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, Urbana-Champaign
501 E. Daniel Street
Champaign, IL 61820
{renear,munoz14,trainor1}@illinois.edu

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 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.

Data curation

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].

DCEP: The Data Curation Education Program at Illinois

Initiated with funding from the Institute of Museum and Library Services (IMLS), DCEP is a specialization within the ALA-accredited master's program at the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign.
DCEP Principal Investigator: Melissa Cragin.

DCEP is coordinated from within the Center for Informatics Research in Science and Scholarship (CIRSS). CIRSS Director: Carole Palmer.

DCEP-H

To prepare information professionals for the unique challenges of humanities data in digital formats, GSLIS received another grant from IMLS to extend the existing Data Curation Education Program (DCEP) [5,6] to include humanities data.
DCEP-H Principal Investigator: Allen Renear; Coordinator: Kevin Trainor.

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.

DCEP-H Principles

The curation of cultural information has much in common with scientific data curation, but 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 it 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. 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.

Many professions and disciplines contribute to the development of data curation—however library and information science (LIS) provides an overarching framework.

LIS is a 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 data curation [4].

Challenges Encountered

Levels of Abstraction

Distinctions such as work vs. text, text vs. edition, and edition vs. copy are familiar, if problematic [3]. Troublesome as they are we must further elaborate these distinctions in order to answer questions such as:

1) When do two files contain different but equivalent representations of the same textual information?

2) When do two different files encode, differently, the same representations?

3) How do we understand identity and change in the digital world?

To discover what a complete account of such distinctions might be we are collaborating with an NSF-funded project carrying out similar work on scientific datasets. [7]

Levels of Interpretation

An important distinction in both the sciences and the humanities is between relatively "raw,"or in some sense given data, and data that is "processed" or the result of interpretation and analysis. A common framework for levels of this sort would be an advantage for the data curation curriculum as well as curatorial policies.

We've found intuitive alignments between the widely used NASA data levels and traditional levels of editorial intervention in textual criticism [8]. This suggests that cross-disciplinary frameworks of curatorial concepts are possible.

But neither NASA nor, for the most part, textual philology provides an adequate conceptual (vs. operational) account of what data levels are, focusing instead on what features should be at what level, rather than the principle for making the selection.

Unique Features of Cultural Data

There appear to be fundamental differences between natural and cultural objects. Cultural objects are constituted by social practices, including the attitudes, intentions, and affective responses, of the communities that create and consume these objects and so are fundamentally communicative in a way that natural objects are not.

In addition, since cultural objects are available to us "from the inside," so to speak, we can reason about them differently. Indeed the humanist's notion of "understanding" seems different from that of the natural scientist.

These differences are poorly understood, resulting in the vague yet undeniable sense that humanities data curation practices, where they are influenced by natural science, do not yet adequately accommodate the *intentionality* of cultural information.

Needs Assessment

To ensure our curriculum includes relevant skills we are carrying out two needs analysis projects.

Survey/Interviews: We are conducting a survey and structured interviews with management and professional staff at selected humanities computing centers to identify problems 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 relevant experience still seems concentrated in these settings.

Position Descriptions Analysis: We are assessing job postings in the sciences, social sciences, and humanities for skills and education deemed relevant to data curation by potential employers in coordination with another DCEP project (led by Melissa Cragin) [2].

Best Practices Guide

These experiences are being reflected in an online guide to data curation 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.

INSTITUTE of Museum and Library SERVICES

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

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.
[7] Sayeed Choudhury and Robert Hanisch. Data Conservancy: Building a Sustainable System for Interdisciplinary Scientific Data Curation and Preservation. Paper given at the Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data (PV) Conference, Madrid, Spain. December 1-3, 2009.
[8] Allen H. Renear, Molly Dolan, Kevin Trainor, and Melissa H. Cragin. Towards a Cross-Disciplinary Notion of Data Level in Data Curation. In Proceedings of the 72nd ASIS&T Annual Meeting. Vancouver, BC, November 8-11, 2009.