humanists mean something different by "understanding" than natural scientists mean. Currently these differences and their significance remain poorly understood, resulting in a vague but undeniable sense that humanities data curation practices, which are influenced by natural science, may not yet adequately accommodate cultural information.

## References

[1] Melissa H. Cragin, P. Bryan Heidorn, Carole L. Palmer, and Linda C. Smith. An Educational Program on Data Curation. Poster presented at the Science and Technology Section of the annual American Library Association conference, Washington, D.C., June 25 2007.

[2] Melissa H. Cragin, Carole L. Palmer, Virgil Varvel, Aaron Collie, and Molly Dolan. Analyzing Data Curation Job Descriptions. Poster presented at the 5th International Digital Curation Conference, London, December 2–4 2009.

[3] International Federation of Library Associations. *Functional Requirements for Bibliographic Records: Final Report*. UBCIM Publications—New Series Vol. 19. K. G. Saur, München, 1998.

[4] Carole L. Palmer, Allen H. Renear, and Melissa H. Cragin. Purposeful Curation: Research and Education for a Future with Working Data. In *Proceedings of the 4th International Digital Curation Conference*, Edinburgh, Scotland, December 1–3 2008.

[5] Allen H. Renear, Molly Dolan, Kevin Trainor, and Trevor Muñoz. Extending an LIS Data Curation Curriculum to the Humanities: Selected Activities and Observations. Poster presented at the iSchools conference, Champaign, IL, February 3–6 2010.

[6] Allen H. Renear, Lauren C. Teffeau, Patricia Hswe, Molly Dolan, Carole L. Palmer, Melissa H. Cragin, and John M. Unsworth. Extending an LIS Data Curation Curriculum to Include Humanities Data. Poster, DigCCurr 2009 conference, April 1-3 2009.