DIGITAL HUMANITIES
DATA CURATION

An Introduction to Humanities Data Curation

Julia Flanders, Center for Digital Scholarship, Brown University
Trevor Muñoz, University of Maryland

§ 1 What is data curation?

At present, there are a number of competing terms used to describe the activity of managing digital materials for research: digital curation, digital stewardship, data curation, digital archiving. There is overlap among these definitions or visions. The variation that does exist is due to more than confusion or carelessness. Each of these terms has significant connotations and attempts to align the relatively new activity of caring for digital materials with an older tradition, discipline, or profession.

The term and the idea of "curation" is experiencing a moment of cultural cachet that simultaneously threatens to empty it of meaning. Nonetheless, curation is a strong and suitably-encompassing term for the practices we want to describe. When we speak of "curation", what set of activities do we have in mind? Etymological guidance from the word's root meaning and early usage gives us the idea of curation as healing, and of the curator as one who "has the care or charge of a person or thing", a "guardian" (OED). As the word has emerged in the contexts of museums and also rare book and manuscript librarianship, "curation" also carries this dual emphasis: on protection, but also on amelioration, contextualization, and effective exposure to an appropriate set of users.

Digital curation: the curation of digital objects and information might be straightforwardly derived from curation as described above, but "data curation" shifts the emphasis and brings an additional consideration into play. The curation of research data – raw and abstracted material created as part of research processes and which may be used again as the input to further research – carries with it the burden of capturing and preserving not only the data itself, but information about the methods by which it was produced. If the methods used to generate the data are algorithmic, the method itself may need to be captured and curated. Because these methods and information on the goals involved in creating the data are often essential to its subsequent interpretation and reuse, they can be considered an important part of the data itself. In addition, because reuse is such a crucial aim, successfully curated data needs to remain functional, and this may require regular changes to its state. At the current time, a key aspect of humanities data curation is thus to ensure that the representations of objects of study in the humanities functions effectively as data: that they are processable by machines and interoperable such that they are durably processable across systems and collections whiles still retaining provenance and complex layers of meaning.

As a compact and provisional definition, we might therefore start by saying that data curation is “the active and ongoing management of datas throughout its entire lifecycle of interest and usefulness to scholarship” (Cragin et al. 2007). To expand on this definition, it may be helpful to gloss some of these terms in more detail:

- **active and ongoing management**: Datas curators intervene in the research process in order to translate or migrate data into new formats, to enhance it through additional layers of context or markup, to create connections between data sets, and to otherwise ensure that data is maintained in as highly-functional a form as possible.
- **entire lifecycle**: As we enter the era of thoroughly digital research, the full lifecycle of digital research data is still not yet known to us. However, we can anticipate that some data (particularly data collected through destructive means, such as archaeological data) will have a very long horizon of usefulnesss (in addition to increased evidentiary value for historical analysis and stewardship of our cultural heritage). The uses of data will likely change over time and with different stages of research.
- **interest and usefulness to scholarship**: The term “scholarship” should be construed broadly, especially since data creation, use, and curation are not limited to the academy. Data curation seeks to retain the interest and usefulness of any data that has a serious purpose to fulfill. Furthermore, even for data created and curated *within* the academy, it is worth remarking that there may be potential user communities *outside* the academy that could provide valuable motivation and even resources for the curation process. Especially given the recent and increasing emphasis on crowd-sourcing, community-driven data, and the permeability of the boundary between the academy and the public sphere, it is reasonable to anticipate that definitions of “scholarship” may continue to broaden rather than narrow over time.

Data curation is, as Terry Cook has written of archives, both as scholarly research area and a praxis and, increasingly, data curation too uses a distinct methodology to advance itself as both a discipline and a professions (in Gilliland-Swetland 2000). Understanding data curation in this dual context means recognizing the range of different kinds of training and expertise upon which it draws, including:

- expertise in humanities subject disciplines
- expertise in library and information science
- expertise in archival science
- expertise in computer science
- expertise in systems/records management

Understanding data curation in a practical context also means knowing how curation intersects with a number of specific actions and processes, including:

- **description**: Documentation of the context of data and the relationships between different forms and functions of research data and its analysis.
- **annotation**: Annotation includes a number of activities that add further information to the data, either in the form of additional markup that identifies data structures at a finer level of detail, or in the form of added information that contextualizes or glosses parts of the content (such as the identification of named entities).
- **collection/aggregation**: Collection refers to the creation of meaning through connecting functions of institutions, projects, or teams to the data they create, or also to the role of some subset of data to a larger information ecosystem.
- **storage**: Storage here stands in for a number of activities related to the encoding of data to physical media and the provisioning and maintainence of concrete systems that underly stability and basic accessibility of data.
- **migration**: Because of constant changes to data formats and data representation standards, long-term data curation necessitates regular migration of data to keep it in formats that can be read and used. Although strategies exist (such as emulation or preservation of original hardware and systems) that permit data to be used in its original form, migration has additional value in allowing data formats to be harmonized, so as to permit more effective usage across data sets.

Resources: Leading Organizations

Digital Curation Centre: Digital Curation Centre.
The Digital Curation Centre (DCC), based in the U.K., is invaluable, first-stop resource

for learning about the major issues in data curation. The DCC mission is to provide expertise and best practice documentation to research communities across higher education. From this site, one can access research on data curation, a more-broadly-targeted guide to best practices, information about upcoming events, and also tools for activities such as data management planning. The primary audience for the DCC is U.K.-based but much of the information foud here is broadly applicable.

National Digital Information Infrastructure and Preservation Program (NDIIPP): Library of Congress.
This is again a compendious resource for information on digital preservation and stewardship. Produced by a collaborative organization hosted and led by the Library of Congress, this site offers case studies, information about events, access to preservation-related standards hosted by the Library of Congress, and an accessible and interesting blog. The beginner can find many useful starting points for further learning here.

Resources: More on Data Curation

Bailey, Charles W. Jr. *Digital Curation and Preservation Bibliography*.

A selective but massive bibliography compiled by Charles W. Bailey, Jr, this resources presents more than 500 digital curation and preservation-related articles, books, and report. This tremendous resource is organized into chapters on subjects such as “copyright”, “formats”, “metadata”, and other core topics. The resource is available as a free, open-access PDF file but may also be purchased in print from a number of worldwide retailers. Bailey’s bibliography provides citations, but due to its size, resources are not annotated. Still, this remains an excellent place to survey much of the available research on a topic related to data curation.

Lord, Philip, Alison Macdonald, Liz Lyon, David Giaretta. *From Data Deluge to Data Curation*.

This influential early paper describes the challenges of managing large amounts of data in digital form and lays out the basic shape of data curation as a response. Becoming familar with this paper will help newcomers orient themselves to the origins of the field—both intellectual and practical. The survey of researcher awareness of data curation needs that is the ostensible subject of this paper led to the formation of the Digital Curation Centre (above).

Ashley, Kevin, editor. *International Journal of Digital Curation*, Vol. 6, No. 2 (2011).

The second number of volume six of the International Journal of Digital Curation offers an excellent overview of the field of data curation through a combination of peer-reviewed papers and topical articles. For those new to the field, browsing through the archives of this journal will form an excellent introduction to both the practice and theory of curation.

§ 2 Unique Features of Humanities Data Curation

In extending our understanding of data curation from thessciences to the humanities, one of the significant challenges is the specific forms of data the humanities disciplines present. By this we do not mean simply data types, although it is worth noting at the outset that humanities data includes (potentially) almost any data type. But in addition, humanities data is presented in specialized aggregations that themselves have significance for understanding, using, and curating the data. Some of these aggregations are digital extensions of long-standing traditional forms: for instance, finding aids, concordances, and scholarly editions, which have a long analog history. Others, like the thematic research collection or digital text corpus, are products of new digital research methods (though they also respond to discipline-specific research needs with a pre-digital history).

Although the field of humanities data is growing steadily, at present we can identify several major types of research objects and collections that present distinctive forms of data and distinctive curation challenges.

- **Scholarly editions**: Again, a high degree of formal structure but with very substantial variability, arising both from disciplinary differences (e.g. documentary editing vs. critical editing vs. genetic editing) and from smaller variances of practice to handle the exigencies of individual cases. Issue of whether there’s any possibility of cross-edition synthesis: do curators treat each edition as a distinct entity? Also, because each edition reflects individual editorial perspectives and expertise to such a degree, effective curation requires highly specialized knowledge. A community of editorial practice has arisen around the use of the TEI Guidelines, which supports the use of systematic encoding of information about individual witnesses, variant readings, and editorial process. However, digital scholarly editions may also be constituted as collections of page images with metadata and annotations (which may or may not themselves contain usefully structured information), or (less rigorously) as web pages representing the text and apparatus of the edition. In cases of the latter kind, the scholarship represented in the content and editing of the text may be of a very high order, even though the digital representation does not use comparably rigorous methods, and in these cases the task of the curator may be to migrate the edition into more robust and maintainable formats.

- **Text corpora**: Because they are among the earliest applications of digital technology to humanities research, text corpora have a long history which also reflects the use of a variety of standards and digitization approaches. The corpus may contain internal markup representing linguistic, grammatical, vocal, or semantic categories, and will probably also contain metadata (at the corpus and record level) documenting the principles of corpus construction (e.g. sampling and segmentation methods, normalization) and also the relevant facts about each item in it (e.g. demographic details about speakers, bibliographic information about sources). From a curation standpoint, it may be desirable to harmonize corpus metadata to permit cross-corpus analysis.
- **Text with markup**: Structured text, and especially documents in which XML markup has been used to represent document structure and content, is becoming an increasingly important part of digital humanities research. In the simplest cases, the markup is confined to simple metadata, but XML languages (especially those like TEI which are intended for scholarly use) are becoming more widely used to represent the details of document structure, editorial annotations, and also aspects of content such as named entities, intertextual references, and thematic or interpretive information. Marked-up text presents some special features of relevance to the data curator, since the markup typically reflects (explicitly or implicitly) a specific editorial frame of reference that itself may be of importance for subsequent re-use of the data. They also present challenges because of the wide variation in how particular markup schemes are applied.
- **Thematic research collections**: where text data, images, and contextual information are bound together in highly structured ways. Such collections can be treated, for some purposes, as a simple aggregation of individual assets (XML files, image files, etc.) but the organization of the collection and the editorial logic that is represented in ancillary materials such as stylesheets and configuration files is likely to be extremely significant, not only in making sense of the collection but also in recovering the editorial intentions and decisions that constitute it as scholarship.
- **Data with accompanying analysis or annotation**: In many cases, humanities data takes the form of a “primary” digital object (such as an image, a map, a virtual 3-D reconstruction) which has been enhanced with annotations or analysis. Annotations may be associated with specific regions of the object, either taking advantage of specific anchors in the data, such as ID values on XML elements, or using external pointing schemes such as the identification of regions within an image by offsets from the origin. In either case the association may be vulnerable to disruption if the target file changes. In this category we can also include scholarly editions that consist of a text annotated with variant readings, emendations, and other editorial

information. With data of this kind, it may be useful to consider the “primary” object as having research value that extends beyond the original intentions of the creator, and apart from the annotations and analysis that express those intentions.

- **Finding aids** and other information maps, such as bibliographies: high degree of formal structure; strongly codified by history of practice, but starting to evolve in light of digital affordances, leading to discrepancies (e.g. between EAD records created from traditional finding aids and those “born digital”) that may need to be remedied in the interest of long-term function.

In addition to these distinctive kinds of humanities data, there are also a few strategic points concerning the treatment of this data that need to be stressed:

- The importance of **interpretive layering**: While data in non-humanities disciplines clearly carries an interpretive framework with it, in the humanities the interpretation is in some cases the primary object of interest, not just a perspective on the data that can be separated from it. For instance, in a digital edition, the source document on its own may be of comparatively little importance (being readily available from other sources) but the compilation of variant readings, emendations, and commentary make the edition distinctively valuable. Similarly, a thematic research collection whose texts are encoded in TEI/XML may include layers of interpretive information: in the markup itself (for instance, in the way that genre is characterized, or in the identification of narrative turns), in the metadata (which may carry details of editorial principles), in the annotations and commentary, in the stylesheets that determine how the collection appears to a user and which textual possibilities are expressed or suppressed, and in the interface design (expressed in programming or configuration files) that opens up or forecloses specific usage options.
- The importance of information about **how the data is captured and prepared**: This information is also crucial to curation of data in the sciences. In the humanities, the details are quite different and the significance of some operations may not be as obvious. Crucial decisions affecting the usability and meaning of humanities data are made at each stage of data creation and management, and the documentation concerning these decisions is likely to be valuable to both future users and curators of the data. These decisions may include the choice of source (the edition or in some cases the specific copy that serves as the basis for the digital object), the calibration of instrumentation (including color balancing, lighting, camera settings, sampling rates), the method of data capture (such as OCR, double-keying, etc.), the details of transcription (including regularization, handling of illegibility and uncertainty, handling of special characters), the details of the encoding scheme used (including whatever extensions or customizations have been made to it), the level and type of

quality assurance testing and error correction, the kinds of editorial oversight that have been exercised, and the details of any subsequent curatorial activity.

- The importance of capturing **responsibility, editorial voice, and debate**: As noted briefly above, in humanities data the interpretive layers that accompany and contextualize the base data may be as important as the data itself. These layers represent scholarly agency and as a result are subject to debate; indeed, as digital humanities scholarship matures, such debates are an increasingly interesting aspect of the data. Some digital representation systems (notably the TEI) already include explicit provision for documenting responsibility in a fine-grained manner, and also for representing certain kinds of editorial debate and disagreement. These affordances are translations or reconceptions of data curation practices from the print tradition of the humanities. Digital research collections may also represent responsibility and debate in ad hoc ways, through annotation or links to other resources that offer an alternative view of the data.

The unique features of curating digital humanities data encompass not only the data themselves but also the research methods and practices. Curatorial practices form part of many humanities research methodologies, and the digital humanities community, in particular, already possesses sophisticated experience in preserving access to digital scholarship. Understanding the relationship between critical or interpretive activities which are also curatorial, and more traditional curatorial activities, which bear more relation to tasks traditionally carried out by libraries and archives, will be important in the context of the humanities. Although digital humanities is comparatively new in humanities terms, it has developed during a period of extraordinary and rapid change in all aspects of digital technology: storage technologies, data formats, hardware design, usage habits, and the emergence of discipline-specific practices. All of these changes have put tremendous pressure on the digital humanities community to be attentive to curation practices, and (perhaps above all) to learn from failure and loss.

There are a number of activities central to humanities research practice which themselves inherently curatorial, and whose value is enhanced if conducted with an awareness of the larger goals of data curation. Textual editing has for centuries sought to stabilize and transmit an accurate textual record, with specific enhancements (such as critical apparatus) which are aimed at making that record more valuable and tractable for research use. Within the field of textual editing, debates about method (and experiments with different kinds of apparatus) have also sensitized scholars to the significance of such choices and the need for explicit documentation; the critical edition as a result is one of the best-documented research artifacts in the humanities, and digital editions typically display this level of meticulous attention as well. Employing tools from corpus linguistics based on statistical analysis of textual features, scholarly editors can expand the scope of their curatorial interventions when working with texts from mass digitization projects. Similarly, the creation of digital thematic research collections constitutes an important kind of data curation, inasmuch as such collections are usually founded on substantial collections of source materials that have been encoded, annotated, and organized to maximize their value to researchers. Items in these scholarly research collections may receive more individual description and discipline-specific contextualization than libraries, managing large collections and facing backlogs in basic cataloguing and description, can provide.

These activities are directed at **creating** new knowledge; however, research practices aimed at interpretation and criticism can be considered curatorial in nature, particularly if their interpretive information is brought directly to bear on the material being curated. Scholarly articles that operate at a distance may offer relevant views but aren't intended as curatorial gestures, whereas interpretive annotations or critical interventions that link directly into the data would constitute very important acts of curation. These more complex interpretive activities also raise the issue of whether we can represent and preserve the motivations and debates that underlie them, since these too may be important for understanding the full semantic field of the thematic research collection.

Resources: Features of Humanities Data Curation

Piccini, Angela. "Locating Grid Technologies: Performativity, Place, Space: Challenging the Institutionalized Spaces of e-Science." *Digital Humanities Quarterly*. 3.4 (2009).

This article offers a very interesting case study of how performance and "live creative practices" operate in a digital arena, and suggests some of the distinctive curation challenges their "media traces" pose. The article reports on a series of AHRC-funded research workshops that situated digital performance within the context of e-Science, and examines the tools, infrastructure, and social/collaborative practices that emerged from these events.

Choudhury, Sayeed and Timothy L. Stinson. "The Virtual Observatory and Roman de la Rose." *Academic Commons*. 2007.

This article explores cultural changes related to digital humanities research and interactions with large datasets through the lens of two projects undertaken at Johns Hopkins University. In some ways, this article represents an early picture of the expansion from eScience to eResearch. Choudhury and Stinson draw useful comparisons between the work of astronomers and medievalists in terms of these researchers' changing relationship to datasets. Even if the specific vision of online spaces for virtual collaboration has not been realized, this article remains an accessible introduction to disciplinary data practices and how they might converge.