In the scholarly communication ecosystem, lectures and conference roundtables offer valuable opportunities to share one's on-going research and reflections with an engaged audience. Although social media, online conference programs, and slideshare sites now boost the signal of scholarly work, talks at conferences are still often limited by the time and place of their delivery. Even as digital humanities forges wider global connections, travel funding at universities and other public humanities institutions is on the decline. Everyone must make increasingly strategic decisions about how and where to invest travel time and money, often to the exclusion of professional events that might be interesting or thought-provoking. Digital networks and the open web, however, transcend boundaries of location and time by allowing for the rapid, broader distribution of transcripts, slides, posters, and audio and video.

In this seventh issue of the *Journal of Digital Humanities*, each featured piece translated what began as an oral presentation at a scholarly conference into another form for a wider audience on the open web. Some were *Digital Humanities Now*'s most frequently nominated content by Editors-at-Large or were the posts most highly-circulated by readers between May and September. Their timeliness, interdisciplinary reach, and popularity among readers make the following works well-suited for formal publication in the *Journal of Digital Humanities*.

In these articles, the authors consider the opportunities, challenges, and wide range of possible forms of digital humanities "data." Speaking about the need for humanists to consider their research materials a type of data available for analysis or manipulation, Christof Schöch first presented "Big? Smart? Clean? Messy? Data in the Digital Humanities" at the *European Summer School in Digital Humanities* in Leipzig, Germany, before sharing a version on his personal blog. Likewise, Trevor Muñoz offered his perspective on data curation as an emerging form of publication to the *CIC Center for Library Initiatives Annual Conference* in Columbus, Ohio, and then posted "Data Curation as Publishing for the Digital Humanities" on his personal website. Nicole Beale's presentation on collaborative explorations of 3D technology by archaeologists and artists was delivered at the *Electronic Visualisation and the Arts* conference in London before it was revised with her co-authors Gareth Beale, Ian Dawson, and Louisa Minkin for open-access publication in the conference's proceedings. Finally, Robert Sanderson's project briefing on RDF frameworks – both recorded audio and accompanying slides — was first delivered in San Antonio at the *Coalition for Networked Information* and was subsequently made available by the conference organizer on their website. In each case, the authors began with an orally-delivered version of their work, which they subsequently
circulated through informal, open-access publication avenues in order to reach an audience that extended beyond those who could be physically present when the paper was first delivered.

This issue of the *Journal of Digital Humanities* also features another genre of scholarly communication that rarely finds its way into formal publication venues: posters. More familiar to attendees of scientific conferences, posters are nevertheless an excellent means for humanities researchers and practitioners to share their approach to a humanities question or to offer best practices learned during their ongoing research. Highly visual, often static, and necessarily focused, posters are produced as exportable, printable files that are efficiently and easily distributed in a digital environment.

The *Journal of Digital Humanities* is proud to debut this genre of gray literature in this first of two installments of posters originally presented at DH2013, the annual, international conference of the Alliance of Digital Humanities Organizations. Presenting a poster at DH2013 allowed practitioners from around the world to introduce their research, tools, and projects to a global audience. Conference participants in July repeatedly celebrated the posters' high-quality and usefulness and expressed a desire to see them again in a format that could be thoughtfully considered and to which they could respond. In addition to public enthusiasm for the posters, we were aware that poster abstracts are written nine months prior to the conference and are often a poor substitute for the evolving final product. Without a venue such as the *Journal of Digital Humanities* to feature the posters themselves, interested readers would be forced to imagine the material from a sometimes out-of-date description.

The posters demonstrate the collaborative nature of digital humanities work by including all the names of the scholars, technologists, graduate and undergraduate students, and other collaborators who come together for project-based work. While some represent the resources of a single creator, other posters demonstrate the possibilities enabled by funding from a national or international organizations. For this reason, we invited the creators of each and every poster from DH2013 to first publish their poster on the open web, and then to submit their poster for publication in the *Journal of Digital Humanities*. Our goal was to improve the visibility of these already peer-reviewed works by offering a sustainable, open publication venue that benefits both those who were able to attend DH2013 and those who were not.

Gray literature originating as lectures, conference presentations, invited talks, and poster sessions offers timely insights into scholarship in progress. By providing a formal publication venue, the *Journal of Digital Humanities* increases the available avenues of access to the most recent developments in our field. While conference presentations, keynote addresses, or seminars will always remain important venues for interactions between presenters and audiences, the *Journal of Digital Humanities* hopes to forge connections between such work and those who are unable to attend. While many readers may have participated in one or two of the conferences in which the posters and talks in this issue were first presented, it is unlikely that any reader attended them all. As you read the features and peruse the poster gallery in this issue, we hope that you find new insights, new tools, or new approaches that are currently in development and of lasting value to you.

Joan Fragaszy Troyano and Lisa M. Rhody, Editors
Features

Big? Smart? Clean? Messy? Data in the Humanities
Christof Schöch .......................................................... 2

Data Curation as Publishing for the Digital Humanities
Trevor Muñoz .............................................................. 14

Making Digital: Visual Approaches to the Digital Humanities
Gareth Beale, Nicole Beale, Ian Dawson, and Louisa Minkin ........................................... 23

RDF: Resource Description Failures and Linked Data Letdowns
Robert Sanderson ......................................................... 33
This paper is about data in the humanities.[1] Most of my colleagues in literary and cultural studies would not necessarily speak of their objects of study as "data." If you ask them what it is they are studying, they would rather speak of books, paintings and movies; of drama and crime fiction, of still lives and action painting; of German expressionist movies and romantic comedy. They would mention Denis Diderot or Toni Morrison, Chardin or Jackson Pollock, Fritz Lang or Diane Keaton. Maybe they would talk about what they are studying as texts, images, and sounds. But rarely would they consider their objects of study to be "data." However, in the humanities just as in other areas of research, we are increasingly dealing with "data." With digitization efforts in the private and public sectors going on around the world, more and more data relevant to our fields of study exists, and, if the data has been licensed appropriately, it is available for research.[2]

The digital humanities aim to raise to the challenge and realize the potential of this data for humanistic inquiry. As Christine Borgman has shown in her book on Scholarship in the Digital Age, this is as much a theoretical, methodological and social issue as it is a technical issue.[3]

Indeed, the existence of all this data raises a host of questions, some of which I would like to address here. For example:

- **What is the relation between the data we have and our objects of study?** - Does data replace books, paintings and movies? In what way can data be said to be representations of them?

- **What difference does it make to analyze the digital representation or version of a novel or a painting instead of the printed book, the manuscript, or the original painting?**

- **What types of data are there in the humanities, and what difference does it make?** - I will argue that one can distinguish two types of data, “big” data and “smart” data. What, then, does it mean to deal with big data, or smart data, in the humanities?

- **What new ways of dealing with data do we need to adopt in the humanities?** - How is big data and smart data being dealt with in the process of scholarly knowledge generation, that is when data is being created, enriched, analyzed and interpreted?

### 1. What is data (in the humanities)?

As a starting point, it is useful to define what we mean by “data” generally and in the context of research in the humanities. First of all, let’s remember how data is generally defined. Information scientist Luciano Floridi defines data at its most basic level as the absence of uniformity, whether in the real world or in some symbolic system.[4] Only once such data have some recognizable structure and are given some meaning, can they be considered information. Floridi’s very general definition of data also shows why data can be represented in many different formats and on many different supports. Digital data is special in that it is discrete rather than continuous, and is usually represented, at its most fundamental level, in the form of a binary
notation involving just two symbols, 0 and 1. On a higher level, digital data are usually represented and processed in data structures that can be linear (for example arrays and matrices, like lists and tables in a data sheet), hierarchical (with a tree-like structure in which items have parent-child or sibling relations with each other, as in an XML file) or multi-relational (with each data item being a node in an interconnected network of nodes, as in graph-based databases).[5]

Some additional distinctions are important. For instance, there is structured and unstructured data as well as semi-structured data. Structured data is typically held in a database in which all key/value pairs have identifiers and clear relations and which follow an explicit data model. Plain text is a typical example of unstructured data, in which the boundaries of individual items, the relations between items, and the meaning of items, are mostly implicit. Data held in XML files is an example of semi-structured data, which can be more or less strictly constrained by the absence or presence of a more or less precise schema. Another important distinction is between data and metadata. Here, the term "data" refers to the part of a file or dataset which contains the actual representation of an object of inquiry, while the term "metadata" refers to data about that data: metadata explicitly describes selected aspects of a dataset, such as the time of its creation, or the way it was collected, or what entity external to the dataset it is supposed to represent. Independently of its type, any dataset relevant to research represents specific aspects of the object of scrutiny, be it in the natural sciences, the social sciences, or the humanities. Data is not the object of study itself, but "stands in" for it in some way. Also, data is always a partial representation of the object of study. In some cases, however, it is our only window into the object of study. Still, this "disadvantage" of partial representation is small compared to the fact that digital data can be transformed, analyzed, and acted upon computationally.

Data in the humanities is a bit special: one could in fact argue that text in a book or a manuscript, or the visual elements making up a painting, are data already. First, however, this is analog, non-discrete data, which cannot be analyzed or transformed computationally; and second, language, texts, paintings, and music are semiotic systems that have dimensions beyond the physically measurable, dimensions which depend on semantics and pragmatics, that is on meaning in context. For this latter reason particularly, speaking of “data” in the humanities is problematic and has been challenged. Criticism has come from mainstream scholars who see “data” and quantitative methods of analyzing them with suspicion, because the apparent empiricism of data-driven research in the humanities seems at odds with principles of humanistic inquiry, such as context-dependent interpretation and the inevitable “situated-ness” of the researchers and their aims.

Some practitioners of digital humanities, notably Joanna Drucker, have argued that the term “data” is actually inadequate. And indeed, the term’s etymology seems problematic in the context of the humanities: it comes from the Latin datum, which means “that which is given.” This means it carries with it the meaning of an observer-independent fact which cannot be challenged in itself. Johanna Drucker prefers to speak of “capta” instead of data, literally “that which has been captured or gathered”, underlining the idea that even the very act of capturing data in the first place is oriented by certain goals, done with specific instruments, and driven by a specific attention to a small part of what could have been captured given different goals and instruments. In other words, capturing data is not passively accepting what is given, but actively constructing what one is interested in.[6]

Similarly, Digital Archivist Trevor Owens has argued that data is not a given, but is always manufactured and created. Moreover, he shows, we can approach data from different perspectives and treat it as an
artifact (something actively and purposefully created by people), as
text (subject to interpretation, for example by scholars), and as
computer-processable information (to be analysed with quantitative
methods). According to Owens, this means that data is not a given and
not some unquestionable evidence; rather, it is "a multifaceted object
which can be mobilized as evidence in support of an argument."[7]

Even without using a new term, we can now redefine what we mean by
data in the humanities. Data in the humanities could be considered a
digital, selectively constructed, machine-actionable abstraction
representing some aspects of a given object of humanistic inquiry.
Whether we are historians using texts or other cultural artifacts as
windows into another time or another culture, or whether we are
literary scholars using knowledge of other times and cultures in order
to construct the meaning of texts, digital data add another layer of
mediation into the equation. Data (as well as the tools with which we
manipulate them) add complexity to the relation between researchers
and their objects of study.

Basically, I would like to argue that there are two core types of data in
the humanities: big data and smart data. These two types of data can
be described in two dimensions: the first dimension describes how
structured, clean, and explicit the data is; the second dimension
describes how voluminous and how varied the data is. I suggest to view
big data, in a first approximation, as relatively unstructured, messy and
implicit, relatively large in volume, and varied in form. Conversely, I
suggest to view smart data to be semi-structured or structured, clean
and explicit, as well as relatively small in volume and of limited
heterogeneity. Although you could say that these are really just
differences of degree, there are more fundamental differences between
them when it comes to looking at how each of them are created or
captured, modeled, enriched, and analyzed.

2. Smart data (in the humanities)

When we move from books to digitized versions of the text contained
in books, we are not necessarily dealing with big or smart data right
away. It may very well be small and simple, not to say "messy" data in
the beginning. This is probably the least useful type of data. So what do
I mean by "smart data?"

First of all, I should mention that "smart data" is not an established or
well-defined term. It is not very widespread and does not have a stable
meaning. Smart data is data that is structured or semi-structured; it is
explicit and enriched, because in addition to the raw data, it contains
markup, annotations and metadata. And smart data is "clean", in the
sense that imperfections of the process of capture or creation have
been reduced as much as possible, within the limits of the specific
aspect of the original object being represented. This also means that
smart data tends to be "small" in volume, because its creation involves
human agency and demands time. The process of modeling the data is
essential to small/smart data; its abstract structure can be defined with
elaborate schemas or as predefined database structures.

A prototypical example of smart data are scholarly digital editions
Technically, TEI documents are usually considered semi-structured;
usually, they follow a data model expressed in a schema, but such
schemas allow for considerable flexibility. In addition to a very clean
transcription of the text, digital editions using TEI can make a lot of
information explicit: first of all, TEI files contain not just the full text,
but also metadata associated with the text (in the teiHeader section);
also, the data is structured and explicit: there is markup making the
structure of the text explicit, identifying parts, chapters, headings,
paragraphs, as well as page and line breaks, for example. Finally, many
more types of information can be specified: for example person names
in a novel or play, place names in a letters or documents, and many more things; and links to other parts of the documents and to external documents. Making all of these things explicit allows to visualize them in specific ways and to index, count and analyze them computationally.

But let's move on to another example of "smart data." This data comes from a study of literary description in the eighteenth century novel which I conducted some years ago.[9] The aim was to identify all descriptive passages in a collection of thirty-two novels published between 1760 and 1800 and to find out how, from the standpoint of literary stylistics, descriptive writing "functioned" at that time. For this, a bibliographic reference management system was used as the front end to a database of descriptive passages which I collected and tagged for dozens of features I considered relevant for the study. For example, all 1,500 pieces of descriptive writing were tagged for the various textual strategies of integrating or legitimizing them in relation to the narrative context. This allowed me to discover previously unnoticed recurring configurations, patterns of usage, and trends over time. For example, although eighteenth-century novels do not "frame" descriptions as clearly and as symmetrically as some nineteenth-century novels do, most of the features of such framing are present. However, they are used in an asymmetrical way that tends to purposefully create a smooth transition between narration and description rather than a sharp framing contrast. In addition, correlations between such integrating strategies and different narrative perspectives were found, correlations which in turn help relate descriptive technique to long-term trends in French narrative fiction.

Using a database made it possible to deal with the 1,500 examples and their many tags. Also, having all excerpts and their tags at my fingertips changed the way I interacted with the data, as opposed to manual annotation and note-taking. Building the database itself was an ongoing process of explicit iterative modeling via an evolving set of tags and their relations, which involved adding more and more descriptions, adding tags to them, revising the tagging system and hence modifying the tags, etc. Recurring patterns and correlations could then be discovered; also, most importantly perhaps, outliers could not just be ignored and the resulting models attempt to cover not just a selection of examples judged to be representative but the full actual practice observed in the corpus.

The iterative process of modeling descriptive techniques with keywords. Image published with a CC-BY license.

Despite these significant transformations of the way we work with texts when they are available as "data," I believe the move from print culture to smart digital data is actually a rather small step compared to the steps required by big data, a subject which will be addressed below. In
the digital medium, we can also read texts, look at images, make annotations, and write down ideas and syntheses.

Now, this is all very well and good: smart data as we find it in scholarly digital editions, in annotated linguistic corpora and in carefully curated image collections is immensely useful. However, there is an issue with smart, clean data: it does not scale well. Although various aspects of creating smart data such as carefully encoded TEI documents can be at least partially automated, ultimately smart data depends on manual work by real people. Classifying descriptions in their context according to formal, semantic and narratologic categories is not something computers can do just yet. This means that it is very time-consuming to create large volumes of smart data.

Of course, there are ways to deal with this, and Machine Learning will no doubt be one of the keys to these challenges. But what if we actually don't really need smart data? What if having a lot of relatively unstructured, relatively messy data is just as useful and much easier to create? This kind of data is called "big data," so let's have a closer look at this alternative model of data.

3. Big data (in the humanities)

Big data is the buzz-word of the decade. Everyone wants big data and big data technologies; big data experts are telling large corporations they won't keep their competitive edge without big data. Areas as diverse as online marketing, stock exchange trading, health care, and political campaigns are driven by big data. The European Commission and the German Ministry of Education and Research hold "big data" conferences and fund big data research.

So, what does it mean for data to be "big"? Jonathan Ward and Adam Barker, the authors of a survey paper on definitions of big data, point out that because the term has been created and defined by industry, media and academia alike, there is a lack of common understanding regarding its definition. Their attempt to define such a common basis results in the following definition: "Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning."[10] Definitions of big data can indeed vary widely depending on the perspective adopted. In a recent best-selling book about the Big Data Revolution, the authors propose a non-technical, outcome-oriented definition of big data: "Big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value."[11] Another high-profile albeit more technical definition of big data, by Doug Laney, points to three key qualities of such data, the three V’s: volume, velocity and variety.[12]

Although the three V’s seem to provide a more precise definition of big data, they also show that big data is in fact a relative term and a moving target, depending on context and available technologies. The idea that big data is defined by its (large) volume is seemingly the most obvious of the three V’s. However, when does a large volume of data really become "big data"? You may consider data you want to analyze to be big when it exceeds the memory of your computer, forcing you to move processing to a grid computing system. However, technologies enabling this are becoming more widespread: A solution like "Hadoop", that allows the distributed but closely coordinated processing of huge volumes of data on hundreds or thousands of machines in a grid, is cutting-edge and fancy now, but will be mainstream very soon.

Second, the idea of "velocity" of data really means two things: first, that data is constantly being generated by sensors (in the natural sciences,
or by public surveillance cameras) or as a by-product of people’s activities in a digital environment (in economics or the social sciences), creating a constant influx of new data. Second, this flow of data is being analyzed in real-time and has to be very quick and responsive. In turn, this allows to react immediately to the data. This aspect of big data is probably the least relevant to data in the humanities, at least today.

Finally, the idea of "variety" of big data means that heterogeneous sources and formats of data are being used together, taking advantage of the links and overlap between such heterogeneous datasets to allow all kinds of inferences. What the idea of "variety" also implies is a variety of ways these datasets are structured, or a relative lack of structure in the datasets. The challenges here lie particularly in the fact that all these various datasets cannot be integrated into one unified dataset. The heterogeneity is probably the biggest challenge of data in the humanities, which may come from a variety of sources, in a variety of formats, and need to be combined flexibly in order to take the greatest possible advantage from them. Similarly to grid computing, however, solutions like "NoSQL"-databases or graph-based databases that avoid some of the limitations that more traditional SQL-databases have when it comes to heterogeneous or unstructured data will soon be part of our normal data analysis toolbox and using them will not be an indicator of "big data" applications anymore.

In addition to this, big data in the humanities is not the same as big data in the natural sciences or in economics. In most cases, velocity does not play a key role in big humanities data right now. Also, the large "volume" is less usefully defined in the humanities by a shift from databases to distributed computing. Variety of formats, complexity or lack of structure does come into play, however. In fact, the distinctive mark of big data in the humanities seems to be a methodological shift rather than a primarily technological one. And it is a huge methodological shift. Paradoxically, the shift from small smart data to big data is much more radical, I would argue, than the shift from print to smart digital data was. Indeed, moving from smart data to big data implies a shift from "close reading" to "distant reading" (in the words of Franco Moretti) or to "macroanalysis" (to use Matthew Jockers' term).[13] In this paradigm, instead of reading a few selected texts, we analyze an entire collection of relevant textual data.

The first consequence of the macroanalytic paradigm in the humanities, where hundreds or even thousands of texts are analyzed at a time, is that instead of operating on the level of literary forms and conventions, of semantics and context, we operate with quantitative measures of low-level features, on the basis of statistics and probabilities. The second consequence is that instead of so-called "representative" texts or paintings, we can now study the entire set of texts or images relevant to a specific research question. Trends in literature can be observed across the entire literary production of a given time and given genre. Questions of representativeness, of canonization, of literary quality play a much smaller, or at least a different, role in this context.

If this sounds too good to be true, that is because it is. Despite massive digitization efforts by private and public actors, we are still far from the complete record of human culture and creativity, even if we are looking just at text. For the British nineteenth century novel, for example, the calculations go as follows: an estimated 20-30,000 novels were published in the nineteenth century; of these, only around 6,000 are estimated to be still existent in the holdings of libraries and private collections worldwide. Of these 6,000 novels only about half, that is 3,500 novels, have been digitized in full text mode and with sufficient metadata; this is the number of novels contained in the "Stanford
Literary Lab Corpus” which Matthew Jockers used in some of the studies described in his book *Macroanalysis*. That corresponds to less than twenty percent of the total production.[14] This would certainly be a good sample size, if it were a random sample, but of course it is not. Rather, it is an opportunistic sample. So, the 3,500 novels seem like a small amount and not a number that actually resolves the sampling, representativeness, and canonization issues. Still, such a change of scale is a huge improvement over the mainstream canon which probably does not include more than one hundred novels for the nineteenth century, and may be visualized as a tiny spot at the bottom of the graph.

The practical limitations in the digital materials available for research mean that examples for "really big" data in the humanities are still relatively rare. Even Google Books is not quite there yet. Google has scanned more than 30 million books, as of April 2013, and continues to scan more. Compared to the estimated 130 million books ever published, this is a large part of the written human record. But of course, this is neither exhaustive nor are books the only medium of print publication there is, so newspapers, magazines and journals would need to be added to this. What really counts, however, from my point of view, is less the volume than the methods used for analysis. And these can be successfully applied to smaller sets of data as well, and imply precisely the methodological paradigm shift I mentioned, from close to distant reading.

I would like to give just one example from my own work, dealing with French drama from the seventeenth and eighteenth century and involving a maximum of 580 individual plays. That's not big data in the technical sense of the three V's, but it requires a radical shift from close reading methods to quantitative, statistical analysis. The basic question I have been addressing for the last year or so is how traditional categories of literary history, such as literary genres, forms and periods, relate to classifications made on the basis of the actual linguistic material. What comes out of this type of analysis, which can be pushed further in a variety of ways, is that there are indeed correlations between linguistic features on the one hand, and large categories from literary history on the other hand; but also, that these are not simple and stable correlations, but highly complex and varying ones.

In one case, for example, I analysed a collection of French plays from the seventeenth century written by Thomas and Pierre Corneille using Principal Component Analysis, a technique which discovers...
correlations in multidimensional data and summarizes such correlations into so-called principal components. The following graph shows how tragedies and comedies by these two authors cluster when plotting them according to the first two principal components.

Some interesting trends become visible: for example, it is remarkable how closely these two components seem to be related one to authorship and one to genre. Most plays in the left half are by Pierre Corneille, with some exceptions especially for the tragedies in the lower half. Most plays on the right side are by Thomas Corneille, again with some exceptions especially in the lower half. So the first component (horizontal axis) seems to be correlated with authorship. The second component (vertical axis), on the contrary, seems to be correlated with genre. Most plays in the upper half of the graph are comedies, and most plays in the lower half of the graph are tragedies. Only a few tragedies by Pierre Corneille and even less by Thomas Corneille appear in the upper half of the graph. Also, the dispersion of the data points (or plays) seems to be greater across PC1 for comedies than for tragedies. The tragedies are somewhat lumped together and authorship distinctions are actually hard to make just on the basis of PC1, so much overlap is there! This is not the case for the comedies in the upper half, where overlap seems to be much weaker. French tragedy as a genre, at least in the 1660s, seems to be more stylistically homogeneous than comedy, that is to be a particularly strongly conventionalized genre, a finding which is well-supported by mainstream scholarship.

**Conclusion: towards smarter big data or bigger smart data**

For most of this paper, I have been opposing big data and smart data. Indeed, big data tends to involve large volumes of raw, plain, somewhat messy text, whereas smart data tends to involve smaller volumes of carefully encoded, very clean text. Big data needs to be analyzed with methods from statistics, such as cluster analysis or principal component analysis and many more, whereas smart data can be analyzed with specific tools allowing to take advantage of structural,
linguistic and contextual markup. Big data requires visualization to even start understanding its possible structure, whereas smart data makes its structures explicit. In big data applications, outliers, errors and ambiguities are said to matter little because they get smoothed over by the sheer quantity of information that is good enough, whereas smart data makes exceptions and ambiguities explicit and effectively reduces possible ambiguities.

That said, I believe the most interesting challenge for the next years when it comes to dealing with data in the humanities will be to actually transgress this opposition of smart and big data. What we need is bigger smart data or smarter big data, and to create and use it, we need to make use of new methods. So, how can we enrich big data sufficiently to make more intelligent queries possible? How can we speed up the process of creating smart data so that we can produce larger volumes of it?

Basically, there are two possible ways to do this: one is automatic annotation, the other is crowdsourcing. Automation refers to various heuristics of discovering implicit units, structures, patterns and relations, and of making them explicit in the data. Crowdsourcing, on the other hand, relies on breaking down a large task into such small units that each of these little tasks can be performed in a distributed way by a large number of volunteers. Various strategies have been developed for breaking up the tasks, for creating incentive structures to motivate volunteers (like "gamification" or "win-win"-constellations), and to reintegrate the added information into the project.

In fact, automation and crowdsourcing will have to work hand in hand. This is what happens with OCR: better and better optical character recognition systems are still no match to manual double-keying or transcription by experts, especially when it comes to print before 1800 or to handwriting. But state-of-the art OCR combined with algorithms to detect potential areas of error and cleverly crowdsourced and distributed error-correction mechanisms such as the ones implemented by "Captcha" go a long way to producing large amounts of more reliable full text. Similarly, automatic linguistic annotation even of basic linguistic features for well-researched languages is still too faulty to be trusted blindly, at least in a "clean smart data" perspective. We will have to find ways of detecting potentially faulty linguistic annotation, then finding and motivating users to check such annotations, and writing the corrections back into larger and larger collections of clean, structured and well-annotated text.

To summarize the story of data in the humanities which I have been trying to tell, one could consider that this story has several steps: The first step leads from the study of creative works in the form of books, paintings and movies to their study based on digital representations of
these works; this is what digitization at its most basic level as brought about; the first of two possible steps from there leads to smart data, that is to data that has been carefully curated, structured, annotated in a way to make explicit a lot of information that is implicit in the "raw" and messy digitized artifacts. This happens, prototypically, in scholarly digital editions of text or music scores. The second of the two possible steps from "raw digital data" leads to big data, simply by accumulating more and more data and letting the algorithms sort it all out, instead of cleaning it up by hand. The last step in this story is to reconcile, or rather to combine, the smart and the big data approaches.

For my own research in computational genre stylistics, having collections of texts at my disposal that are both larger and smarter than what we have now will be crucial. Collections need to be large, because as soon as you focus on more specific cases, such as a specific sub-genre from a specific period, even a relatively large collection of texts will only yield a small number of samples. And when the number of samples gets too low, statistical approaches loose their robustness and reliability. And similarly, more nuanced and interesting analyses of large text collections depend on having a large array of metadata and annotations regarding each text, including things like the proportion of verse and prose in a play, or of description and narration in a novel. Such information needs to be available so that correlations between stylometric findings concerning a text and relevant attributes of the text in question, can be discovered.

In other words, we need smart big data because it can not only adequately represent a sufficient number of relevant features of humanistic objects of inquiry to enable the level of precision and nuance scholars in the humanities need, but it can also provide us with a sufficient amount of data to enable quantitative methods of inquiry that help us transgress the limitations inherent in methods based on close reading strategies. To put it in a nutshell: only smart big data enables intelligent quantitative methods.


Notes:

[1] This contribution is a revised version of a talk I gave at the European Summer University "Culture & Technology" organized by Elisabeth Burr at the University of Leipzig in Germany, on July 26, 2013. The Summer University brings together graduate students and researchers from a wide range of disciplines in the humanities and in computer science from many different countries around the world. The talk was one of several plenary talks intended to introduce the audience to various topics of more general concern, in addition to the more specific workshop tracks. The talk was the result of a cooperation between the European Summer University and the German branch of the DARIAH initiative in which I am a research associate and where I am mostly concerned with understanding digital methods and with supporting and training mainstream humanities scholars to use such methods. A slightly revised version of the talk was documented on my blog, The Dragonfly's Gaze, in early August 2013. The research reported here has been supported by DARIAH-DE with funding provided by the German Federal Ministry of Education and Research (BMBF) under the identifier 01UG1110A-M.

[2] There are many such sources of digital data for research in the humanities: we have large text archives such as Google Books, Hathi Trust, the Internet Archive, or Gallica; we have scholarly digital text archives such as TextGrid's Digital Library of German-language
literature from 1500 to 1900 or the Théâtre Classique collection of French Drama, and many others; and we have hundreds of elaborate digital editions of literary and historical texts, such as the Van Gogh Letters, the Walt Whitman archives, or Rousseau-Online. We have image databases like “Flickr Commons,” the distributed image archive Prometheus, Getty, and others. And we have sites like Youtube, Open Culture and the Moving Image Archive.


[9] This study was done as part of my doctoral dissertation; its results have been published as Christof Schöch: La Description double dans le roman français des Lumières 1760-1800. Paris: Classiques Garnier, 2011.


Data Curation as Publishing for the Digital Humanities

“Publishing” has assumed a large role in discussions of how scholarship is changing. One reason is that, in these discussions, the mechanisms of publishing come to stand in for the larger and more complex processes of creating, vetting, and circulating knowledge. Some of the sense of unmet need that arises in considerations of the emerging, alternative publishing methods for those working in digital humanities comes from the problems with this shorthand.

Meeting the publishing needs of digital humanities scholars is challenging not only because the outputs may take new forms—digital, “database-driven,” or somehow online—but also because some of the publishing that digital humanists want or need to do encompasses processes of knowledge creation, dissemination, and exchange that “publishing” does not encompass in its current forms. To put it another way, if we are focused on outputs then we can probably accept “publishing” as a shorthand for the larger processes of scholarship: workflows, peer-review, marketing, production need to be adjusted certainly for new forms but the framework of the conversation can remain the same. Yet, if we examine the work that humanists are doing—in something like the way that scholars in the field of Science and Technology Studies (STS) have done for science—by looking at their culture of material practices, then the familiar framework of “publishing” does not serve us well. The interplay of theory, data, and computational methods in a significant portion of digital humanities scholarship works in such a way that to publish this scholarship requires that we add some new dimensions to our ideas of “publishing.”

I want to suggest that the theory and practice of data curation can augment our notion of “publishing” in a way that will serve the needs of the digital humanities community. The work of data curation —“active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; ... activities [which] enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time” (Cragin et al. 2007) — should be legible as “publishing” work for libraries and scholars to do in much the same way that well-understood tasks related to preparing and circulating monographs or journals are already legible as publishing work. Moreover, I argue that articulating connections between “publishing” and data curation is important in the context of strategic decisions that libraries make about how to participate in “publishing.” Data-curation-as-publishing is publishing work that draws directly on the unique skills of librarians and aligns directly with library missions and values in ways that other kinds of publishing endeavors may not.[1]

The link between data curation and publishing is not wholly new. In 2009, Joyce Ray, Sayeed Choudhury, and Mike Furlough presented a paper at the Charleston Conference, summarizing several strands of contemporaneous work. The paper was entitled “Digital Curation and E-Publishing: Libraries Make the Connection.” In this paper, Ray,
Choudhury, and Furlough describe how data curation and publishing can be mutually-reinforcing activities. They write:

We have on the one hand, a community, or a subset of several communities, that has been working on the “back end” of digital production from the generation of raw data to the construction of an organized product that can be accessed, and, on the other hand, another community—publishers—who work on the “front end” of scholarly communications, from manuscripts to publication (Ray, Furlough, Choudhury 477).

Making the connection of the title involves bringing these communities together as complementary elements of a service portfolio, staffing model, and infrastructure that justifies the funding and the relevance of libraries in a changing scholarly environment. This is a good argument and some innovative libraries (among them Penn State, Johns Hopkins, New York University, and Purdue) seem to be having some success with this as a strategy. The main thrust of the argument that Ray, Choudhury, and Furlough advance is managerial, bringing together libraries and (university) publishers (under the aegis of the library) as an attempt to rationalize the “business” that the combined library/publisher is in. Treating data curation and publishing as kindred services may offer the prospect of expanding a library’s stable of “innovative” offerings while not straining resources because there are management efficiencies in having both the “front end” and “back end” people in the library. However, in this model, neither libraries nor publishing seems truly transformed and this is a problematic mismatch when so many other aspects of scholarly work are being transformed.

I argue that it is possible, even preferable, to treat the connection between data curation and publishing as more fundamental. As Ray, Choudhury, and Ray themselves say: “Digital curation is a useful label for that collection of challenges newly located at the intersection of publishing, collections development, preservation, and the humanities” (Ray, Furlough, Choudhury 479).

Data Curation and Digital Humanities

As an enterprise, “digital humanities” (formerly “humanities computing”) dates back to the late 1940s (debatably, even earlier) and, since at least the 1980s, the curation of digital humanities research data has been an associated area of research, activity, and concern. Many of the early genres of digital humanities scholarship re-enforced this connection between digital humanities scholarship and data curation: “development of indices, annotated linguistic corpora, and digitally encoded texts—in other words, the preparation, collection, organization, and maintenance of datasets” (Palmer, Weber, Muñoz, and Renear). For a fuller account see Palmer, Weber, Muñoz, and Renear, 2013. In a piece for the Chronicle of Higher Education in 2002, Jerome McGann predicted that “in the next 50 years, the entirety of our inherited archive of cultural works will have to be re-edited within a network of digital storage, access, and dissemination” (B7). Many have taken McGann’s forecast as both a description of and a call to the work of digital humanities. This vision of digital humanities contains clear parallels with the goals of data curation practitioners and teachers aiming “to build and maintain not only digital libraries and curated data sets, but also the associated indexing systems, metadata standards, ontologies, and retrieval systems” (Palmer, Renear, and Cragin 2008 3). The point here is not that digital humanities and data curation have a special affinity—there are long and rich histories of data curation in the sciences and social sciences. My point is that data curation has been part of the ambit of digital humanities for a long time and this should guide us in thinking about how to publish digital humanities work.
In referring to “data curation,” I am speaking specifically of information work that integrates closely with the disciplinary practices and needs of researchers in order to “maintain digital information that is produced in the course of research in a manner that preserves its meaning and usefulness as a potential input for further research.” (Munoz and Renear 2011) This distinguishes data curation from many near synonyms: digital curation, digital stewardship, digital preservation. The emergence of a specific discourse on “data curation” in the sciences (with accompanying policy development, funding, and new research interest) provides a framework for pulling together diffuse and disparate activities in the humanities and describing these activities under the new rubric.

Thus, digital humanists in particular are becoming increasingly aware of data curation issues and data curation needs as part of the way they (we) work. This element of digital humanities work is becoming prevalent enough that I selected examples more or less at random from items that I came across in my professional networks and social media feeds over the period of a week in the spring of 2013. Lincoln Mullen, a PhD candidate at Brandeis University, posted on his blog about using the statistical programming language R for historical research. As part of his discussion, Mullen describes how he converted the tables from a monograph he found in his research to a series of comma-separated-values (CSV) files in order to produce graphs and charts of the changing demographics of American religion. Along with his analysis and the blog post about his methods, he posted the (small) data set to Github, a platform for sharing open source software code and open data. Ted Underwood, Associate Professor of English at the University of Illinois, has made the work he and a graduate assistant have done building, cleaning, normalizing, and labeling a data set drawn from the HathiTrust corpus a significant part of the output of his “Uses of Scale” project and other professional presentations. Kathleen Fitzpatrick has argued that humanists “might … find our values shifting away from a sole focus on the production of unique, original new arguments and texts to consider instead curation as a valid form of scholarly activity” (Fitzpatrick 79). Fitzpatrick uses “curation” here as a near synonym of selection after the manner of an gallery or museum curator selecting art for an exhibition—a slightly different meaning than I have been developing. The examples above of computational work with datasets draw in additional meanings of curation related to information science.

It is also increasingly common to see the release of open data sets as enticement to attract digital humanists to work on particular sets of questions, or in partnership with cultural heritage organizations—see, for example, the IndexCat data from the National Library of Medicine, a small collection of catalog records for a historical library of children’s literature, data from some of the crowdsourcing projects run by the New York Public Library, the Smithsonian Cooper-Hewitt, National Design Museum collection data, and many more examples.

Data Curation-as-Publishing

At least part of the professional activity of the digital humanists and organizations above involves making data available and suitable for re-use. As any of the researchers involved would no doubt say, curation of these data sets takes time, effort, and money. Libraries getting involved to help digital humanists do this kind of work would be offering something of value. This would be “publishing” not only in the sense of registering and “making public” a product of scholarly work, but also in the sense of ensuring quality and disseminating outputs to interested communities.[2] By recognizing data curation work as a publishing activity, libraries would have a “market opportunity” to address unmet needs in the digital humanities community (among others).
Distinguishing data-curation-as-publishing (a new and more-broadly conceived activity suited to the kinds of knowledge production and dissemination happening in digital humanities) from data as merely another form of publication is a crucial point. In a recent publication in *Data Science Journal*, Mark Parsons and Peter Fox explore “data publication” as a metaphor for the kind of things that scholarly communities want to see happen with data. They explain that “Data Publication builds from the familiar and conceptually simple model of scholarly literature publication” (WDS37) and they capitalize the terms deliberately to indicate the status of this phrase as “a recognized metaphor and data management paradigm” (WDS33). Parson and Fox’s paper elaborates on what are some significant problems in adopting this metaphor. In the limited space available I want to focus on just one of these problems. Parsons and Fox note that under the model of Data Publication “publishers are distributed and can act autonomously or in concert” (WDS37). Thus, they write:

> there is ... little emphasis on data discovery and interoperability across systems. Data are often presented as they were created without explicit considerations of data integration or significant reuse. ... The attention is on preservation and formal recognized scholarly contribution with less attention to ... issues such as latency, rapid versioning and reprocessing, and computational demands (WDS37).

To understand data-curation-as-publishing (which I’m advocating as a way to serve digital humanities scholars) only as “Data Publication” expands recognizable publisher and library activities to a new class of scholarly objects (data) but in many ways perpetuates the (flawed) status quo.

Within the critique of “Data Publication” there are glimpses of what it could mean to treat the activities of data curation as “publishing” activities in a way that would benefit both scholars and libraries. The first part of Parson and Fox’s critique is that under the model of “Data Publication” there is “little emphasis on data discovery and interoperability across systems” (WDS37). Various examples from the media landscape beyond scholarly publishing suggest the truth of this claim. In the realm of ebooks, the importance of outlets like Amazon and other digital dissemination channels has recently forced publishers to pay greater attention to “discovery” and to devote more resources to things like metadata. But at the same time, the fracturing and proliferation of ebook reading platforms is an ongoing example of problems of interoperability across systems in a publishing marketplace (there is a similar shape to the story of the relative fortunes of the on-demand video company Netflix and various real or rumored video platforms implemented by specific studios or content creators). This leads to the question of whether lack of emphasis on discovery and interoperability are intrinsic to the business of publishing (presumably because the energies of publishers are directed elsewhere to activities considered more vital to mission and survival)? Attention to “discovery” and related issues of interoperability across systems are traditional and persistent features of library work.

This is where Ray, Choudhury, and Furlough see opportunity for libraries who can “make the connection” between publishing and data curation—to excel where traditional publishers have not—by having both “back end” (librarian) and “front end” (publisher) organizational capacity. In their discussion of “back end” and “front end,” these authors map the organizational rationalization of connecting data curation and publishing as library activities onto existing lifecycle models of data (this is explicit in the paper) and onto an extended lifecycle model for scholarly publications (in the example they give, journals and monographs) that encompasses both distribution and long-term preservation. “Putting a standard monograph series online didn’t make the Library a publisher,” they write, “but it linked the
Library’s role as a preservation agent more directly to its emerging role as a distributor” (Ray, Choudhury, and Furlough 480).

Yet, for as much as these early sketches of programs (for which the authors deserve credit as pioneers in making any such moves in this direction) emphasize the interdependence and mutual reinforcement of curation and publishing, this vision of scholarly work with data is still somewhat disappointingly static and familiar. Publishers add value to end products through peer review and high quality production and presentation. Libraries standardize and preserve these outputs and continue to make them available to a community over time. Organizations which comprise both library and publisher can imagine this as a unified suite of services that cover the entire data lifecycle. However, if Data Publication, rather than data curation is the governing metaphor, this alignment, just having both “back end” and “front end” of the process, may not be sufficient to avoid falling into traps such as neglecting discovery and interoperability of digital humanities work with data. It is worth noting, too, that a Data Publication model does not easily encompass “issues such as latency, rapid versioning and reprocessing, and computational demands” that resemble precisely the kinds of demands that digital humanists are likely to make. (Parsons and Fox WDS37).

This leads to the next part of Parsons’s and Fox’s critique—under a model of Data Publication, “data are often presented as they were created without explicit considerations of data integration or significant reuse.” (WDS37) Data “presented as they were created” sounds like a description of researcher self-deposit into (institutional) data repositories—currently the most common form of library engagement with data curation and embedded in both the Johns Hopkins and Penn State models. Libraries cannot adopt a position of becoming data publishers (via repository provision) in the way some are seeking to become journal publishers through the provision of platforms like Digital Commons and similar initiatives. That Parsons and Fox single this problem out in a discussion of why Data Publication is a problematic metaphor from the perspective of solving the real information needs of researchers suggests that while the provision of institutional data repositories is necessary and important it is not sufficient to “purposeful work” with data. So, libraries cannot stand pat; they cannot maintain only the “back end” of these processes but must make the connection to more active engagement.

Ray, Choudhury, and Furlough present a fully-developed model of the data lifecycle (developed by the UK Digital Curation Centre) but offer only a loose schematic of a publishing lifecycle. To suggest how data-curation-as-publishing may help expand the notion of “publishing” in a way that would allow libraries to break out of their traditional role at the “back end,” I would add another model to the discussion. Historian Robert Darnton proposed a model of what he-called “the communication circuit” in a seminal paper from 1982 entitled, “What is the History of Books?”. Darnton’s model has been critiqued and elaborated by other book historians (including Darnton himself) since the first publication but the original version will suffice to advance the argument about data curation and publishing. What is salutary about Darnton’s model and what makes it an interesting partner for data lifecycle models is that it includes, in addition to authors and publishers, printers, shippers, booksellers and readers as part of the scope for understanding books as texts and objects.

Part of the unique contribution of book history as “an interdisciplinarity run riot” has been to reveal the agency of those who package, categorize, organize, disseminate, receive, re-use, and interpret in the co-creation of meaning and knowledge (Darnton 67). I am not convinced that Ray, Choudhury, and Furlough’s merged
organization capable of both “front end” and “back end” work is sufficient to cover the stations of this richer model. I believe libraries should treat data curation activities as “publishing”—worthy of new enthusiasm and new resources—but they (we) should be wary of framing the endeavor as “data publishing” (an analog to journal and monograph publishing). By taking “publishing” as a category to be re-imagined rather than a pre-existing workflow to stepped into, libraries can and should take a more active role in work with data—this is data curation-as-publishing.

**Developing Capacity for Data Curation-as-Publishing in Libraries**

What form might this actually take? To offer a specific example from digital humanities, data-curation-as-publishing might look something like the Alexandria Archive Institute’s [Open Context project](#). Open Context provides review, documentation, and publication of research data, mostly in the discipline of archaeology. The “About” page of the project web site speaks of “data sharing as publication” and a flavor of the work the project carries out can be gleaned from a representative sample of the [editors’ blog](#), which discusses matching date files with code books, cross-checking values, annotating, and describing the data set. The editors [remark](#) that “data sharing requires similar levels of effort and professionalism as other more conventional forms of publication.” Open Context is hosted and administered by the non-profit Alexandria Archive Institute and thus represents a kind of freestanding example of an organization doing data-curation-as-publishing. What would it mean to locate this effort in libraries?

This question recalls a point that Choudhury, Furlough, and Ray make in passing. In describing the creation of the Data Conservancy architecture and service at Johns Hopkins, they write: “It is especially important to note the role of a particular individual at AAS who acted as the human ‘interface’ between the various players. This individual could easily be classified as a ‘data scientist’ – an individual with knowledge of a specific domain or discipline yet also a deep knowledge of data management” (479). They go on to remark that “libraries would be wise to consider developing such expertise and capacity in-house” (479). I contend that Open Context, and its editors, represent another example of this kind and that libraries should be figuring how to set up and host such activities. Developing the capacity to partner in this more broadly conceived version of publishing that digital humanities and other data intensive disciplines increasingly need will require libraries to alter how they relate to collections.

At the [University of Maryland Libraries](#), those working on data curation are beginning to work on the question of how to make a case to subject selectors (who control collection budgets to support various disciplines) to spend collection funds on curation work for significant data sets. These discussions are still at early stages—there is lots to figure out including what specifically should appear on “the invoices” for such data curation work that selectors are being asked to pay—but libraries who wish to engage seriously with support for data-intensive research (like the digital humanities) will increasingly need to sell and buy such services. These funds will need to come from collections (because that is still where the bulk of the budgets reside) and accomplishing this shift will entail [breaking up many of the present economic “realities”](#) that shape libraries’ collection development.

I am not envisioning allocating funds to buy datasets—perhaps from a vendor or platform that makes them available. The valuable work, the work that libraries should own, is the type of activity like those the editors of Open Context perform. The current situation in which libraries purchase subscriptions to large databases of, for example,
journal articles, represents not only an unsustainable economic situation but also an unsustainable professional one in which libraries outsource the expertise and experience of collecting, normalizing, organizing, and making available scholarly information. Librarians should spend more time on creating metadata, building catalogs, developing and refining indexes, and building, organizing, and maintaining collections than on negotiating publisher contracts or teaching the details of interfaces created by vendors. Extending library, archive, and information science practices for data may include aggregating data sets, cleaning and normalizing values, and annotating data with controlled vocabularies and ontologies. The issues of description, organization, and access for data are still largely unsolved and libraries should demonstrate their expertise in solving these challenges through developing and sustaining data curation-as-publishing programs.

Data Curation-as-Publishing Aligns with Library Missions and Values

There is a clear need for data curation work. Perhaps this should be its own strategic initiative for libraries to pursue in parallel with “publishing” initiatives? (This is in fact what many libraries are doing.) Yet, with the financial support for libraries in flux, how many strategic initiatives can libraries count on and expect to do well? Data curation activities are fully legible as “publishing”—meeting the same ends and goals and potentially contributing to scholarship in the same kinds of ways. Also “library publishing” is a site of buzz and activity and potential investment—partly, as I have argued, because the processes and products of publishing stand in for “scholarship” writ large. I would argue that if libraries are going to invest resources in “publishing,” then that money should be spent partly on doing data curation work because data curation-as-publishing offers the most value to both researchers and libraries.

Data curation-as-publishing is the right form of publishing for libraries to be in because the work of data curation aligns with libraries’ missions and values in ways that other kinds of publishing ventures do not. There is much about scholarly “publishing” as it exists now that is not about making knowledge public or ensuring quality of that knowledge or disseminating it to those who need and could use it. There is a great deal of “publishing” that is about issues of prestige, labor, and equity of the disciplinary professions. In my opinion, libraries don’t really have a dog in that fight and shouldn’t spend resources trying to fix those problems.

In a recent paper in the library and information science literature on assessing data value, Carole Palmer, Nic Weber, and Melissa Cragin remind us that “the library and information science meta-science perspective articulated by [Marcia] Bates has always been fundamental to the role of providing broad, useable information collections and services, especially to support interdisciplinary research” (1999). Doing data curation work (like that described above) needs the unique training and skills of librarians and other information professionals and it supports the goals and values of the profession in making information accessible and usable to communities of users who need it. Making data curation fully legible as publishing, and investing in data curation-as-publishing, can help make problems of data discovery, interoperability, and re-use less daunting and show a clear way for the library to be a publisher in ways that research communities like digital humanities need.

Notes:

[1] This piece was originally posted as an edited version of a presentation given at the CIC Center for Library Initiatives Annual Conference, May 22–23, 2013. The theme of the conference was “alt.pub.edu: Emerging Options for Scholarly Publishing” and I was delighted to be part of a panel with Matt Gold (CUNY) and Matthew Jockers (Nebraska) on “Digital Humanities, Alternative Publishing Needs of Faculty.” I have made additional revisions for publication in JDH. My thanks again to all the staff of the CIC Center for Library Initiatives and to the members of the Program Committee for the 2013 Annual Conference for inviting me to speak.

[2] Thanks are due to Shana Kimball for prompting this extension of the argument in discussion after my original talk.

Works Cited:


Making Digital: Visual Approaches to the Digital Humanities

The *Making History Project* is an attempt by artists and archaeologists based within the [University of Southampton](http://www.southampton.ac.uk) to collaboratively develop innovative uses for 3D technologies. Techniques such as high resolution data capture and 3D printing represent a new era in digital imaging. As these technologies become increasingly affordable they are coming to play a more significant role in archaeological and artistic practice. Both art and archaeology are currently involved in attempting to realise the full implications and potential of these technologies. This paper describes a project undertaken by the [Archaeological Computing Research Group](http://www.archaeoleng.ac.uk) and [Winchester School of Art](http://www.winchester.ac.uk) at the University of Southampton which seeks to address this moment of technological disruption in order to collaboratively develop creative and methodologically innovative approaches to the use of these technologies.

1. INTRODUCTION

The introduction of 3D data capture technologies and 3D printing have begun to have an impact on practice in art and in archaeology. The increased availability of these techniques, as software usability improves and hardware costs lessen, has been accompanied by a high level of interest from not only the art and archaeology communities, but also from the general public. This interest is without doubt justified. These technologies present the possibility of tactile engagement with data in a way that has never previously been possible. It is essential that engagement with this technology is critical, creative and grounded within good disciplinary practice, and that we avoid what Huggett termed the technological fetishism which often accompanies the introduction of technology to a discipline or area of study.[1]

During this period of inflated expectations for 3D technologies, it is important that we identify tangible benefits resulting from the uptake of these technologies for both art and archaeology. This is in line with Gartner’s technology adoption hype cycle.[2] We must avoid a decrease in the use of potentially valuable tools as interest diminishes and enthusiasm wanes.

The *Making History Project* seeks to drive the innovative and relevant use of these technologies. The project is formed of a number of event-based collaborations and sub-projects each of which give participants an opportunity to collaborate and to critically assess the potential use of technologies in their own, and in each other’s, disciplines. The events are participatory and highly discursive, allowing project members to think critically about the use of technology while sharing resources, expertise and disciplinary insights.

1.1. Why Art and Archaeology?

Archaeology and art as disciplines have been intertwined for as long as both have existed. At the most fundamental level, both have a shared focus on substance and materiality.[3] The value of the interplay between the two areas lies in the inherent reflexivity of the relationship. From the first works of antiquarian scholarship through
to the present day, the influence of the arts upon archaeology has been profound.[4]

The way in which the past is represented and perceived has been characterized by a constant dialogue between art and archaeology and in many cases the active involvement of artists in the archaeological process.[5] The styles and conventions which characterize archaeological representation of all kinds (including illustration, printing, photography, and computer graphics) were not developed in isolation, but were honed and developed in response to, or alongside, changing artistic practice and the changing aesthetic norms of society at large.

Archaeology has also had a powerful influence upon art. Art objects might be considered historic documents in their own right, illuminating their context through the place, circumstance, and conditions of production. Furthermore, the archaeological process itself is one of the key metaphors of modernity in art practice, encompassing measured relations to place and landscape as a cultural substrate. Use of the archaeological metaphor also has the capacity to invoke myths of origin, psychoanalytic excavations, and fragments of the Romantic. Artists have long been fascinated by the processes of archaeologists and their antecedents. This association has been commonplace in twentieth century practice, from Picasso’s use of ‘primitive’ ethnographic artifacts, to Hepworth, Moore, and Nash’s reflections on the ancient landmarks of the British landscape, and Deller’s re-enactment of the Battle of Orgreave. Artists and archaeologists share a concern with the visible. The movement of images and gestures across cultures and historic eras demonstrates both an interest in the operations, ideology and figures of other times and spaces, and a sense of the importance of the historic transfer of process and technology and its operation in transforming practices and knowledge.

There can be little doubt then that the connections between archaeology and art are valuable in both a profound and highly practical sense. These connections have inspired thinkers and practitioners in both fields and have the capacity to continue to do so. The differing but closely related skill sets of artists and archaeologists can motivate creative approaches to practice on both sides. The potential which artists and archaeologists may identify in new technologies will almost certainly not be the same. It is highly likely that this difference of perspective will cause collaborators to review their assumptions and to think more creatively about the possibilities offered by these technologies. The main challenge of the Making History Project is to catalyse and to support multidisciplinary collaboration. In line with this, one of the goals of the project has been to develop the potential for collaborations of this type to occur within the university setting.

1.2. Technologies and Academia: Adoption and Adaption

Academic practice traditionally has been bound within disciplines, with each discipline preferring particular methodologies for knowledge production and knowledge sharing. Multidisciplinary approaches have long been recognized as essential for solving research problems.[6] At most universities this kind of working is encouraged and supported. But the institutional, cultural, and funding settings must be developed to support rather than impede the type of relationships that will lead to such work being possible.[7] The relationship between art and archaeology is an example of this collaboration. Whether acknowledged or implicit, there has always been a link between these disciplines, and this link, and the potential for further explicit collaboration, is explored by the Making History Project.
Administratively, universities must be managed by research group, department or faculty, which often leads to barriers to working multidisciplinarily. The *Making History Project* aims to transverse these boundaries by providing opportunities for ideas and approaches to occur organically, and allowing for iterative development of project management. Within *Making History*, two departments with different approaches to adopting technologies embarked on a journey to use an event-based relationship to enrich both departments’ understanding of the possibilities of 3D technologies.

One key part of the project is to examine the use of 3D technologies within art and archaeology together. This investigation attempts to find a way to incorporate approaches from each department into each other’s work, whilst also collaboratively creating outputs together, with both students and staff. The intention is to find new ways to use 3D technologies that would not have otherwise been identified.

### 2. MULTIDISCIPLINARITY IN ACTION

Traditionally multidisciplinary work involves project specific collaboration, closely tied to pre-agreed aims and objectives. Most projects are driven by one department, with other disciplines becoming involved as specialisms are required. We did not want our collaboration to be based on merely explaining to each other how we work. It was essential that both disciplines have direct involvement in decision-making and in the identification of activities. In order for this to occur, we opted to organise a series of events at which archaeologists and artists could learn through participating in typical art and archaeology activities.

The project began in April 2012 with meetings to discuss the potential for working together and arranging opportunities for staff and students to meet. We began the project believing that art and archaeology have different but overlapping goals. For archaeology, a key goal was to develop new approaches to the creation and use of 3D data, and the use of 3D technologies, through collaboration and problem solving. For art, the goals could be described as being more vampiric than overlapping; participants from WSA sought to discover how archaeologists use new technologies for recording and exhibiting. It was hoped that the collaboration might provide a new body of material for creative consumption, and that students and staff would discover new processes to adopt and adapt.

During initial planning meetings, the staff from WSA commented that artists thrive on dialogue, re-purposing ideas and methodologies. The project embraced this as an essential part of the approach, encouraging students to fold these technologies into their vocabularies of making with confidence and agency. This is one of the first projects attempting to develop innovative approaches to cross-faculty collaborative research and education based around shared technology and expertise. The project has at its heart the development of a better understanding of the opportunities and limitations that 3D technologies offer for art and archaeology, both separately and together.

### 3. TECHNOLOGY: DISRUPTION, INNOVATION AND PRACTICE

3D technologies have the capacity to revolutionise artistic and archaeological practice. The implications for each area differ greatly, but the opportunities for both areas can be more fully understood and more fully realised through co-exploration. The following section will outline the opportunities which project participants perceive that 3D technologies will offer in their subject areas.

#### 3.1 Opportunities for Archaeological Practice
3D data capture and 3D printing are relatively recent additions to the toolkit used by archaeologists. The potential of high resolution 3D data capture has been identified and exploited in a range of archaeological settings. Primarily the technique has been employed as a means of recording archaeological objects or environments. Despite the growth in the volume of 3D data being produced by archaeologists, the range of techniques available which allow these data to be explored, interrogated or disseminated remains limited.

Traditional computer graphics techniques have been widely employed as a means of exploring and disseminating large 3D data sets. Dissemination has generally taken the form of animations or still rendered images which offer the viewer few options for engagement or interaction. 3D printing offers an alternative mechanism with which to explore and disseminate 3D data, however this technique has not yet been widely adopted. 3D printing offers the possibility of tactile, multi-sensory engagement with 3D data. 3D printed objects invite us to consider a range of factors relating to the feel, weight and materiality of the object in ways which have not previously been possible. In order to fully realise the potential of this technology and to fully appreciate its limitations it is important that archaeology confers with other disciplines, such as art, which have a similar focus on objects and their materiality.

### 3.2 Opportunities for Art Practice

The past decade has seen an exponential increase in 2D image transfer technology, from capture apparatus to compression codecs. New languages have opened up through the accessible transmission, manipulation and mediation of images, the boundaries of art practice have fallen. Technologies of reproduction produce societal change. They have affect. 3D technologies open up new ways again for us to think about fabrication, seriality, dissemination and multiples in art practice. There is an inherent shift from conventional structures of disciplinarity within an art school even in naming a ‘3D print’. The shift from discipline to digital infrastructure opens up the possibility of some inventive configurations, building unforeseen conjunctions and new spaces.

As artists it is novel to think that in observing, drawing or modelling we might be capturing, analysing and processing data. We are familiar with copies, traces, imprints, indexes. We know about practices of auto-composition that have a random or machinic element. Our work has been infected by the spaces of computer games. Military technologies have changed the configuration of spatial and pictorial practices, re-imagining the thresholds of ‘virtual’ and ‘real’. Complex, contradictory and fractured positions are generated, focusing viewer and author into the same space.

A 3D printer can realize impossible objects, those which cannot be cast otherwise; it can also make available replica objects to hold. Combined with traditional processes, it means that a fragile object can be 3D scanned, printed and then cast conventionally in the foundry to produce a material facsimile without damage or loss. It also allows objects produced by software to be transmitted and materialized.

For both artists and archaeologists alike, our image world is constructed by film industry CGI, advertising manipulations, social networking data bodies, spectacular projection mapping. We see an increasing use of consumer software to model spaces and objects, producing multiple camera positions, complex remixed spaces, armatures constructed as frameworks and skins. 3D scanning as a drawing tool allows us to insert complex representations into these geometric/axonometric worlds, and vice versa. It seems vital to make the technology that produces such strong and enchanted worlds available, hackable, and demystified.
4. THE EVENTS

The *Making History Project* is primarily event-based. A series of major events have been carried out since April 2012, and are outlined below. With each event the collaboration develops further and additional ideas are explored and new pathway possibilities drafted.

4.1 ACRG Lab Workshop: 3D Technologies for Artists

A workshop organised by project members from the Archaeological Computing Research Group with attendees from WSA and the ACRG. The workshop was designed to encourage inter-disciplinary thinking on the potential uses of 3D technology for exhibition design for art and cultural heritage. Attendees formed groups and designed a hypothetical exhibition featuring 3D technologies such as 3D printing, projection mapping and computer graphics. The event drew upon the skills of attendees which included; exhibition design, sculpture, art history, printing, 3D graphic design, web design and theoretical approaches to art and archaeology.

![Figure 1: Students and staff from WSA worked in teams to produce maquettes for 3D technologies use within cultural heritage themed exhibitions.](image)

4.2 WSA Workshop: Printing for Archaeologists

Archaeology staff and PhD students spent a day in the Printmaking Workshop at WSA as part of our interdisciplinary exchange. The group brought images from their research including some from Basing House where our joint project will be conducted over the coming year. These images were layered up in screen-print to reinterpret the data files and offer new possibilities for composition and meaning. Copper plate etchings were made with the tools and objects that the archaeologists use in the field, leaving traces of trowels, measuring tools, underwater markers. The fingerprint and footprint appear regularly as traces in the sensitive surface of the soft ground, alongside contours from maps, details of diagrams, and archaeological illustration.

As a counterpoint, WSA staff and students made a study visit to the Victoria and Albert Museum in London. Students were asked to locate, observe and describe historic artefacts in the museum display cases. The same objects were then reinterpreted by another member of the group, producing different takes on the same material and making visible different aspects in which an object might be viewed.

![Figure 2: PhD Archaeology students with WSA staff.](image)
4.3 Portus Head: Art and Archaeology

The Portus Head is a carved marble head, excavated in 2012 by the University of Southampton team working at Portus, the ancient port of Rome. It is an iconic object, and has become the symbolic representation of the University’s involvement in the project. Scanned and data captured by the ACRG and 3D printed by WSA, it is the first collaborative project between the two departments. Multiple copies were printed in various scales and were handled, discussed and disseminated throughout the university and externally. A print of the head was reformed as part of an artistic practice and a film of the 3D print process; the layering of the resin powder within the print bed and then the excavation of the 3D print became poetic counterpoints to the original excavation of the Portus head.

4.4 Demo at WSA End of Year Show

Our first semester this year was spent reconstructing a C19 3D prototyping apparatus. Reconstruction is more familiar as a speculative practice in archaeology, but has proved a valuable pedagogical tool in art school. Students worked collaboratively to investigate and construct an ambitious, functioning apparatus, deducing its operation from photographs and period descriptions. In the process they learned construction techniques, documentation skills and research methodologies. Archaeology staff joined us for the final presentation with their RTI dome, a contemporary take on the same multi-viewpoint technology.

Figure 3: Ian Dawson’s work on display in the Humanities building at the University of Southampton.

As part of the demonstration organised at WSA, artworks created by art students were recorded by ACRG PhD students. They used a form of computational photography, Reflectance Transformation Imaging (RTI), as an experiment in the re-use of archaeological photographic

Figure 4: Art students simultaneously trigger camera, photographing a model seated inside the photosculpture apparatus. Photograph courtesy of Lesley Blakelock.
recording techniques for the recording and sharing of three dimensional artworks.

4.5 Stored Collections Visit

WSA students and staff visited the stored collections of the Hampshire County Council Museums Service. Students were introduced to the stores, their taxonomy and structure as well as questions of collecting, archiving and accessioning artefacts. The day was framed around ways of looking at objects, collecting information and technical mediations. Groups rotated through RTI, 3D laser scanning and making drawings of objects from the Basing House site. Interestingly, each process initiated different ways of handling the objects as well as documenting them, producing different forms of attention in the students. The artefacts themselves were very variable in their material qualities. The carved stone will inspire a new workshop at WSA.

Figure 5: A model poses inside the photosculpture apparatus. Photograph courtesy of Matthew Goodsmith.

Figure 6: Recording art student artworks at WSA.

Figure 7: WSA students laser scan an artefact at the Hampshire County Council Museums Service collections stores in Winchester. Photograph courtesy of Alick Cotterill.
4.6 ‘How to’ Advice for Art/Archaeology Transition Students

An unanticipated outcome of the Making History Project has been the increase in requests from students for career or study advice relating to subjects outside of their originating discipline. Undergraduate art students have asked for advice on taking up postgraduate studies within archaeology, and postgraduate archaeology students have requested training in art techniques to complement their research.

4.7 Basing House Excavation

One major event planned as part of the Making History Project was a collaboration between art and archaeology students in July and August 2013. During an undergraduate archaeology student training excavation at Basing House in Hampshire, art students and staff worked with archaeology students and staff to produce creative interpretations of many of the artefacts and features at the site, including decorative stonework, corbels, standing building remains, earthworks, and the excavations themselves.

The outputs involved the use of technologies such as computational photography (including photogrammetry and RTI) along with on-site workshops introducing 3D technologies software (such as 123DCatch and Google SketchUp). Sub-projects included the development of models for printing and 3D projection that are based on laser scans of a large LEGO model of a reconstruction of the site from the sixteenth century, and recording corbels using photogrammetry for the reworking of meshes using 123DSculpt. There was also an Artist in Residence scheme over the duration of the excavation, giving art students an opportunity to work closely with archaeologists to make alternative recordings of the excavations.

4.8 Exhibition at Winchester Discovery Centre, Milestones Museum and Willis Museum

Outreach and education events for the public will be organised by archaeology and art students regularly, with an emphasis on working together to produce reusable outreach activities for Hampshire County Council Museums Service. Students will work with the county Conservator and Keeper of Archaeology, as well as specialists from the curation and exhibition team at Hampshire County Council. One major outreach event is a planned exhibition of the outputs from the students’ time at Basing House, scheduled to be staged late summer at various cultural heritage exhibition spaces. The exhibition will have components that will be on display at both WSA and the ACRG computer laboratory.

4.9 INTECH Science Centre After Dark

Public Outreach events are planned throughout the excavation at Basing House for archaeology and art students to experiment with producing models and images using 3D technologies for art and archaeology subjects. In addition to this, a one-off experience is scheduled at INTECH Science Centre’s After Dark event, targeted at adults wishing to learn more about science in creative ways. Working together, archaeology and art students will develop activities to introduce 3D technologies to visitors to the event.

4.10 Development of Art/Archaeology Course

Collaborations have led to a planned course to be jointly taught by WSA and Archaeology staff, exploring the relationship between art and archaeology.
5. REFLECTING ON COLLABORATION

5.1 In Each Other’s Shoes

One of the most effective dimensions of the collaboration has been the opportunity to engage in the working practice of a separate discipline. Activities such as the workshops held at WSA and ACRG have given researchers the chance to work with techniques and concepts with which they are not necessarily familiar and to reflect upon their own disciplinary practice from the outside. The value of these activities to participants is demonstrated in the range of collaborations which have emerged from initial workshops and activities. These have led to several instances in which researchers and students have modified or adapted their working practice in line with new insights. Many of these activities are outlined above, including the creation of collaborative exhibitions such as the Portus Head installation and the print exhibition at the Vice Chancellor’s residence. These activities will culminate in the first University of Southampton joint Archaeology and Art Summer Fieldwork Project at Basing House.

5.2 Outputs and Communication

Communication sits at the heart of both art and archaeology. Communicating using 3D technologies (or communicating the ways in which we use them) must now be at the heart of both disciplines.

Established archaeological modes of representation have been called into question through collaboration with WSA. Artists might be equally interested in the failures of communication, glitches, noise, overloads, and the impulsive, while archaeology tends to focus upon the use of established modes of representation in order to convey information. Considering archaeological processes from the perspective of the arts has had a significant impact upon archaeological practice and has led to development of a more critical approach towards the capture, processing and dissemination of data. An example of this has been the processing of 3D data. A 3D scan or printed model has to be cleaned up in the same way that a bronze cast must be worked on after initial casting. The Portus Head installation provided a vehicle through which to explore this observation and to convey these ideas to the public. The exhibition was based upon the notion that irritation can be as useful as rapport. In providing a visual counterpoint between conventional archaeological representation and a highly modified, personalised version of the same object, the exhibition invited the public to consider the nature of mechanical reproduction using 3D printing while also calling into question normative modes of archaeological representation.

This is just one example of the disruptive influence which collaboration can have upon working practice. Collaboration has offered participants the chance to question the technical inscription of cultural objects as models, projections, and simulations, and to explore these ideas in a public forum communicating with a wider audience than would ordinarily be the case. The collaborative development of outputs which are designed to engage audiences outside of art and archaeology will remain at the heart of the Making History Project and will expand as the Basing House fieldwork project begins this summer.

5.3 New Methodologies

Archaeology and art both have established and conventionalised methodological approaches. One of the dangers implicit in the adoption of new technology is that established methodological approaches are uncritically applied in scenarios where they are inappropriate or do not adequately confront the demands or exploit the possibilities of new technology. A central challenge of technological innovation in art and in archaeology is to ensure that methodological approaches are appropriately modified while also avoiding uncritical
acceptance of the methodological, cultural and conceptual baggage which often accompany technological change. Sterne (2003) argues that the best way to avoid succumbing to the accepted wisdom and hype which surrounds new technology is to adopt a reflexive approach to methodological development; much like the reflexive sociology championed by Bourdieu. The *Making History Project* should be seen as an attempt to enhance the capacity of art and archaeology to be receptive to the benefits of 3D technologies while remaining critically aware and so (we hope) remaining partially immune to the blind methodological alleys which technological novelty and/or methodological dogma can lead. There is great creative potential in the development of hybrid methodological approaches through critical interdisciplinary discourse. Each of the events and sub projects which constitute the *Making History Project* has led to the development of innovative and self-aware methodological approaches. The scale, scope and lifespan of these methodologies is highly variable and often highly unpredictable. It will perhaps be possible to judge the degree to which these approaches are successful in retrospect, but for now the goal of this project should be to continue to provide a space within which these interactions can continue to take place and within which ideas can be developed, expanded, and adapted.


Notes:


For those unfamiliar with Linked Data, a selection of introductory material is referenced below.[1] A full understanding of the technology is not necessary to engage with the presentation, but knowing the general background would be advantageous to get the most benefit. URIs for the topics referenced in the presentation are also provided after the pdf if they are not on the slides.

### Recorded Presentation

![RDF Recorded Presentation](https://vimeo.com)

**Editors’ Note:** If you are reading this as a PDF on a device with an internet connection you can view the video [here](https://vimeo.com).

### Slides

![RDF Slides](https://vimeo.com)

**Interactive Features.2 RDF Slides**

---

**Introduction**

The topic of the Semantic Web, more recently rebranded as Linked (Open) Data, has been discussed in digital humanities and related disciplines since its inception. Now, more than a decade later, the technology has had more than enough time to run through the hype curve towards the Plateau of Productivity. But where are we really in that timeline? Some sectors have embraced it, and others are catching up quickly, but significant obstacles to adoption remain.

This presentation looks at the major difference between Linked Data and more traditional document-centric technologies, contrasting the benefits with the difficulties in an effort to provide an explanation as to when it is valuable and when the overheads are currently too high. It is intended as an overview to enable the audience to decide for themselves when Linked Data is appropriate for their use case, and what challenges they may run into along the way. Readers may either watch the slides with the recorded presentation or view the slides independent of that recording.
Editors’ Note: If you are reading this as a PDF on a device with an internet connection you can download the slides here.

Slide References

• (Slide 5) “SPARQL 1.1,” http://www.w3.org/TR/sparql11-overview/
• (Slide 12) Sanderson, R. Ciccarese, P. Van de Sompel, H. "Open Annotation,” http://www.openannotation.org/spec/core/
• (Slide 32) “What is the range of the HTTP dereference function?” http://www.w3.org/2001/tag/issues.html#httpRange-14

Notes:


Posters

Digital Humanities Keywords: A Collaborative Community Web-based Project
Susan Garfinkel

Networking the Belfast Group through the Automated Semantic Enhancement of Existing Digital Content
Rebecca Sutton Koeser and Brian Croxall

Programming with Arduino for Digital Humanities
Kazushi Ohya

Text Mining Tools in the Humanities: An Analysis Framework
John Simpson, Geoffrey Rockwell, Ryan Chartier, Stefan Sinclair, Susan Brown, Amy Byrbye, and Kirsten Uszkalo

Voyant Notebooks: Literate Programming and Programming Literacy
Stefan Sinclair and Geoffrey Rockwell

Architecture to Enable Large-Scale Computational Analysis of Millions of Volumes
Yiming Sun, Stacy Kowalczyk, Beth Plale, J. Stephen Downie, Loretta Auvil, Boris Capitanu, Kirk Hess, Zong Peng, Guangchen Ruan, Aaron Todd, and Jiaan Zeng

KORA: A Digital Repository and Publishing Platform
Rebecca Tegtmeyer, Dean Rehberger, Catherine Foley, and Ethan Watrall

Textal: A Text Analysis Smartphone App for Digital Humanities
Melissa Terras, Steven Gray, and Rudolf Ammann

TEI Boilerplate
John Walsh and Grant Leyton Simpson
Abstract
As the field of digital humanities has expanded dramatically in recent years, it has also struggled to define—or perhaps more correctly, to redefine—itself across a seemingly divergent set of its practitioners' backgrounds, interests, priorities, methodologies, and institutional settings that yet still have something fundamental in common. From critical code studies to corpus linguistics to tool creation to online pedagogy, digital humanities practitioners surely share the digital and the humanities yet may work within different fundamental paradigms, asking different questions and using different methodologies in the service of achieving very different final goals (Gold, Kirschenbaum). How, beneath the shared surface of this broadly cast "big tent" of digital humanities, can such a diverse community of practice come together in fruitful and mutually beneficial ways? What can we as scholars do to facilitate this convergence in a broad discursive space where the ethic of "more hack, less yack" can sometimes discourage the meaningful exchanges of ideas that still need to happen, the collective negotiations of purpose that so often serve as the source of new insights?

In his seminal book *Keywords: A Vocabulary of Culture and Society* (first published in 1975), Raymond Williams defines his project as "the record of an inquiry into a vocabulary" (Williams, 15) and presents that inquiry as a series of short interpretive essays on carefully selected but commonly used significant words. Between them, his chosen keywords comprise "a general vocabulary ranging from strong, difficult and persuasive words in everyday usage to words which, beginning in particular specialized contexts, have become quite common in descriptions of wider areas of thought and experience" (Williams, 14). They are “significant, binding words in certain activities and their interpretations,” and they are “significant, indicative words in certain forms of thought" (Williams, 15). Keywords,
then, are those terms that contain and naturalize the categories through which we form our ideas; in our daily speech they embody the assumptions through which our views of the world around us emerge.

In Williams’s approach to the meaningful keywords of a community's shared discourse is an implicit challenge to any diverse yet conscientious community of practice: to deliberately and skillfully make ourselves aware of the multiple, hidden, sometimes divergent, and often hegemonic meanings of our basic shared terminology. Our task becomes, like Williams’s, to inquire into our shared vocabulary rather than merely to use it. So, what are our “big tent” digital humanities community’s strong, difficult, persuasive, yet overly-familiarized keywords? They are item-based terms like code, data, object, document, archive, corpus, and collection. They are activities such as digitization, preservation, encoding, visualization, interpretation. They are action words: hack, make, curate, catalog, blog, and tweet. They are metadata; they are keyword. They may also reside as absences, a significant subset of keyword yet to be fully explored (Klein).

The Digital Humanities Keywords project, launching in June of 2013 at http://www.dhkeywords.org, seeks to take up that challenge by creating a shared space and collaborative conversation for working through and exposing the underlying assumptions—the points of agreement, the sites of tension, the unanswered questions—of the still-emergent and surprisingly complex digital humanities community, through focused attention to the building blocks of its shared discourse.

Like other keywords projects started in the wake of Williams's work (Bennett, Grossberg and Morris; Burgett and Hendler; García and Faherty), Digital Humanities Keywords will be a multi-authored work, but in addition will also engage the flexibility and collaborative features that many peer-based digital humanities projects now offer. As an open access, lightly edited, and openly peer-reviewed Web-site-turned-publication, Digital Humanities Keywords will function not only as an information resource, but as a site of active dialogue and a temporal record of shared communal content as it emerges over time. The project provides guidelines but no hard-and-fast-rules for contributions and it openly solicits beneficial interventions. Comments and dialogue and feedback will reside permanently side-by-side with the primary essays created for the site. Not so much a "how to" as a "how to think about," the Digital Humanities Keywords project explicitly seeks to bring some well-considered yack back to hack, in ways that take fullest advantage of the digital humanities community's existing strengths in open-source online collaboration.

The Digital Humanities Keywords project is sponsored by the Digital Humanities Caucus of the American Studies Association, but is offered to the entire Digital Humanities community with no expectation of an American Studies focus. The current poster presentation works to introduce the project to a broad cross section of the digital humanities community and to solicit participation—which is actively invited at all levels of involvement.

Originally presented by Susan Garfinkel at DH2013 on July 17, 2013.

References


Kirschenbaum, M. (2012). "What is Digital Humanities and What is it Doing in English Departments?" *Debates in Digital Humanities.* University of Minnesota Press. 3-11.

Williams, R. (1985). *Keywords: A Vocabulary of Culture and Society.* Oxford University Press.


Bennett, T., L. Grossberg, and M. Morris (eds). (2005). *New Keywords: A Revised Vocabulary of Culture and Society.* Blackwell.
Networking the Belfast Group through the Automated Semantic Enhancement of Existing Digital Content

Abstract

There is increasing work on and interest in social networks in the digital humanities community (Meeks 2011). Analysis is frequently done on digital content — including images (Akdag Salah et al. 2012); email (Hangal et al. 2012); and citation networks (Visconti 2012) — because the data lend themselves to aggregation, conversion, and analysis. Yet despite this flurry of activity, the possibility exists for an exponential jump in network analysis. After all, the holdings and catalogs of galleries, libraries, archives, and museums (GLAMs) include traces of vast paper-based networks, but the data are locked away in forms that don’t easily lend themselves to analysis. What if we could open up that content? In this poster, we will report on an attempt to provide tools for archivists to expose the information embedded in the descriptions of their collections as well as a test case for analyzing that data: an examination of the networks of the Irish poets collectively known as “the Belfast Group.”

Our goal is to develop software tools and design a workflow to enhance TEI and EAD — documents that are already commonly created and maintained by archivists and text centers — without radically increasing the time and effort involved. The software tools (http://github.com/emory-libraries-disc/name-dropper) consist of a plugin for the Oxygen XML editor and command line scripts that will, first, make use of DBPedia Spotlight to identify and annotate recognized names and other resources within the text and, second, connect to linked-data systems (starting with the Virtual International Authority File [VIAF]) to provide authoritative, scholarly identifiers.[1] The scripts will allow technical users to inspect and tune the results or to automatically tag high-certainty resources, and the plugin will provide a user-friendly interface to review and accept suggested names while editing a document. The enhanced documents should provide significant benefits to GLAMs, allowing them to connect disparate
types of content (e.g., digitized texts or photographs from an archival
collection) and augment it with data from other linked data systems.
Furthermore, the enhanced documents will make it possible to expose
these data in more machine-readable and research accessible formats.
Our tools and workflow could be applied to resources held by different
archives (for a different approach, see Blanke et al. 2012). What’s
more, enhancing these documents helps GLAMs provide a means for
researchers to do non-consumptive, social network research on the
metadata of collections that might otherwise be closed or problematic
in other ways (e.g., restricted correspondence from living authors).

Although our tools are not yet complete, we have already begun
preliminary visualization and analysis of network relationships using
data that mirrors what we will generate automatically by Summer
2013. The difficulties of defining “the Belfast Group” make for a
compelling test case for our attempt to understand networks via data
that are newly machine readable. The Group is a contentious network
since the label has been variously applied to a weekly writing workshop
that ran from 1963-1972, the most famous poets who attended that
workshop — including Seamus Heaney, Michael Longley, and Paul
Muldoon — or more loosely applied to all of the writers who “put
Belfast on the literary map” (Clark 6). The significance of the writing
workshop is debated by critics and often rejected by the poets
themselves, sometimes vehemently. In contrast to a more formalized
group, some scholars identify “an informal community” of poets
evidenced by their letters, promotion of each other, and poems
dedicated to each other (Drummond 32), connections which are richly
documented by archival materials held at Emory University.

Using preliminary data manually generated from a subset of the
correspondence EAD, our data suggests a wider set of connections in
the Group than traditional scholarly approaches. The latter selectively
emphasize the relationships of the most prominent authors and the
role of the writing workshop (see fig. 1). Since our data is based on a
much larger set of artifacts, as well as their complete metadata, we find
that the locus of poetic activity in Belfast is not so oriented around the
workshop (see fig. 2). Once we collect the full dataset via our
completed tools and workflow, we will compare it with models
generated by traditional scholarly methods, to identify significant gaps
and discrepancies in either model.

Figure 1. Graph of relationships inferred from Heather Clark’s
Ulster Renaissance. Nodes are sized by degree and colored
by hub score. The writing workshop is the strongest hub; the
trio of large nodes represent Michael Longley, Derek Mahon,
and Seamus Heaney.
Providing not only this new analysis of the Belfast Group’s network and a report on the development of our tools, our poster presentation at DH 2013 will also include a hands-on demonstration of the software tools and interactive visualizations of network data.

**Notes:**

[1] It is in the use of existing systems (DBpedia) and vocabularies (VIAF) that distinguishes this project from the Entity Authority Tool Set (EATS), which involves setting up and maintaining one’s own authority server. See Litta Modignani Picozzi, Norrish, and Monteiro Vieira (2012).

**References:**


---

Figure 2. Relationship graph based on preliminary correspondence data, sized and colored as in figure 1. Based on this data, the writing workshop does not function as a hub at all, and Paul Muldoon becomes the largest node.

Originally presented by Brian Croxall and Rebecca Sutton Koeser at DH2013 on July 17, 2013.


Visconti, A. View DHQ: Citation Network Visualization for Digital Humanities Quarterly. http://digitalliterature.net/viewDHQ/. (accessed 1 November 2012.)
Editors’ Note: If you are reading this as a PDF on a device with an internet connection you can download the slides here.

Background
The department of library, archive, and information studies at Tsurumi University provides courses on computer science for humanities students. The courses include one introductory and two intermediate programming courses. There are project-based courses on Digital Humanities instead of an advanced programming course. In the introductory course, Scratch was used and in the intermediate courses, Java has been adopted.

Difficulties in teaching programming in an introductory course to humanities students have been to let students find interest or enjoyment in (1) symbol manipulation and (2) grouping tasks with functions or other similar units.

Humanities students tend to expect big results in programming. They are disappointed at small results from source codes in exercises. It is usually difficult to expect them to find interest in a small result from symbol manipulation in introductory programming.

As far as our experience goes, humanities students seem to struggle to envision the existence of a computational world in their minds. The formats and abstract behaviors of typical programming patterns appearing in structured programming such as assignment, iteration, condition, and flow control themselves are not difficult for humanities students to understand. The grammar is easy to comprehend. The problem they face is to understand the ways to use them as substantial components making up the whole code in actual programming.\[1\]

Embedded System: Arduino
In order to respond to the aforementioned difficulties or problems, we set two requirements for the programming environment:
1. letting students be interested in something moving, and
2. letting students be satisfied with small results.

Then, we decided to introduce embedded systems to an introductory programming course. We expected embedded systems to bring the real world into a scene of learning programming.

Current students have been immersed in digital or virtual environments from an early age. They are not easily satisfied with computational results on screens. On the other hand, interestingly, they show interest in physical phenomena even if it is slight.[2] They are very sensitive to physical stimuli. Embedded systems can be expected to let students be interested in something moving that may be small. And, it can contribute to helping students envision a computational world in their minds.

As an embedded system, we adopted Arduino, which provides a good developing environment:

1. there is no need to prepare a device to install native codes to ICs,
2. there is an easy IDE based on Processing, and
3. the price of Arduino is affordable.

A concern we noticed about introducing Arduino was that students have to learn about electricity to some extent. It is an unfamiliar subject to humanities students.

The IDE for Arduino on Processing provides simple descriptive rules reminiscent of BASIC, and easy to view and write grouping tasks in methods. We expected this feature would make it easy for students to concentrate on finding categories of processes and making groups with functions.

Course Design

One semester consists of 15 classes, which are lectures on and practices of programming. The practices consist of four themes: LED handling, variable register, sound handling, and binary display.

In all practices, we used only three circuit patterns based on a voltage divider. The typical four programming patterns in procedural languages were learned with eight source codes using one circuit pattern of an LED. We used a variable resistor as an input device and a game controller in learning structural programming. An array is not a difficult topic in programming, but it is not easy for students to understand the usefulness of the array. Arrays working as music scores to handle sound seemed to be a good example. We think that even though the topic of a binary digit may not be useful in Digital Humanities, students should learn it to feel the philosophy behind symbol manipulation. We used bit operation to control LEDs of a binary digit display, and provided a chance to learn binary digit.

Observation and Future

Fortunately, many students seemed to find enjoyment from a small result in physical programming and grouping tasks. However, as we expected before this experiment, learning electricity seemed to be difficult for students. For example, Ohm's law was difficult for students who had learned it in junior high school. It might be possible to teach electronic circuit like LEGO block without any explanation of theory or background knowledge. However, learning scientific theories is inevitable in science education. We are planning to devise course materials to reduce the offset of learning electricity for next year.

Originally presented by Kazushi Ohya at DH2013 on July 17, 2013.
Notes:


References:


Processing. [http://processing.org/](http://processing.org/)

Scratch. [http://scratch.mit.edu](http://scratch.mit.edu)
John Simpson, Geoffrey Rockwell, Ryan Chartier, Stefan Sinclair, Susan Brown, Amy Byrbye, and Kirsten Uszkalo

Text Mining Tools in the Humanities: An Analysis Framework

Abstract[1]

The most extensive compendium of text mining tools to date includes 71 tools and summarizes each based on ten criteria.[2] While extensive, this listing of tools and their properties is general in its review criteria and does not offer any testing-based observations to help users assess actual usability. Humanists looking to try text analysis, visualization, and mining tools for research need better information that is relevant to their needs and reviews of tools that help them make choices. This poster presents the testing framework developed for the TAPoR 2.0 portal reviews. The poster will cover:

1. The need for tool reviews
2. The information gathered about tools
3. The testing and reviewing process
4. Conclusions about the state of text tools

The poster will be accompanied by a demonstration of TAPoR 2.0 so that users can see the reviews in context.

1. The Need for Tool Reviews

A humanities researcher new to computing methods and looking for reviews of text tools by peers on the internet is going to be disappointed. There is nothing like The New York Review of Books for tools, though in the early days of humanities computing you could find short announcements about tools in journals like Computing in the Humanities. We, however, believe that certain text tools are intellectual contributions to the field that should be reviewed not just to help people choose what tools to use, but also as a way of engaging these tools in a dialogue around computer-assisted interpretation.[3]
While there are individual blog entries about tools scattered across the web, each is from the perspective of a single user with an entirely different dataset, making comparison difficult. If we want to make computing methods accessible and encourage colleagues to use tools, we need a more systematic approach. This is especially true of text mining tools that can’t simply be tried with a text at hand.

2. Information Gathered about Tools

TAPoR 2.0 is a portal for text analysis, visualization, and mining tool discovery and review. TAPoR 2.0 is a complete redevelopment of the original TAPoR portal that has focused the portal on discovery and review instead of trying to provide access only to web services.[4] As part of the redevelopment of TAPoR 2.0 we used a persona/scenario usability design approach to identify attributes by which users might want to discover tools.[5] Further, we built TAPoR 2.0 so that editors can add new attributes without the database having to be reprogrammed. Some of the attributes we currently record for tools include the author(s), ease of use, type of analysis, type of license, and so on. We also provide links to related tools. Our poster will be accompanied by a demonstration of TAPoR 2.0 so that visitors can explore what we have and how we represent it.

3. The Testing and Reviewing Process

Recording basic information about tools is not enough, especially for sophisticated text mining tools like Mallet that take time to learn and that can be used in different ways. With text mining tools users need longer narrative reviews. For this reason we developed processes for testing and reviewing tools. For simpler text analysis and visualization tools this involved developing a set of different texts with which to test tools so we could compare their use. For text mining we had to go further and are working with the CWRC project (Canadian Writing Research Collaboratory) to develop a number of literary corpora with experts we can draw on to help assess the value of results. As of this writing, we have three corpora drawn from the Orlando Project and one of Victorian children’s literature. We expect to have two more by the time of presentation. The poster will discuss the criteria used to develop these open test corpora.

The reviews take the form of comments that have been pinned to the top of the list of comments available. This allows others to leave comments, though we haven’t seen much activity by people not connected to the project (with the exception of spammers who seem to feel there is a connection between text analysis tools and various stimulants). We have developed guidelines for reviews so as to make
them accessible and comparable. The poster will outline our guidelines.

4. Conclusions from Testing and Reviewing

Having tested and reviewed a variety of tools and text mining systems, we see some common barriers to access. Most of these tools have been developed for use by the developers and are poorly documented for people not involved in the development. Further, many tools, including those we are involved in, are in continuous development, resulting in documentation that is out of date. We will therefore end this poster with lessons learned while testing and reviewing text mining tools, with particular attention to removing usability barriers for novice users.


Notes:

[1] The authors would like to thank and acknowledge support from both the INKE Research Group and the Text Mining & Visualization Project, funded by the Social Sciences and Humanities Research Council of Canada.


Editors’ Note: If you are reading this as a PDF on a device with an internet connection you can download the slides here.

Abstract

Some thirty years ago Donald Knuth, a computer scientist, proposed literate programming as a better way of organizing narrative and code (1984). Knuth argued that more emphasis should be placed on explaining to humans what computers are meant to do, rather than simply instructing computers what to do. Knuth was especially interested in weaving together macrostyle code snippets with prose that provided a larger narrative context, not merely functional comments of specific lines of code that are the distilled remnants of an intellectual process.

Literate programming has been more influential in theory than in practice (Nørmark), despite several utilities and environments including Mathematica, Knuth’s (C)WEB, Sweave for R, and Marginalia for Clojure. Perhaps the exigencies of programming in the real world correspond poorly with the vision of Knuth of the programmer as author: "the practitioner of literate programming can be regarded as an essayist, whose main concern is with exposition and excellence of style" (1992, 1). However, that balance of essayist and coder strikes us as perfectly appropriate for the digital humanities, a natural blend of the expression of intellectual process with the exposition of technical methodologies. The prose can gloss the code, or vice versa, in a symbiotic relationship that serves to strengthen an argument and demonstrate its own workings.

One of the most significant potential benefits of the literate programming paradigm is pedagogical: these works can both explain an interpretive insight and present the methodology for reproducing the data or results that were part of the process. Many widely-read digital humanities blogs already present these characteristics of exploration, explanation, interpretation, and step-by-step instructions (see for example blogs by Ted Underwood, Benjamin Schmidt, Lisa...
Rhody and Scott Weingart). Literate programming can be more self-contained and more useful for those learning new methodologies and new programming techniques. This is about the principles of literate programming, but also about the potential for increasing programming literacy.

This poster will introduce Voyant Notebooks, a web-based literate programming environment designed for the digital humanities (see Appendix A). There is already a working prototype, and we anticipate having a more feature-rich version available by July 2013. Voyant Notebooks inherits many of the characteristics of the Voyant Tools environment, including a concern for usability and flexibility (researchers and students should be able to use it with minimal or no training and with their own texts of interest). Voyant Notebooks also addresses one of the main weaknesses of Voyant Tools: the fact that most tools are constrained by assumptions about how they would be most commonly used. For instance, the Wordle-like (word cloud) Cirrus tool is designed to show the top frequency terms from a corpus or document; but what if the user instead wants to visualize the top frequency nouns, or people, or repeating phrases? All of that functionality could be built into the tool, but possibly at the cost of usability (endless menus and options), and it could still never address all of the possible use cases. Voyant Notebooks, by contrast, empowers the user to customize some of the functionality by leveraging the analytic capabilities of the Voyant back-end and the visualization interfaces in the front-end (like Cirrus). Our poster will have two parts, a) a usable demonstration on one or more laptops and b) a poster that illustrates how Voyant Notebooks implements Knuth’s concept of literate programming. In addition to these conceptual aspects, the poster will outline technical details about the Voyant Notebooks prototype for those interested, including the technologies used for both client-side (browser) and server-side components. Some of the technical challenges that will be described include:

- managing the flow of code execution in an asynchronous architecture
- using web workers to avoid browser freezes during longer executions
- mitigating the security risks of user-defined and persistent Javascript code
- code variable scoping across editor instances and window components
- embedding of Voyant tool panels (visualizations) and other services
- developing a flexible API for different programming levels and styles
- developing an API that includes both client-side and server-side operations
- ensuring efficiency of repeated code snippets during writing and viewing

And of course, visitors to the poster session will be warmly encouraged to play with Voyant Notebooks.
Appendix A: Mockup of Voyant Notebooks (previously called Voyeur Notebooks).

![Mockup of Voyant Notebooks](image)

Figure 1: Mockup of Voyant Notebooks

Originally presented by Stéfan Sinclair and Geoffrey Rockwell at DH2013 on July 17, 2013.

References:


Abstract

The HathiTrust Research Center (HTRC) is a collaborative research center that provides Digital Humanities researchers access to not only millions of volumes from the HathiTrust (HT) digital library, but also cutting-edge software tools and cyber infrastructure to perform advanced computational analysis over the corpus at an unprecedented scale.

The corpus at the HTRC currently consists of over 3 million public domain volumes and anticipates access to an additional 6 million in-copyright volumes. In their raw form at the HathiTrust, these volumes are stored as files on special hardware using an internal Pairtree structure. The internal HathiTrust structure is optimal for its primary function of the digital page image delivery to digital library patrons for viewing; however, it does not support well the large-scale computational analysis which is the primary function of the HTRC. Navigating the Pairtree and uncompressing the text data would encounter major performance and scalability issues. While researchers from other scientific communities have been addressing aspects of the “Big Data” problem with success, the large corpus that HTRC hosts to support computational analysis presents a unique setting in that it consists of a massive number of small text-based files, whereas most solutions from the scientific communities are tailored towards large files and non-text-based content. In this poster, we will present the approach the HTRC takes to solve this problem — the HTRC keeps the Pairtree only for the purpose of synchronization with the HT, and processes and pushes the volume data from the local Pairtree to a NoSQL storage cluster using Apache Cassandra hosted on conventional hardware during the ingest process. In order to balance the data store and ingest workload, the developers at the HTRC and the HT also devised a very simple yet effective way to parallelize the rsync of the
single source Pairtree at the HT on all Cassandra nodes by starting rsync at lower branches instead of at the root.

The use of a NoSQL cluster adds more complexity to the architecture than traditional file systems, but such complexity is transparent to the Digital Humanities researchers as most of the HTRC components with which user algorithms have interaction are RESTful web services, such as the Data API for accessing the data. The HTRC uses Blacklight, an open source bibliographic search and display interface, backed by a Solr index, to let users search for volumes for analysis and create collections. To apply analytical techniques to the data, a user may choose from a number of provided algorithms from the web portal, including SEASR/Meandre flows. In addition, the HTRC is actively researching and developing a secure computation environment (dubbed the Sloan Cloud) to support large-scale non-consumptive research over copyrighted volumes, and an experimental release is scheduled for end of March. This Sloan Cloud will allow researchers to deploy their own analysis algorithms against a corpus like the HT data, and to save intermediate data for later reuse, as well as to include custom worksets for the computation. We will present our early findings of the experimental Sloan Cloud and hope to get feedback from the digital humanities research community.

Editors’ Note: If you are reading this as a PDF on a device with an internet connection you can download the slides here.

Abstract

MATRIX has developed an open source application that cultural and educational institutions can use to preserve digital materials and display them online. The application, KORA (http://kora.matrix.msu.edu), is particularly well-suited for working with digital objects of all media types and for easily creating displays of these objects in multiple ways that enhance their educational and research value. MATRIX, a humanities computing research center at Michigan State University, has built and enhanced this application in the course of seven years of research with support from the National Science Foundation.

Designed for long-term preservation and access, KORA includes unique features that meet two important needs of institutions that have limited technological resources: (1) simple design of the digital repository and the ingesting of data, and (2) the ability to display digital materials online in diverse ways, such as image galleries, multimedia educational activities, or story chapters.

The KORA architecture is unique in that it can accommodate any set of metadata schemes (or tables) in individualized digital libraries. Users can easily create metadata elements (database fields) using a simple point-and-click interface, select the type of form control for each element (e.g., required formats for date, URL, file upload, etc.), and then determine whether the element is required for each record, whether it should appear in search returns, and other features. KORA then automatically generates storage structures, ingestion (data entry) forms, and validation requirements for each metadata scheme.

Because the back-end of projects can be created in minutes by people without technical training, the overhead for getting projects started is reduced immeasurably compared to beginning with a blank SQL or
other database. And because KORA is an online application, multiple users can develop a collection from separate locations at the same time. Also, KORA can ingest materials from any standardized repository and can output XML that can be harvested by these repositories.

KORA also includes an easy-to-use "associator" tool for creating relationships that combine objects of various media types. As demonstrated by diverse websites built with KORA, many creative displays are possible using this open source application.

In keeping with the need to ensure authenticity and integrity of files ingested into KORA, as described in the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) guidelines, automatic fixity checking has been built into KORA to verify that data has been kept free of tampering and corruption. Long-term access to digital material can be assured by storing this preservation information in the digital repository, as described by the ISO Reference Model for an Open Archival Information System (OAIS) model and Preservation Metadata: Implementation Strategies (PREMIS).

**New Release: KORA 3.0**

KORA 3.0 will be released this spring with a host of new features, including many major changes:

- all new user experience design;
- independence from MYSQL so it can be used with other database management systems;
- enhanced Multilanguage capabilities;
- rebuilt on Symphony, KORA will have enhanced plugin capabilities;
- and many more features.

Originally presented by Rebecca Tegtmeyer, Dean Rehberger, Catherine Foley, and Ethan Watrall at DH2013 on July 17, 2013.
Abstract

This poster introduces Textal, a text analysis application for iOS, and text analysis service infrastructure, which is currently in development at UCLDH and UCLCASA and will be freely available beginning Summer 2013. This poster will present findings evaluating the development, launch, and reception of the app, indicating how smartphone technology can increase the potential for public engagement within the Digital Humanities.

Textal (soon to be launched at www.textal.org, currently on twitter at @textal) will be a freely available smartphone application which allows users to create, share, and explore word clouds of a document, website, or tweet stream. Those in visualization and Digital Humanities have tended to sneer at the popular use of word clouds, given we are used to applying robust text analysis tools (such as http://voyant-tools.org/).[1] However, Textal turns word clouds into an intuitive, visually-oriented interface: once a Textal of a chosen text is generated, users can click on words to access underlying statistics, such as frequency and collocates, and so we believe that the pinch, stretch, and click potential in smartphones, along with our judicious design, can fix the elements of word cloud visualization which are currently held to be problematic and act as a bridge between those who have never encountered text analysis techniques, and the more detailed approaches undertaken by researchers in Digital Humanities. All Textal visualizations, including word-clouds, graphs, charts, and word lists, can be shared via social media such as Twitter and Facebook, and the resulting interface word clouds will also be available online. Textal is powered by server-side processing of linguistic data (users can submit any text material they want, by URL, or copy, or paste). The resulting sever architecture will also serve as an API for those wishing to carry out on-the-fly generation of text analysis statistics, which can be used in conjunction with other web services.
We envision the Textal iPhone, and iPad, app as a fun text-analysis-in-your-pocket product, which can raise the profile of this technique. We have built Textal with the general audience in mind, to bring Digital Humanities approaches to as wide an international audience as possible (we will be translating the interface into many languages). With an increasing move towards smartphone rather than desktop technologies there is a need to understand how mobile technologies fit within the Digital Humanities remit.\[2\] We believe we are one of the first teams to build, from scratch, a stand-alone app that brings Digital Humanities techniques to a wider, mobile based, audience. (Previous apps, do exist, such as the DH2012 conference app, which is an app based version of the conference programme. Geostoryteller is a platform for history walking tours that allow smartphone users to interact with multimedia historical information as they move around a neighbourhood.\[3\] Others have used augmented reality viewers for historical and archaeological sites (see http://www.dead-mens-eyes.org/), often built on existing commercial platforms. We don’t believe, however, that others have built smartphone apps that allow the user to do much data analysis or processing in the way we describe).

We are building Textal from the ground up using our own server infrastructure, with the app programmed in house in Objective-C. Textal will be available for iOS only, with a plan to build a stand-alone application for use with Apple laptop and desktops. Depending on reception, we may then build an app for other operating systems. Given that we own the infrastructure, we will be able to view and analyze how, why and, when people are using text analysis: we will be tracking use and users, including geo-locating text analysis, to ascertain the potential audience for this type of service and to understand more about the kind of texts people want to analyze, allowing us to undertake a reception study into Textal’s uptake, which will be of great interest to the wider Digital Humanities audience.

Although the app will not be launched until Summer 2013, this is not a promissory abstract: most of the development, including both technical infrastructure, server architecture, and design-work on the app is now complete and at time of submission we are moving into alpha-testing with a core group of users interested in text analysis. This poster will be an up-to-the-minute account of a very recent development in Digital Humanities: what ramifications do apps hold for Digital Humanities as a discipline or a field of practice? We will report using up-to-date statistics generated from Textal as a case study, and demonstrate Textal at the poster session.

Originally presented by Melissa Terras, Rudolph Amman, and Steven Gray at DH2013 on July 17, 2013.

Notes:


**Introduction**

TEI Boilerplate is a lightweight, HTML5 compliant framework for publishing TEI documents. TEI Boilerplate (TEIBP) is designed to bridge the gap between the browser-friendly features of HTML and the semantic richness of native TEI documents.[1]

Although TEI provides mechanisms for describing the design, presentational, and material features of the source document, projects and individual scholars that use TEI are responsible for developing their own methods, or implementing existing solutions, for converting the TEI to a presentation-ready state for the web or print.[2] Two potential paths to reach this goal are:

1. Transforming TEI to HTML using XSLT and styling the HTML output with CSS.
2. Styling the TEI directly with CSS by referencing a CSS stylesheet from within the TEI document.

Both approaches have advantages and disadvantages. Although HTML is the language of the web and is well supported by browsers, HTML’s descriptive capabilities are less expressive than TEI’s. When TEI is transformed to HTML, much of the richness of the TEI is lost in the resulting HTML. However, the browser understands HTML very well and knows, for example, when to initiate retrieval of a document based on certain user events, such as clicking a link. The second option, CSS-styled TEI, delivers the TEI document directly to the browser. However, while the browser may apply CSS to format and style a TEI document, the browser does not understand the semantics of TEI. For instance, the browser does not understand that TEI’s <ptr> and <ref> elements are linking elements.

---

**Editors’ Note:** If you are reading this as a PDF on a device with an internet connection you can download the slides here.
TEIBP bridges the gap between these two approaches by making use of the built-in XSLT (1.0) capabilities of browsers to embed the TEI XML, with minimal modifications, within an HTML5 shell document. Features expected of web documents, such as clickable links and display of linked images, are enabled through selective transformation of a very small number of TEI elements and attributes. Both the HTML5 shell and the embedded TEI are styled using CSS.

TEIBP gives HTML/CSS/JavaScript documents direct access to original TEI content, and it gives TEI documents direct access to the substantial capabilities of HTML, CSS, and JavaScript — the dominant document format, styling language, and (client-side) programming language of the web. TEIBP aims for simplicity and elegance, but it also facilitates complexity and innovation by exposing TEI content directly to the capabilities of JavaScript, the many powerful JavaScript frameworks, and CSS.

Theoretical Motivations

The power of TEI lies in the richness of its vocabulary. But much of that richness and expressiveness is lost in the translation to HTML. TEIBP largely preserves the integrity of the TEI document. Because the TEI document is delivered directly to the browser, that source TEI document — unchanged by any XSLT transformation — can be easily accessed and saved.

Scholars labor over the intricate encoding of TEI documents, encoding that may represent sophisticated readings and analysis. But with the typical XSLT publishing solution, much or all of the richness of the TEI content is lost. Furthermore, the presentation of the document is targeted at the HTML surrogate rather than the encoded TEI document. This results in a conceptual disconnect between the design of the document and the original TEI encoding. By exposing the TEI itself to the browser, one may format the TEI directly, applying intentional design to a sophisticated document model.

TEIBP respects the integrity of the TEI document, and keeps the TEI document central throughout the publication process. TEIBP takes advantage of the separation of form and content inherent in XML, XSLT, CSS frameworks. However, like Liu (2004), Galey (2010), and others, the authors of TEIBP view that separation with suspicion.[3] TEIBP attempts to weaken that separation of form and content in the typical TEI-to-web design and delivery model by largely removing the HTML layer, exposing the TEI-encoded text directly to the browser, and providing scholars with more immediate access to the readings, models, and analysis embedded in the TEI-encoded document.

Our proposed poster will provide an overview of the TEIBP system and explore in more detail the theoretical motivations behind the project.

Originally presented by John Walsh and Grant Leyton Simpson at DH2013 on July 17, 2013.

Notes:
Contributors

Rudolf Ammann is an honorary research associate at the UCL Department of Information Studies and designer at large at the UCL Centre for Digital Humanities. Current projects include a book on the emergence of blogging.

Loretta Auvil works at the Illinois Informatics Institute (I3) at the University of Illinois at Urbana Champaign. She received a MS in Computer Science from Virginia Tech and a BS in Applied Mathematics and Computer Science from Alderson-Broaddus College. She has worked with a diverse set of application drivers to integrate machine learning and information visualization techniques to solve the needs of research partners.

Gareth Beale, Archaeological Computing Research Group, University of Southampton

Nicole Beale is studying a PhD in improving access to cultural heritage using the Web. She is based at the University of Southampton’s Web Science Doctoral Training Centre and the Archaeological Computing Research Group. She is interested in open data, particularly how this can be used by museums and for public archaeology projects and is currently a CHASE Going Digital Scholar. She has a degree in History of Art, a Masters in Archaeological Computing and a Masters in Web Science, blogs at The Cultural Heritage Web, and tweets as @nicolebeale.

Susan Brown is Professor of English at the University of Guelph and Visiting Professor at the University of Alberta. Her research interests include the digital humanities, Victorian literature, and women’s writing. All of these interests inform Orlando: Women’s Writing in the British Isles from the Beginnings to the Present, an ongoing experiment in digital literary history published online by Cambridge University Press since 2006 that she directs and co-edits. She leads the Canadian Writing Research Collaboratory infrastructure project.

Boris Capitanu is a Research Programmer at the Illinois Informatics Institute of the University of Illinois at Urbana-Champaign. His research interests include large scale data analysis, machine learning, data mining, and educational technologies. Boris holds a B.S. and M.S. in Computer Science from University of Illinois at Urbana-Champaign, and a MBA+MHRIR from the same institution.

Ryan Chartier, University of Alberta

Brian Croxall is Digital Humanities Strategist and Lecturer of English at Emory University in Atlanta, Georgia. Along with
developing and managing digital scholarship projects in the new Emory Center for Digital Scholarship, he teaches courses on digital humanities, media studies, and American literature. He has co-edited an issue of Neo-Victorian Studies on steampunk; is co-editing a book on the same subject; is a cluster editor at #alt-academy; serves on the Executive Council of the Association for Computers and the Humanities and the Committee on Information Technology at the Modern Language Association; and is a writer for the group blog ProfHacker.

Ian Dawson, Winchester School of Art

J. Stephen Downie is Associate Dean for Research and a Professor at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. Downie is the Illinois Co-Director of the HathiTrust Research Center (HTRC). He is also Director of the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) and founder and ongoing director of the Music Information Retrieval Evaluation eXchange (MIREX).

Amy Dyrbye, University of Alberta

Catherine Foley is Director of Digital Library and Archive Projects at MATRIX. Foley has considerable experience working with KORA. She has managed NEH-funded projects to preserve and provide access to the public television series American Black Journal and to create a database to provide access to and analysis of multiple datasets about African slaves in North and South America. She manages several digital repository projects with multimedia resources from Africa, including the Community Video Education Trust digital archive, Diversity and Tolerance in the Islam of West Africa, African Oral Narratives: life histories, interviews, folklore & song from sub-Saharan Africa, Pluralism and Adaptation in the Islamic Practice of Senegal and Ghana and Biographies: The Atlantic Slave Data Network.

Susan Garfinkel, Research Specialist, Digital Reference Section, Library of Congress

Steven Gray, University College London

Kirk Hess is the Digital Humanities Specialist at the University Library at the University of Illinois Urbana-Champaign. He holds a MS in Library and Information Science from the University of Illinois Urbana-Champaign and a BA in History from Carleton College. His current research with HTRC focuses on enhancing discovery of the Solr index using a custom implementation of the Blacklight digital library platform.

Stacy T. Kowalczyk is an Assistant Professor in the Graduate School of Library and Information Science at Dominican University. Her research focuses on the problems of research data, big data, and curation, specifically looking at the intersection of social and technical issues. In her current work, she is investigating the research practices of scholars, the lifecycle of research data including data reuse, and the antecedents, barriers, and threats to preservation of research data.

Rebecca Sutton Koeser, Emory University

Louisa Minkin, Winchester School of Art
**Trevor Muñoz** is Associate Director of MITH as well as Assistant Dean for Digital Humanities Research at the University of Maryland Libraries. Trevor holds an MA in Digital Humanities from the Department of Digital Humanities at King’s College London and a MS in Library and Information Science from the Graduate School of Library and Information Science at the University of Illinois, Urbana-Champaign. He specializes in issues related to curation of humanities data including data management planning, systems, and best practices. Trevor is co-editor of a forthcoming guide to resources for Data Curation in the Humanities and co-organized the first Humanities Data Curation Summit intended to promote the development of a sustainable plan for preserving digital humanities research.

**Kazushi Ohya** is an associate professor at Tsurumi University, Yokohama, Japan. He received a M.A. in Linguistics and a Ph.D. in Computer Science, both from Chiba University, Japan. He has researched markup languages such as HyTime and XML, language documentation with linguists studying endangered languages, and experimented on corpora e.g. Old-Map, Manga, and discourse data. He is a member of IEEE, ACM, IPSJ, and ADHO.

**Zong Peng** is a Ph.D. candidate in Computer Science in the School of Informatics and Computing at Indiana University Bloomington. He is a Research Assistant in Data to Insight Center of the Indiana University Pervasive Technology Institute, under supervision of Professor Plale. His current research with the HTRC focuses mainly on the building, deploying and maintaining the index infrastructure for the entire HTRC collection providing security, auditing, and customized functions.

**Beth A. Plale** is a professor of Computer Science as well as the Director of the Data to Insight Center of the Pervasive Technologies Institute and the Director of the Center for Data and Search Informatics at Indiana University. Plale is the Indiana Co-Director of the HathiTrust Research Center (HTRC). She has broad research and governance interest in long-term preservation and access to scientific data, and enabling computational access to large-scale data for broader groups of researchers.

**Dean Rehberger** is the Director of MATRIX and also Associate Professor in the department of Writing, Rhetoric and American Cultures at MSU. Dean specializes in using online technologies and developing educational resources for the World Wide Web. He has run numerous faculty technology and workshops and given presentations for educators and cultural heritage workers from local, national and international audiences. Dean oversees MATRIX project planning, research and development, coordinating many of the grant-funded projects for the Center. His primary areas of research include: information design and architecture; digital libraries, museums and archives; Internet technologies in the classroom; and hybrid learning environments. Dean teaches course for a variety of courses at MSU for Professional Writing Program, American Studies, the Graduate Study in Rhetoric & Writing, and Museum Studies. He also helps to design and develop a number of online course for the department of History.

**Geoffrey Rockwell** is a Professor of Philosophy and Humanities Computing at the University of Alberta, Canada. He received a B.A. in philosophy from Haverford College, an M.A. and Ph.D. in Philosophy from the University of Toronto and worked at the University of Toronto as a Senior Instructional Technology Specialist. From 1994 to 2008 he was at McMaster University where he was the Director of the
Humanities Media and Computing Centre (1994 - 2004) and he led the development of an undergraduate Multimedia program funded through the Ontario Access To Opportunities Program. He has published and presented papers in the area of philosophical dialogue, textual visualization and analysis, humanities computing, instructional technology, computer games and multimedia. He is the project leader for the CFI (Canada Foundation for Innovation) funded project TAPoR, a Text Analysis Portal for Research, which has developed a text tool portal for researchers who work with electronic texts and he organized a SSHRC funded conference, The Face of Text in 2004. He has published a book *Defining Dialogue: From Socrates to the Internet* with Humanity Books.

**Guangchen Ruan** is currently working toward the Ph.D. degree in Computer Science in the School of Informatics and Computing at Indiana University Bloomington. He is a Research Assistant in Data to Insight Center of the Indiana University Pervasive Technology Institute, under supervision of Professor Plale. His research interests include big data analysis and management, distributed data intensive computing, data visualization, and data mining.

**Dr. Robert Sanderson** is an information scientist in the Research Library at Los Alamos National Laboratory and previously a Lecturer in Computer Science at the University of Liverpool. His research focuses on the areas of scholarly communication, especially with regards to digital humanities and large scale data mining. He was won international awards for his research, including the 2010 Digital Preservation Award and both the Vannevar Bush Best Paper award at JCDL2011 and Best Poster Award at JCDL2012. Between 2009 and 2011, he was the UIUC GSLIS Honorary Research Fellow for his interdisciplinary work in digital humanities. Dr Sanderson has been co-PI for national level projects, such as the Open Annotation Collaboration in the US and FORESITE in the UK, and named Investigator on multiple EU funded projects in the FP7 and eContentPlus streams. He is an editor of several international specifications including, most recently, the W3C Open Annotation Community Draft, IETF Memento Internet Draft, and NISO Resource Synchronization. He also has close ties with the very large scale digital library community, including working with the San Diego Supercomputer Center, UC Berkeley, Stanford, Europeana and DPLA, as well as being a founding member of the UK's National Center for Text Mining.

**Christof Schöch** is a researcher at the Chair for Computational Philology, University of Würzburg, Germany. His interests in research and teaching are French Literature (Age of Enlightenment, contemporary novel) and Digital Humanities (quantitative text analysis, digital scholarly publishing). He writes about his research on a blog called "The Dragonfly's Gaze." You can find out more about him on his personal website.

**Grant Leyton Simpson**, Indiana University

**John Simpson** is a postdoctoral fellow at the University of Alberta, splitting time between two projects: INKE and Text Mining & Visualization for Digital Literary History. He teaches, codes, and carries on research in the digital humanities in topics related to visualization, text mining, the semantic web, programming, gaming, and philosophy of science and philosophy of computing.
**Stéfan Sinclair** is an Associate Professor of Digital Humanities at McGill University. His primary area of research is in the design, development, usage and theorization of tools for the digital humanities, especially for text analysis and visualization.

**Yiming Sun** is a Senior Software Architect of the HathiTrust Research Center project and works as a full-time staff at the Data to Insight Center of the Indiana University Pervasive Technology Institute. He has a M.S. degree of Computer Science from the School of Informatics and Computing at Indiana University and is currently a PhD candidate in the same school. His research interest is in the long-term preservation of scientific datasets.

**Rebecca Tegtmeyer** is Assistant Professor of Graphic Design in the Department of Art, Art History, and Design at Michigan State University. Rebecca also serves as Director of Visual Interaction Design Projects at MATRIX: The Center for Digital Humanities and Social Sciences at Michigan State University. Rebecca is an accomplished designer in both the professional and academic spheres. Before coming to MSU, she obtained her Masters in Graphic Design (terminal degree) from NC State University’s College of Design. Prior to graduate school she worked as a designer and art director at Hallmark Cards Inc. At Hallmark she had various roles that primarily involved managing product development and designers in merchandising, marketing, and in-house communications. Currently she serves on the board for AIGA Detroit (the professional association for design) as the Education Director.

**Melissa Terras** is the Co-Director of University College London (UCL) Centre for Digital Humanities, and Reader in Electronic Communication in the Department of Information Studies at UCL. She teaches Internet Technologies, Digital Resources in the Humanities, and Web Publishing, and supervises a range of Masters and PhD students. Her research interests involve applying computational technologies to Humanities problems, to allow research that would otherwise be impossible. As such, she's interested in -- and been involved in -- a variety of research areas that span many areas of Digital Humanities. Current projects include QRator, Digital Transformations, and Textal. Previous projects include Log Analysis of Internet Resources in the Arts and Humanities, Virtual Environments for Research in Archaeology, eScience and Ancient Documents, and Researching eScience Analysis of Census Holdings. Terras is the general editor of *Digital Humanities Quarterly*, and is currently Secretary of the Association of Literary and Linguistic Computing and serves as an executive of the Alliance of Digital Humanities Organizations.

**Aaron Todd** is a Ph.D. candidate in School of Informatics and Computing at Indiana University Bloomington. He is a Research Assistant in Data to Insight Center of the Indiana University Pervasive Technology Institute, under supervision of Professor Plale. His research interests include concurrency, parallelism, compilers, parallel runtime systems, logic programming, type systems and effect systems.

**Kirsten Uszkalo**, University of Alberta

**John Walsh**, Indiana University

**Ethan Watrall** is Assistant Professor in the Department of Anthropology and Associate Director of Matrix: The Center for Digital Humanities and Social Sciences at Michigan State University. In
addition, Ethan is Director of the Cultural Heritage Informatics Initiative and the Cultural Heritage Informatics Fieldschool at Michigan State University. Ethan’s research interests fall in the domain of cultural heritage informatics and digital archaeology, with particular focus on serious games & meaningful play, mobile & geospatial (mostly within the context of public outreach and engagement), and linked open archaeological data. Ethan is co-editor of Archaeology 2.0: New Tools for Communication and Collaboration, an open access volume published by the UCLA Cotsen Institute of Archaeology Press. In addition to his scholarly work, Ethan has written trade books on web and interactive design, including Head First Web Design published by O'Reilly.

Jiaan Zeng is a Ph.D. candidate in School of Informatics and Computing at Indiana University Bloomington. He is a Research Assistant in Data to Insight Center of the Indiana University Pervasive Technology Institute, under supervision of Professor Plale. His research focuses on large scale distributed system and data management, specially looking at how to run data analysis on large data set.