One Culture

Computationally Intensive Research in the Humanities and Social Sciences

A Report on the Experiences of First Respondents to the Digging into Data Challenge

by Christa Williford and Charles Henry
Research Design by Amy Friedlander

June 2012

Council on Library and Information Resources

One Culture

Computationally Intensive Research in the Humanities and Social Sciences

A Report on the Experiences of First Respondents to the Digging into Data Challenge

by Christa Williford and Charles Henry
Research Design by Amy Friedlander

June 2012

CLIR wishes to acknowledge the support and assistance of the four funders of the Digging into Data Challenge

NATIONAL ENDOWMENT FOR THE Humanities

JISC

NSF

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada

as well as the lead program officers for the 2009 grant program:
Alastair Dunning (JISC), Mika Oehling (SSHRC),
Jennifer Serventi (NEH), and Elizabeth Tran (NSF)

Council on Library and Information Resources
Washington, D.C.

ISBN 978-1-932326-40-6
CLIR Publication No. 151
Published by:

Council on Library and Information Resources
1707 L Street NW, Suite 650
Washington, DC 20036
Web site at http://www.clir.org

Additional copies are available for $20 each. Orders must be placed through CLIR's Web site.
This publication is also available online at http://www.clir.org/pubs/reports/pub151.

The paper in this publication meets the minimum requirements of the American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials ANSI Z39.48-1984.

Cover photo: © iStockphoto.com/kr7ysztof

Contents

Tables and Figures

About the Authors

Christa Williford is a program officer at the Council on Library and Information Resources (CLIR). She has co-coordinated the Cataloging Hidden Special Collections and Archives grant program since 2008, in addition to working on programs related to the future of research and the professions of scholarship. She held a CLIR Postdoctoral Fellowship in Academic Libraries from 2004 to 2006 at Bryn Mawr College and a fellowship in Theatre and I.T. Modelling at the University of Warwick in the United Kingdom from 1999 to 2004. Williford holds an M.L.I.S. from the University of Washington Information School and a Ph.D. in Theatre History, Dramatic Literature, and Criticism from Indiana University.

Charles Henry is president of CLIR. Before coming to CLIR, he was provost and university librarian at Rice University, where he was responsible for library services and programs, including the Digital Library Initiative and the Digital Media Center. He served as publisher of Rice University Press, the nation's first all-digital university press, and was a member of the American Council of Learned Societies (ACLS) Commission on Cyberinfrastructure in the Humanities and Social Sciences. He currently serves on the advisory board of Stanford University Libraries, and is also a board member of the National Institute for Technology in Liberal Education (NITLE) and of the Center for Research Libraries. He is a member of the Scientific Board of the Open Access Publishing in the European Network (OAPEN) project. In collaboration with NITLE, he is currently publisher of Anvil Academic Publishing, which focuses on new forms of scholarly research and expression.

Amy Friedlander is on a temporary federal appointment at the National Science Foundation (NSF) where she has worked with the assistant director who heads the Directorate for Social, Behavioral, and Economic Sciences (SBE) to coordinate a strategic visioning exercise to articulate the driving questions in the SBE sciences for the year 2020 and beyond. This resulted in the report, *Rebuilding the Mosaic: Fostering Research in the Social, Behavioral, and Economic Sciences at the National Science Foundation in the Next Decade* (2011). Ms. Friedlander is currently on a detail to the Education and Human Resources directorate at NSF, where she works on strategic communications. In May 2012, she stepped down as editor-in-chief of the *ACM Journal on Computing and Cultural Heritage*. Before joining NSF in June 2010, she was director of programs at CLIR where, among other projects, she participated in the two-year Blue Ribbon Task Force on Sustainable Digital Preservation and Access and established the external evaluation of the National Endowment for the Humanities–led joint agency Digging into Data Program.

Acknowledgments

The Council on Library and Information Resources is grateful to the National Endowment for the Humanities for its sponsorship of the data collection, travel, and publication costs of this project through Cooperative Agreement HC-50007-10. Special thanks are due to Brett Bobley and Jennifer Serventi at the Office of Digital Humanities for their guidance and support throughout the research process, including reading multiple drafts of this document. Their clear visions for the Digging into Data Challenge program generally and for this assessment specifically were invaluable to shaping this report's arguments. Their colleagues at the Joint Information Systems Committee in the United Kingdom, the Social Science and Humanities Research Council in Canada, and the National Science Foundation were equally encouraging and enthusiastic throughout; we are especially thankful for their help in understanding the complexity of the international landscape of e-research. Above all, the many researchers whose groundbreaking work this report describes gave generously of their time over a two-year period that included telephone interviews, site visits, and focus groups. These individuals are listed at the end of this report. Many of the most insightful and eloquent thoughts expressed here are theirs; any mistakes are the responsibility of the authors. Kathlin Smith and Brian Leney at CLIR edited and helped prepare both the abridged print and lengthier web version of this report.

Executive Summary

How many lifetimes? This question often arose when the authors of this report pondered the extraordinary scale and complexity of research conducted in the Digging into Data Challenge program. Analyzing and extrapolating patterns of meaning from tens of thousands of audio files; nearly 200,000 trial transcripts; millions of spoken words, recorded over many years; and hundreds of thousands of primary and secondary texts in ancient languages would, if undertaken using printed resources and analog materials, have required the lifetimes and generations of scholars. Because the resources in question were digital, the time of analysis and discovery was compressed into months, not decades. By choosing to work with very large quantities of digital data and to use the assistance of machines, the Digging into Data Challenge investigators have demarcated a new era—one with the promise of revelatory explorations of our cultural heritage that will lead us to new insights and knowledge, and to a more nuanced and expansive understanding of the human condition.

As articulated in section one, the Digging into Data projects are built on collaborations that are neither contrived nor strained. These collaborations include humanists, social scientists, computer scientists, and other specialists working together toward shared goals that also meet their individual research aspirations. Rather than working in silos bounded by disciplinary methods, participants in this project have created a single culture of e-research that encompasses what have been called the e-sciences as well as the digital humanities: not a choice between the scientific and humanistic visions of the world, but a coherent amalgam of people and organizations embracing both.

Within this one culture are many important differences and distinctions (think of a magnifying lens adjusting to expose increasing levels of granularity). Regardless of their disciplinary significance, at the lowest level all data in a digital environment are zeros and ones, a flattening of information that, while necessary for its storage within a computer's architecture, is not particularly meaningful to humans. At an intermediate level, the human user can appreciate the diversity of digital resources. Data for the humanities and social sciences comprise many media and formats; among the types examined by the Digging into Data investigators are digital images of American quilts, fifteenth-century manuscripts, and seventeenth-century maps; conversations recorded in kitchens; news broadcasts; court transcripts; digitized music; and thousands upon thousands of digitized texts in many languages. Text, speech, music, image, and linguistic data offer rich opportunities for close, careful examination as well as rapid, large-scale computational analysis.

Research at these scales, speeds, and levels of complexity encourages new methodological approaches and intellectual strategies. As recently as 20 years ago, social science researchers typically used analog resources and some computational analysis of data collected in a laboratory or in the field, while humanists worked predominantly with library and archival materials. The Digging into Data Challenge presents us with a new paradigm: a digital ecology of data, algorithms, metadata, analytical and visualization tools, and new forms of scholarly expression that result from this research. The implications of these projects and their digital milieu for the economics and management of higher education, as well as for the practices of research, teaching, and learning, are profound, not only for researchers engaged in computationally intensive work but also for college and university administrations, scholarly societies, funding agencies, research libraries, academic publishers, and students.

Recommendations

This report results from a study of eight international projects that have uncovered previously unimagined correlations between social and historical phenomena through computational analysis of large, complex data sets. The following recommendations are based on this study; they are urgent, pointed, and even disruptive. To address them, we must recognize the impediments of tradition that hinder the contemporary university's ability to adapt to, support, or sustain this emerging research over time. Traditional organizations and funding patterns reflect a much more strictly delineated intellectual landscape. It is time to question which among these boundaries remain useful, which should be more porous, and which no longer serve a useful purpose.

1. Expand our concept of research

To realize the benefits of data-intensive social sciences and humanities, institutions and scholarly societies must expand their notions of what kinds of activities constitute research and reconsider how these activities are supported, assessed, and rewarded. Computationally intensive research projects rely upon four diverse kinds of expertise, each described in detail in section two of this report: **domain (or subject) expertise**, **analytical expertise**, **data expertise**, and **project management expertise**. The active engagement of each of these kinds of experts in the research enterprise is essential. A re-evaluation of hiring practices, job requirements, and tenets of promotion is requisite.

2. Expand our concept of research data and accept the challenges that digital research data present

The digital raw materials upon which today's humanists and social scientists rely are every bit as heterogeneous, complex, and massive as "big data" in the sciences.[1] Not only do humanists and social scientists work with big data, their research also *produces* large data corpora. In fact, some scholars engaged in computationally intensive research see the new data they create as their most significant research outcome. The academy risks losing valuable data unless someone takes steps to care for them in an intelligent manner; to test them with an appropriate degree of skepticism; and, where needed, to correct, enhance, and integrate them with other data in ways that make them meaningful, reliable, and useful to others.

3. Embrace interdisciplinarity

The scholars participating in the first eight Digging into Data projects are active members of multiple academic communities that cross traditionally bounded fields. Their need to work across disciplines mirrors a larger need for organizational flexibility and possible restructuring of institutions of higher learning to promote successful working partnerships between differently trained scholars and academic professionals. Interdisciplinary collaboration benefits not only researchers but also students. Today's colleges and universities must equip students with skills appropriate for a rapidly changing and diverse workforce: the intellectual flexibility that an interdisciplinary perspective cultivates is an excellent foundation for developing these skills.

4. Take a more inclusive approach to collaboration

As the subjects of this report attest, humanists and social scientists engaged in computationally intensive work benefit intellectually and professionally from sustained collaborations with others outside their academic departments and institutions. Library, information technology (IT), and other academic staff; graduate and postdoctoral fellows; undergraduates; and even citizen scholars have roles to play in such research projects. These roles need to be articulated and supported. Section three of this report explores this challenge and other challenges arising from collaborative, multidisciplinary research.

5. Address major gaps in training

The complexity of digital research requires an ongoing commitment to professional development in order to maintain expertise in rapidly accruing resources and tools. Faculty, staff, and students need strong, reliable training programs that correlate sound methodological strategies with appropriate new technologies.

6. Adopt models for sharing credit among collaborators

Institutions of higher learning can more forcefully encourage engagement across disciplinary, institutional, and professional divides by noting and appropriately rewarding their faculty, staff, and students for making substantial contributions to collaborative efforts.

[1] John Coleman, Mark Liberman, Greg Kochanski, and colleagues make compelling comparisons between the sizes of major data corpora in the sciences and humanities on page 3 of their white paper *Mining Years and Years of Speech*. See http://www.phon.ox.ac.uk/files/pdfs/MiningaYearofSpeechWhitePaper.pdf.

Few large-scale digital projects can succeed if individual researchers remain solely responsible for them. If collaborative credit sharing enhances, rather than detracts from, the assessment of an individual's work, more scholars will be willing to work collaboratively, and, ultimately, both the quality and the long-term impact of digital projects in the humanities and social sciences will grow.

7. Adopt models for sharing resources among institutions
The level of investment required to support computationally intensive research is large and growing. It makes no sense to replicate resources, skills, and services at all colleges and universities. Instead, institutions have an opportunity to establish explicit, long-term agreements to work with one another for mutual benefit. There will be serious challenges to overcome—including maintaining appropriate controls over network security, data privacy, and intellectual property—but these challenges must be met to sustain digital research efficiently and affordably.

8. Re-envision scholarly publication
Institutions, scholarly societies, libraries, and funding agencies are all positioned to expand the range of available publication outlets for scholars. Many meaningful outcomes of computationally intensive research, such as data-rich visualizations, cannot be distilled into conference presentations, journal articles, or monographs. Taking advantage of current web technologies, leaders in the academic sector can create new models for publication that incorporate rigorous review processes while at the same time inviting diverse data-rich and multimedia contributions to the academic record.

9. Make greater, sustained institutional investments in human infrastructure and cyberinfrastructure
Computationally intensive research demands a sustainable, redundant network for the preservation of information, as well as trained research professionals to manage this network intelligently. The network's infrastructure should facilitate sophisticated knowledge management and extraction for both anticipated and unanticipated future research. Gateways into that infrastructure will need continual refinement. With investments in innovation and the refinement of user tools, researchers will be able to engage a broader public in their work. Maintaining a digital infrastructure in which collaborative research can flourish will require major commitments from individuals, institutions, governments, and other funders of higher education. It is time for each of these stakeholders to make these commitments.

Recommendations by Stakeholder Group

For researchers:

- Look for opportunities to develop expertise in areas beyond a single discipline, including other related disciplines, data management, data analysis, and project management.
- Create opportunities for students to develop these kinds of expertise.
- Be willing to collaborate both within and outside your discipline, particularly in cases where researchers in other disciplines use similar methodologies.
- Be willing to collaborate both within and outside your institution.
- Be willing to share credit for collaborative work and to recognize others' collaborative efforts.
- Cite digital resources, including tools and data, that you use just as consistently as you cite published articles, conference papers, or monographs.
- Contribute to new forms of digital publication, as authors, editors, and as peer reviewers.

For administrators:

- Commit to investing in the long-term management and preservation of data.
- Create opportunities for humanities and social science faculty, adjunct faculty, staff, and students to develop skills in the management, analysis, and interpretation of these data.
- Offer incentives for engagement in collaborative research initiatives.
- Develop models for the assessment of collaborative work.
- Develop partnerships with institutions with complementary strengths.
- Adopt clear policies for sharing hardware, software, and data resources among on- and off-campus researchers that maximize openness yet protect privacy and intellectual property.

For scholarly societies:

- Cultivate and critically assess new research methodologies with potential benefits for your discipline.
- Promote the value of computationally intensive research methodologies within your discipline to researchers outside the discipline and to the wider public.
- Create opportunities for members to develop skills in data management and analysis.
- Encourage cross-disciplinary engagement among members as well as non-members with relevant expertise.
- Build alliances with other societies with similar needs and interests.
- Support new models for scholarly communication and peer review.
- Commit to supporting the long-term preservation of key digital resources in your discipline.

For academic publishers:

- Seek to publish content that crosses disciplinary boundaries and embraces newer computationally intensive methodologies.
- Encourage the submission of work by multiple authors and ensure that publications give credit to all contributors to this work.
- Seek ways to incorporate digital data and multimedia into online publications, and adopt models for assessing such work.
- Commit to supporting the long-term preservation of and access to your publications.
- Deepen partnerships with academic institutions and scholarly societies in the service of preservation and access.
- Where possible, increase transparency in your business practice so that other academic stakeholders understand the true costs of publication and how these costs are changing over time.

For research libraries:

- Recruit and develop staff prepared to engage as active partners in computationally intensive research initiatives, particularly by offering expertise in data management, data analysis, or the management of collaborative projects.
- Recruit and develop staff capable of contributing to the peer review of new forms of online scholarship.
- Offer consultation services to researchers that help them manage, maintain, and, if warranted, transfer responsibility for valuable research data to library repositories.
- Offer consultation services to researchers that help them identify appropriate publication venues for non-traditional forms of scholarship.
- Encourage cross-disciplinary engagement among researchers and students at your library, such as through public programs or workshops related to data-intensive research tools.
- Establish partnerships with other institutions to promote the long-term preservation of and access to scholarly publications and the digital data upon which they rely.

For funding agencies:

- Acknowledge the high costs of curating reliable large-scale digital data sets for the humanities and social sciences and create incentives for researchers, institutions, and scholarly societies to accept responsibility for these costs.
- Support robust, thoughtful approaches to computationally intensive research in the humanities and social sciences that incorporate disciplinary rigor as well as sound data management, analytical, and project management practices.
- Support training and professional development opportunities related to computationally intensive research for students, staff, and faculty.
- Support new models for academic publication and peer review.
- Encourage cross-disciplinary and multi-institutional research initiatives that take advantage of academic professionals' and institutions' complementary strengths.

1. The First Challenge and Its Respondents

1.1 One Culture

> *I have, of course, intimate friends among both scientists and writers. It was simply through living among these groups, and much more, I think, through moving regularly from one to the other and back again that I got occupied with the problem of what ... I christened to myself as the "two cultures."*

C. P. Snow thus differentiated two distinct intellectual communities—what we would call today humanists and scientists—that had lost the ability to communicate across their disciplinary boundaries and, for all the similarities in intellect, background, and social standing, lived and worked in worlds that could not be bridged. Encounters between these two societies were often hostile and dismissive.

Interestingly, one of the topics that Snow chose to highlight in his description of the divergent worldviews of the "two cultures" was the Industrial Revolution. Snow claimed the revolution was largely a product of science and engineering, and also claimed that writers and humanists had largely ignored it. The crux of this observation is an assumption that science embraces technology and the humanities does not, or does so more slowly and reluctantly.[2]

Such an assumption is no longer valid, as this report shows. The Digging into Data projects are successful collaborations, and through their success give evidence of shared intellectual values, rigorous methodological approaches, and common ground across scientific and humanistic disciplines. Researchers from these disciplines rely deeply upon one another for insight and discovery when confronted with very large-scale, complex challenges.

Nevertheless, these researchers still work in environments that, at least implicitly, admit the residual truth to Snow's argument. Their academic departments, scholarly and professional societies, colleges, and universities are set apart, clustered, and structured according to the traditional "two-culture" perspective Snow describes. The grant programs and funding agencies that have traditionally supported their work are similarly focused. The Digging into Data program has challenged this bifurcation by insisting on collaboration across disciplines and by funding the projects through an amalgam of sources that cross these borders.

While it is not the intent of this report to dredge again the merits and failings of Snow's now famous declamations, the projects it describes suggest a very different academic landscape supporting *one* culture in the pursuit of knowledge. The eight teams of researchers have built collaborations that are neither contrived nor strained. In assessing the project teams' work, we have come to understand that the one culture of e-research—encompassing what have been called the e-sciences as well as the digital humanities—involves not

[2] Snow, C. P. *The Two Cultures* (Cambridge: Cambridge University Press, 1998), p. 2. Based upon a talk given by Snow at Cambridge University on May 7, 1959, first published in the same year by Cambridge University Press. Cited in Patricia Waugh, Review of *The Two Cultures Controversy: Science, Literature and Cultural Politics in Postwar Britain*, by Guy Ortolano (Cambridge: Cambridge University Press, 2009).

a choice between the scientific and humanistic visions of the world, but an imperative that people and organizations fully embrace both. In these projects, highly organized teamwork, such as might characterize a scientific laboratory, is as significant as more free-form contemplation. It is in working together *and* apart[3] that we will see digital scholarship flourish.

1.2 Background: "What Do You Do with a Million Books?"

In 2006, Tufts University professor Gregory Crane posed the "million book" question in an article[4] exploring the potential for doing large-scale investigations of text corpora. Crane identified several problems facing computationally intensive research with texts, including insufficient funding for digital text repositories, variable quality and granularity of repository content, inaccuracies arising from errors made in optical character recognition (OCR) and metadata generation, research plans that are too narrowly defined to appeal to a broad audience, and access restrictions imposed by pay walls and copyright laws.

The one culture of e-research involves not a choice between the scientific and humanistic visions of the world, but an imperative that people and organizations fully embrace both.

The launch of the Digging into Data Challenge in 2009 was in part a response to the "million book" prospectus. Very large data sets are susceptible to scholarly inquiry but are dependent on computational tools and equipment for execution and analysis. What are the intellectual benefits, and what are the risks? How does this new research align within the traditional context of scholarship and how might it be distinct? Mass-digitization projects such as Google Books, which had by then prompted widespread excitement and speculation about its use for research,[5] and HathiTrust, a not-for-profit library-based alternative, made the challenge more intuitively feasible. "Reading" large text corpora by machine—encompassing an amount of information exponentially greater than would be possible for any individual to take in and process in a lifetime—was then, as now, a subject at once intriguing, daunting, and unsettling.

Under the leadership of Brett Bobley, chief information officer and director of the Office of Digital Humanities, National Endowment for the Humanities (NEH-ODH), the Digging into Data Challenge was framed broadly, to encompass *any* type of digital or digitized content used by researchers in the social sciences and humanities. In discussions before the program's launch, leading researchers and other funders had stressed that establishing reliable methodologies for analyzing large quantities of non-text digital

[3] A 2009 CLIR report titled *Working Together or Apart: Promoting the Next Generation of Digital Scholarship*, which was the outcome of a symposium planned by former CLIR Director of Programs Amy Friedlander, provided the foundation for the study upon which this report is based. The insights of its authors, who write from specific disciplinary perspectives, resonate well with the findings here.

[4] Crane, Gregory. "What Do You Do With a Million Books?" *D-Lib Magazine* 12.3 (March 2006). Available at http://www.dlib.org/dlib/march06/crane/03crane.html.

[5] This speculation was in addition to the class action lawsuit filed by the Author's Guild and the Association of American Publishers, still in litigation after a proposed settlement agreement was rejected by the New York Southern District Court in March 2011.

content—including audio, image, and audiovisual data, was as important as learning to machine-read large bodies of texts. These advisers proposed that the Challenge be supported by a group of funders rather than adopted as the responsibility of a single agency. The United States National Science Foundation (NSF), the Joint Information Systems Committee in the United Kingdom (JISC), and the Canadian Social Sciences and Humanities Research Council (SSHRC) joined the NEH in preparations to coordinate grant calendars, guidelines, and a review process for the new program. By requiring international collaboration, the four agencies hoped to fund projects that would have high visibility and broad appeal; by actively recruiting the managers of significant data repositories to signal support for the program through making their holdings accessible, the agencies hoped to encourage openness. Eight proposals were funded for the first round; they are the focus of this report.

Table 1. Digging into Data Chronology

November 2007	The National Endowment for the Humanities Office of Digital Humanities (NEH-ODH) begins exploring the idea of a new funding program focused upon computationally intensive humanities research
May 8, 2008	NEH convenes the "Million Book Challenge" planning meeting with scholars and other funding agencies
January 16, 2009	Four cooperating funders announce first Digging into Data Challenge[6]
September 10–11, 2009	Joint review panels determine first award recipients
December 3, 2009	First awards announced[7]
March 16, 2011	Eight cooperating funders announce second Digging into Data Challenge[8]
June 9–10, 2011	Digging into Data Challenge conference held[9]
December, 2011	Second round of awards announced[10]

1.3 The Context of this Study

At its inception, this study posed two fundamental questions to the eight research teams:

1. Why do you as a scholar need a computer to do your work?; and
2. What *kinds* of new research can be done when computer algorithms are applied to large data corpora?

The questions imply a distinction between "new" computer-based and "traditional" non-computer-based research in the humanities and social sciences. Early on, that distinction became problematic. While natural and perhaps necessary to pose the old and new in opposition to one another to better understand the changing landscape of scholarship and the transformative potential of new

[6] http://www.neh.gov/news/press-release/2009-01-16-0

[7] http://www.neh.gov/news/press-release/2009-12-04

[8] http://www.neh.gov/news/press-release/2011-03-16

[9] See http://www.diggingintodata.org/

[10] http://www.neh.gov/news/press-release/2012-01-03

technologies, there was never clear separation between past and present, traditional and digital, or other bounded concepts that very quickly felt artificial and unhelpful. Many of the researchers interviewed for this study assiduously avoided making such distinctions.

The framing questions thus quickly and unintentionally exposed an important aspect of collaborative, computationally intensive research initiatives. The eight projects that are the subject of this report reflect more complex, iterative interactions between human- and machine-mediated methods than are implied by our second question. Rather than being a combination of fixed, clearly defined entities—the researcher's question, the algorithm, and the corpus—the projects are structures built with continually moving parts. Certain research questions require major investments of human labor in amending corpora; others require intense testing and reworking of algorithms to adapt to new and varied data. It is the *combination* of algorithmic analysis and human curation of data that helps humanists

Table 2. Digging into Data Projects

Project	Disciplines	People	Data Type/Size	Method of analysis	Tools
Using Zotero and TAPOR on the Old Bailey Proceedings: Data Mining with Criminal Intent (DMCI)	History; Philosophy; Humanities Computing; Communication Studies and Multimedia	19	Text from Proceedings from the Old Bailey Online,[11] including 197,000 trial records with full transcripts from 1674–1913	Text mining and multiple types of visualization	Zotero,[12] Voyeur/Voyant Tools,[13] Mathematica,[14] custom application programming interface (API) for Proceedings from the Old Bailey Online[15]
Digging into the Enlightenment: Mapping the Republic of Letters	French; English Literature; History; Computer Science; Academic Technology	23	Transcriptions and metadata from more than 50,000 letters within the Electronic Enlightenment[16]	Geographic analysis; text mining; visual analytics	Improvise,[17] Electronic Enlightenment Correspondence Visualization Tool[18]
Towards Dynamic Variorum Editions (DVE)	Classics; Computer Science; Computational Linguistics; Artificial Intelligence; Library Science	11	1.2 million-volume collection of digitized texts from the Internet Archive, Google, and the Perseus Digital Library[19]; tests of Greek OCR conducted using 158 19th-century Greek texts[20]	OCR; morphological analysis	Gamera document analysis framework[21]
Mining a Year of Speech	Phonetics; Linguistics, Philology; Computing and Information Science	10	9000 hours (~100 million words) of recorded American and British speech, including the British National Corpus[22] and holdings of Penn Linguistic Data Consortium	Forced alignment of transcription to audio	Penn Phonetics Lab Forced Aligner[23]

[11] http://www.oldbaileyonline.org/

[12] http://www.zotero.org/

[13] http://hermeneuti.ca/voyeur

[14] http://www.wolfram.com/mathematica/

[15] http://www.oldbaileyonline.org/static/API.jsp

[16] http://www.e-enlightenment.com/

[17] http://www.cs.ou.edu/~weaver/improvise/index.html

[18] http://www.stanford.edu/group/toolingup/rplviz/

[19] http://www.perseus.tufts.edu/

[20] See Robertson, Bruce. "Optical Character Recognition of 19th-Century Polytonic Greek Texts: Results of a Preliminary Survey." Perseus Digital Library (Jan. 19, 2012). Available at http://www.perseus.tufts.edu/publications/dve/RobertsonGreekOCR/

[21] http://gamera.informatik.hsnr.de/

[22] http://www.natcorp.ox.ac.uk/

[23] http://www.ling.upenn.edu/phonetics/p2fa/

and social scientists refine their existing questions and articulate new ones. Furthermore, many of these projects show collaborators making significant advances in the field of computer science as well as within the relevant subject domain. Conducting research "at scale," especially across the unstructured and heterogeneous data upon which humanists depend, can inspire new and more nuanced applications of computer tools, which can in turn lead to new questions.

1.4 The Eight Projects

The web-based version of this report is available at http://www.clir.org/pubs/reports/pub151.

The web-based version of this report includes individual case studies that describe key findings as well as some of the challenges each project team encountered. This printed report describes the cases in aggregate, extrapolating the commonly shared, characteristics. Table 2 notes the represented disciplines, numbers of researchers, data types, methodologies, and tools used for each project.

Table 2, continued

Project	Disciplines	People	Data Type/Size	Method of analysis	Tools
Harvesting Speech Datasets from the Web	Speech and Language Processing; Linguistics; Computing and Information Science	3	Project involves harvesting short snippets of audio from numerous very large audio corpora across the web	Forced alignment of transcription to audio; acoustic extraction; machine learning classification	ProsodyLab Aligner;[24] HTK Speech Recognition Toolkit[25]
Structural Analysis of Large Amounts of Music Information (SALAMI)	Computational Musicology; Music Technology; Library and Information Science	16	1,383 individual musical pieces analyzed both by students and computer	Computer-aided analysis of musical structures	The Music Information Retrieval Evaluation eXchange (MIREX);[26] custom interactive music visualizer
Digging into Image Data to Answer Authorship Related Questions (DID-ARQ)	French; Cultural History; History of Cartography; Computer Science; Assessment; Museum Science; History; Art; Art History; Environmental Literature	34	6,000 high-resolution page images of 15th-century French manuscripts; images of 40 maps of the Great Lakes region, 1650–1800; more than 56,000 low-resolution images of 19th- and 20th-century quilts	Adaptive image analysis; machine learning	Image 2 Learn Toolset;[27] Virtual Vellum;[28] Medici image repository; custom code repository[29]
Railroads and the Making of Modern America	History; Geography; Computer Science and Engineering	14	unstructured text from books, newspapers, and railroad-related periodicals and ephemera, alphanumeric census and non-census data sets, GIS data sets, maps	Hand-corrected data and metadata on discrete topics made available for geospatial and temporal exploration in web-based "apps"	The Aurora Engine,[30] a set of customized data exploration tools specifically produced for this project

[24] http://prosodylab.org/tools/aligner/

[25] http:// htk.eng.cam.ac.uk

[26] http://www.music-ir.org/mirex/wiki/MIREX_HOME

[27] http://isda.ncsa.uiuc.edu/Im2Learn/

[28] http://www.shef.ac.uk/hri/projects/projectpages/virtualvellum

[29] http://did.ncsa.illinois.edu/svn/did/trunk/

[30] http://auroraproject.unl.edu/

2. Characteristics of the Eight Projects

> *It became clear in our work that humanists, who are often exceptional experts in their fields, often have a difficult time describing how they go about their work and analyses. Having humanists work in teams and with computer scientists required them to explain and detail their processes. The work we have done has made us more sensitive to this issue and opens up many new areas of research—How can we develop better collaborative models that help humanists explicate their processes? Can we build tools to help capture the way humanists work? How can we enhance digital archives to facilitate the ways humanists work with objects?*
>
> —Dean Rehberger, *Digging into Image Data to Answer Authorship-Related Questions*

2.1 Structural Commonalities and Notable Differences

A broad range of topics and methodologies are represented in the eight inaugural Digging into Data initiatives. Nevertheless, when considered as exercises in research *practice*, the initiatives reveal some shared characteristics. All projects:

1. Engage with data corpora that are much larger than what might be read, seen, heard, or experienced by a single individual. These data range from highly structured, uniform, and topically specific to completely unstructured and heterogeneous corpora.
2. Apply some form of computational analysis—whether described as a tool, an application, or merely an algorithm—to these corpora. These tools, applications, and algorithms vary from the highly specific to the more general, from the most experimental to the mature. Some are widely accessible, and others require the expertise of computer specialists.
3. Require continual refinements to tools and data, which in turn requires collaboration and coordination of multiple project participants with varied backgrounds and skills.
4. Conducted a research process that incorporates most or all of seven stages:
   a. hypothesis and/or question formation;
   b. selection of a corpus or corpora;
   c. exploration of a corpus or corpora;
   d. querying and correcting, modifying, or amending the data as needed;
   e. pulling together subsets of data relevant to a given question;
   f. making observations about those data; and
   g. drawing conclusions from and/or interpreting those data.

The case studies show that computationally intensive research mirrors other kinds of inquiry, although they suggest that a dependency on digital tools and resources requires more explicit documentation and communication about methodology than has typically been the case in the humanities and qualitative social sciences. Because digital research methodologies are still maturing, it is important to consider carefully the rationale for the investigators' choices of analytical tools and the evidence produced—that is, to reflect upon the significance of a specific tool applied to a specific corpus. To borrow the words of one project team, who worked together on *Using Zotero and TAPOR on the Old Bailey Proceedings: Data Mining with Criminal Intent (DMCI)*,[31] "The methodology is part of the message."

Social scientists are generally comfortable foregrounding explanations of methodology in discussions of their research; humanists, by contrast, tend to foreground the argument or interpretation resulting from scholarly investigation rather than the research methods. Asserting the value of one's approach to research as a model for others is a more comfortable position for social scientists than for humanists. Humanists often see greater understanding of the subject matter with which they are concerned as their primary contribution to their fields, or at least a more important contribution than the preparatory work necessary to describe new findings and support new claims.

2.2 Interdisciplinarity

Crossing disciplinary boundaries often increases the impact of computationally intensive scholarship by exposing it to greater numbers of researchers, students, and the public. At the same time, it complicates project management: traditions, concepts, and research vocabularies must be adapted to accommodate other points of view. When the common ground for a collaboration is methodological (the "how") rather than driven by a shared desire for a particular discovery or outcome (the "why"), partners must be prepared to work in ways that do not neatly fit the models they have been trained to emulate. This results in products for which partners cannot take sole credit, some of which defy traditional kinds of peer review. The level of stress this transformation may create for the researcher varies by discipline, by institution, and by individual, but acceptance of this change is obligatory.

These projects point to new avenues for investigation more often than they provide conclusive answers to their original framing questions. This is not surprising, given that topics as complex as patterns of human creativity, authorship, and the continuity of culture over time often elude conclusive explanation. But many practitioners of computer-assisted investigation contend that in time, with enough attention to the curation of valid data, the formation of suitably complex and replicable methods of analysis, and the framing of increasingly precise questions, it may be possible to combine computer-based analysis of large data corpora with the creativity and critical power of the human researcher to promote a greater understanding of our society and culture than has ever been possible. The prospects of new discovery at such a scale seem achievable only through continued collaboration across disciplines.

[31] http://criminalintent.org/.

2.3 The Spectrum of Data and Its Consequences

The quality, quantity, and utility of data is unquestionably the most complex determining aspect of these projects. Within an umbra of many shared characteristics, important differences surfaced as an effect not only of different disciplinary traditions but also of the choice of collaborators deemed most suitable for the media, scale, and organization of the targeted data sets, the proportion of manual to automated work, the need for continual adaptation of analytical tools, and the likelihood of achieving major outcomes in a brief (15-month) grant period. In other words, it is not just the specificity of the question or the maturity of a tool that determines what computationally intensive research might achieve, but also the state of the raw material from which it is produced.

In discussions and subsequent exchanges at the Digging into Data program's culminating meeting in June 2011, University of Portsmouth Professor Richard Healey, co-principal investigator of *Railroads and the Making of Modern America*, suggested that the framing of the original Challenge oversimplified the kinds of work that computationally intensive research encompasses. He writes, "I think there may have been something of an implicit original assumption behind the initiative, at a broad level, that since there were multiple millions of digital text/image/data files 'out there' …all the focus would be on the use of data mining and other algorithms to tease out new signals from the noise." Rather than a "one-size-fits-all" model for data-intensive humanities and social sciences, Healey proposes many different levels of data-related operations. He describes these levels as a "data hierarchy" and characterizes them as follows:

> Level 0: Data so riddled with error that they should come with a serious intellectual health warning ... ! (We have much more of this than most people seem willing to admit. ...).
>
> Level 1: Raw data sets ... corrected for obvious errors.
>
> Level 2: Value-added data sets (i.e., those that have been standardised/coded, etc., in a consistent fashion according to some recognised scheme or procedure, which may require significant domain expertise/ training and the exercise of judgement. …).
>
> Level 3: Integrated data resources ... these will contain value-added data sets but the important additional aspect of these resources is that explicit linkages have been made between multiple related data sets (or have been coded/tagged in such a way that the linkages can be made by software. ...).
>
> Level 4: "Digging Enabler" or "Digging Key" data/classificatory resources … these require extensive domain expertise and use/ analysis of multiple sources/relevant literature to create. They facilitate extensive additional types of digging activity to be undertaken on substantive projects beyond those of the investigators who created them, i.e., they become "authority files" for the wider research community. Gazetteers, structured occupational coding systems, data cross-classifiers, etc., fit into this category. ...

> There are important questions also about how such resources acquire authority status (e.g., through quality of referencing back to original sources, through collaborative work by leading research groups in the field, by peer review, by crowd sourcing from citizen scholars).[32]

These distinctions make clear that to realize the benefits of data-intensive social sciences and humanities, institutions and scholarly societies must expand their notions of what kinds of activities constitute research, and must reconsider how these different activities are supported, assessed, and rewarded.

2.4 Expertise

Investigators stressed frequently that the research they pursued would not be possible without extensive collaboration with partners who contributed many kinds of expertise working in what Peter Ainsworth (*Digging into Image Data*) called a "transformative, symbiotic partnership." Collaborators' expertise and training overlapped more in some cases (such as *Mining a Year of Speech*, *Data Mining with Criminal Intent*) than in others (such as *Digging into the Enlightenment, Digging into Image Data*). When the teams included experts with complementary, rather than overlapping, strengths, the coordination and management of the project, including communications among the partners and dividing responsibility for shared resources, was especially vital, as were significant investments of time in planning for and framing the project.

Four generic kinds of expertise were represented among partners in each project: domain expertise, data management expertise, analytical expertise, and project management expertise. Participants in all the projects shared an appreciation for each of these kinds of skills. While not always represented in the same proportions, each of these areas was represented in the eight projects by one or more individuals. These categories of expertise seemed important counterbalances to one another, as if they were four supporting legs of a table (Figure 1).

Although the investigators agreed that the four categories were equally important, some observed that the contributions of researchers with more than one of these kinds of expertise were most critical to project success. Dan Edelstein, who worked on *Digging into the Enlightenment*, put it this way: "What made our project possible was that we had these hybrid people with more than one leg of the 'table'. Those people are very hard to find. They don't do well naturally in a university setting." Students, short-term project staff, and junior faculty all played crucial roles, often in a "hybrid" capacity.

[32] E-mail from Richard Healey to Christa Williford, June 11, 2011.

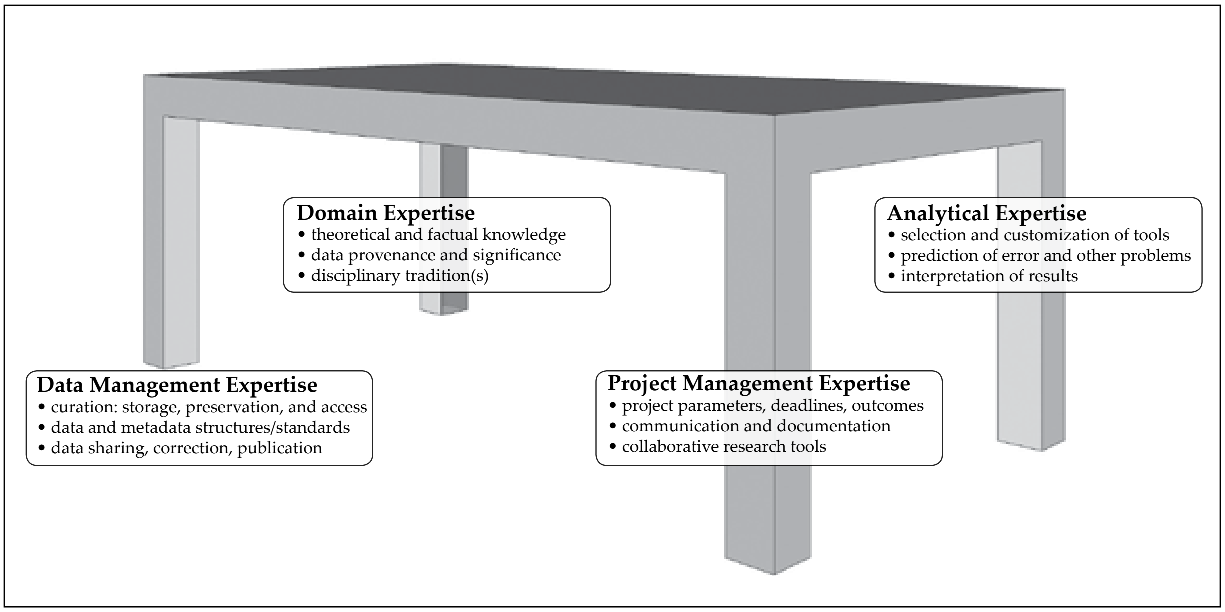
Domain Expertise
• theoretical and factual knowledge
• data provenance and significance
• disciplinary tradition(s)
Analytical Expertise
• selection and customization of tools
• prediction of error and other problems
• interpretation of results
Data Management Expertise
• curation: storage, preservation, and access
• data and metadata structures/standards
• data sharing, correction, publication
Project Management Expertise
• project parameters, deadlines, outcomes
• communication and documentation
• collaborative research tools

Fig. 1: Expertise represented among project partners

2.4.1 Domain Expertise

Domain expertise incorporates theoretical as well as a factual understanding of the humanities or social science research traditions relevant to the project. It was usually represented in the projects at the principal investigator level, an indicator of its critical importance. Beyond this, outside contributors also played important roles; for example, in the *Data Mining with Criminal Intent* project, a number of outside experts tested and evaluated the project's tools and methodology. The theoretical component of *Structural Analysis of Large Amounts of Music Information* is an ontology contributed by experts at the universities of Oxford and Southampton, and was fundamental to shaping the direction of that project. Domain expertise requires familiarity with the kinds of data to be examined, the ways in which disciplinary specialists have interpreted them in the past, and the ability to identify key knowledge gaps and questions toward which computationally intensive methodologies can be applied. Other relevant skills include an understanding of the provenance and materiality of digitized evidence and the imagination to make connections between research concerns and the concerns and practices of related disciplines. Familiarity with the relevant disciplinary literature, its

Domain experts have:

- a deep theoretical and factual knowledge of relevant field(s)
- familiarity with types of data to be examined, their provenance, and their significance to the relevant field(s)
- the ability to identify knowledge gaps
- familiarity with disciplinary literature and conventions
- the ability to teach others from different backgrounds to appreciate all of the above

conventions of citation and publication, and an ability to teach others to appreciate the importance of each are included in this skill set. Critically, these experts must be comfortable teaching others from diverse educational backgrounds, including students at all levels; computer scientists, programmers, and developers; and members of the general public.

2.4.2 Data Expertise

> *As a team we noticed an interesting interaction where we had to accept each other's approaches. This was particularly important in that those in the Old Bailey who had come in with an appreciation for their structured data had to come to understand how the Old Bailey could be seen as a mass of unstructured data for text mining. The text miners in the group in turn had to look more closely at what could be done with structured data. This was a fruitful exchange.*
>
> —*Data Mining with Criminal Intent* team

Data expertise is defined by an understanding of how data have been collected and curated, the relationships between material objects and digital representations of those objects, relevant data models and conventions for description, and storage systems and how they affect the way in which data are accessed and preserved. Understanding of information-seeking behaviors across diverse disciplines and an ability to predict future or alternate uses for data consumed or produced by the project are also relevant. Devising ways to manage the hand-correction of erroneous data efficiently is another important contribution of data experts, since such tasks can consume a major share of labor on a digital project when such correction is necessary.

A data expert must have sufficient technical knowledge of storage systems to help others comprehend how they might affect compatibility or interoperability with other systems and standards. She or he must also understand new forms of publication that can integrate data resources with narrative and interpretation. These experts make important contributions in teaching and advising other participants to adopt research practices that maximize readiness of relevant data for publication and reuse.

Data expertise was often represented in the projects by the managers or curators of the corpus or corpora to be investigated. The British Library partners who manage the British National Corpus that is part of the basis for *Mining a Year of Speech*, the University of Oxford–based creators of the *Electronic Enlightenment* for *Digging into the Enlightenment*, the Tufts University managers of the Perseus Digital Library for *Towards Dynamic Variorum Editions*, and the multiple partners who share responsibility for creating the *Old Bailey Online* (*Data Mining with Criminal Intent*) and the *Quilt Index* (*Digging into Image Data*) are examples. The level of engagement of these partners in the day-to-day operations of each project varied according to how structured or accessible their data were initially and how closely project activities aligned with the priorities of the institution responsible for maintaining the corpus.

Data experts have:

- an understanding of how data have been collected and curated and of relationships between material objects and digital representations of those objects (if applicable)
- familiarity with data models and/or conventions of description
- an understanding of how relevant data are accessed and stored
- the ability to facilitate data sharing and manual error correction, both during and after the project
- the ability to predict future or alternate uses for data
- an understanding of new forms of publication that can incorporate data

2.4.3 Analytical Expertise

> *We realized that we needed to be better about opening up the black boxes of algorithms. For many humanists, they remain a mystery in which one feeds things in one end and an "answer" comes out the other end. But algorithms are more like recipes, and it is important to have humanists be part of every stage of the process. We need to determine the ingredients (features) that will be used in the process. We need to make it clear that the actual "cooking" process of the algorithm can be changed or tweaked depending on the input and output. And finally, the output is not an answer but another kind of "text" or "visualization" that needs to be interpreted or analyzed. Algorithmic literacy means not only learning how to interpret results but to understand the whole "cooking" process of algorithm development.*
>
> —Dean Rehberger, *Digging into Image Data to Answer Authorship-Related Questions*

Technologists, scholar-technologists, information scientists, and computer scientists contributed analytical expertise. While the role of the analytical expert was important in all of the projects, it was paramount in those relying on high-performance computing infrastructures and in those developing cutting-edge methodologies (such as visual analytics for *Digging into the Enlightenment*, computer-aided structural analysis of music for the *SALAMI* project, and adaptive image analysis for *Digging into Image Data*). Analytical expertise is not limited to specialized programming and computation, and for data-intensive work often requires a much broader understanding of research methodologies than is common for programmers or developers. Gregory Crane, one of the investigators who led *Towards Dynamic Variorum Editions*, emphasized the importance of this distinction: "We often do not get access to people working at a sufficient level of expertise [in computational analysis] to get real work done."

Analytical expertise includes understanding the strengths and weaknesses of an array of research tools relevant to a project. These may include generic statistical, visualization, geographic information, and optical character recognition tools as well as the numerous specialized algorithms used by the Challenge investigators. Analytical experts select the most appropriate tools and are able to customize and improve them for specific research tasks. These experts can test the efficacy of an analysis, validate results, and teach less-experienced partners to read and interpret visualizations, charts, and statistics. Measuring the performance of new methods against traditionally collected "ground truth" data in order to validate those methods was a key component of the *SALAMI* project and *Harvesting Speech Datasets for Linguistic Research on the Web*.

Analytical experts have the ability to:

- understand the strengths and weaknesses of individual research tools
- select and customize appropriate tools to support research goals
- predict problems that might arise with using the selected tool to perform project tasks
- predict and detect error rates in data and data analysis algorithms and to choose statistical methods that account for these errors when appropriate
- teach others to interpret results of analysis

2.4.4 Project Management Expertise

> *This project has offered me the unique opportunity, as both a junior faculty member and a female in digital humanities, to simultaneously develop leadership and research skills. The project PIs have given me and other junior faculty the opportunity to integrate ourselves fully into the project not just as a source of labor (be it intellectual or task-based) but also as a participant in shaping the future of our research. They have mentored us throughout the project, created specific pathways to publications and presentations, and allowed us equal ownership of the project.*
>
> —Jennifer Guiliano, *Digging into Image Data to Answer Authorship-Related Questions*

Without project management expertise, none of the inaugural Digging into Data projects could have succeeded. The inherently experimental nature of these projects made coordinating parallel work streams complicated. Project managers had to track the achievements of their collaborating partners on an almost daily basis, especially in cases where large numbers of people were involved. Thorough and consistent project documentation, so necessary for the products of such initiatives to be useful to other scholars, is an additional component that requires a skilled manager's coordination. For the projects funded through the Challenge, there were the additional burdens of reporting about the same project to several funding agencies. The work involved in compiling such reports is significant.

Project management responsibilities were often distributed to several members of the team, most commonly to principal investigators. Occasionally one of the collaborating institutions assumed leadership in this area. For one of the larger initiatives, *Digging into the Enlightenment*, the team chose to assign major project coordination tasks to the in-house academic technology specialist at the Stanford Humanities Center; having access to a professional with expertise in grant management and coordination was invaluable. For *Digging into Image Data*, another large collaboration, the experience and support of staff at the Institute for Computing in Humanities, Arts, and Social Science (I-CHASS) at the University of Illinois' National Center for Supercomputing Applications (NCSA) were fundamental. Here, where sharing hardware, software, and data among distant collaborators was critical to project success, the partners crafted a formal memorandum of understanding for the project that eliminated confusion about participants' individual roles and responsibilities, freeing partners to focus on their work.[33] The agreement also addressed legal and ethical issues, including setting standards for citation and credit sharing among participants in post-project presentations and publications as well as for respecting intellectual property restrictions. In addition, the agreement prescribed methods for communication and documentation for the project and a policy for licensing of any software deliverables.

Project managers have:

- an ability to frame project parameters
- an ability to set appropriate goals and deadlines and to coordinate parallel work streams if necessary
- an ability to select the most appropriate communication and documentation strategies for the project
- a mastery of collaborative research tools
- a strong desire to work toward outcomes that benefit all team members

[33] Simeone, Michael, Jennifer Guiliano, Rob Kooper, and Peter Bajcsy. "Digging into Data Using New Collaborative Infrastructures Supporting Humanities-Based Computer Science Research." *First Monday* 16.5 (May 2, 2011). Available at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3372/2950. Section 3.2 of this article, titled "Legal and Ethical Aspects of Scholarly Collaborations," is especially salient here.

3. Reflecting and Looking Ahead

> *I think that historians need a computer to document even more fully their research process and to allow others into that process. I believe we will move into a process of scholarly production and communication that is cumulative, discipline based, and verifiable in digital form. Beyond that I imagine that readers of history want to have access to materials that historians work with and interpret. The historians should use the computer to open up their work and make it more visible, relevant, and meaningful to the public.*
>
> —William G. Thomas, *Railroads and the Making of Modern America*

3.1 Why Computers? What Kinds of New Research?

We have explored the common characteristics of the eight inaugural Digging into Data projects and hinted at some of their diversity and complexity, but what of their significance? What kinds of new discoveries in the humanities and social sciences have these investigators reported to date?[34] What are the answers to this assessment's initial questions: Why do you as a scholar need a computer to do your work? and What *kinds* of new research can be done when computer algorithms are applied to large data corpora?

One consistent metaphor in this study likens the computer to a moveable and adjustable lens that allows scholars to view their subjects more closely, more distantly, or from a different angle than would be possible without it. Daniel Cohen, a principal investigator for *Data Mining with Criminal Intent*, described two "use cases" for investigating the massive corpus of 197,000 digitized and coded trial transcripts within *The Old Bailey Proceedings Online*.[35] The first, which he called "hunt and peck," involves picking out a few examples of specific phenomena from within a vast data corpus; the second, which he called "slices," is a way to look for trends and anomalies across larger amounts of data. In their white paper, the group cites examples of both types of use. By "hunting and pecking" for cases with references to "poison," historian and developer Fred Gibbs discovered frequent co-occurrences of "poison," "drank," and "coffee," suggesting to him that coffee was the poisoner's medium of choice in eighteenth- and nineteenth-century London. By contrast, Tim Hitchcock and William Turkel have extracted vast "slices" from the Old Bailey corpus to create scatter plots of trial transcript lengths over time, as in Figure 2, which represents Old Bailey trials from the 1860s.

The results are intriguing: "As far as we know, no one has ever observed that the printed trials in this decade and a number of others were either shorter than about one hundred words, or considerably longer, but almost never around 100 words long. We are currently

[34] At time of writing, some of the Digging into Data projects funded in 2009 are still under way and have yet to report final results.

[35] Tim Hitchcock, Robert Shoemaker, Clive Emsley, Sharon Howard, Jamie McLaughlin, et al., *The Old Bailey Proceedings Online, 1674-1913* version 7.0, 24 March 2012. Available at http://www.oldbaileyonline.org.

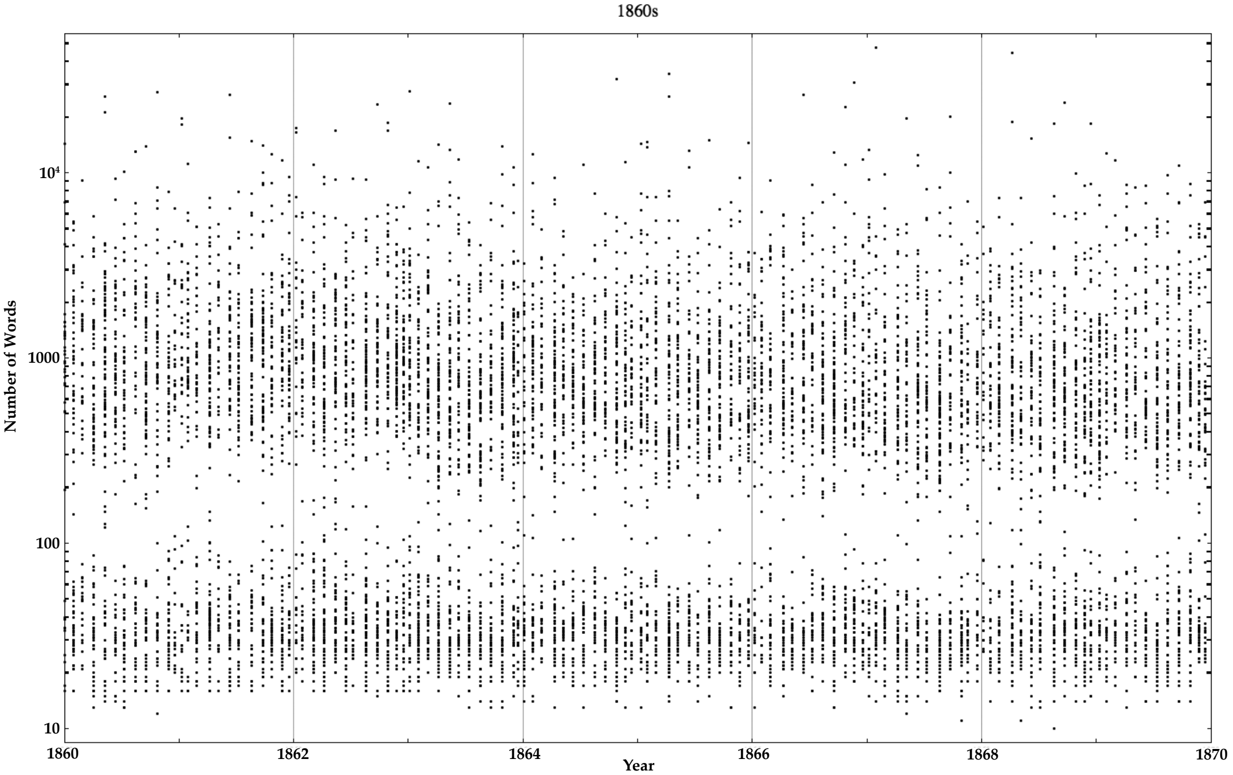

Fig. 2: Scatter plot created in Mathematica *showing distribution of Old Bailey trial lengths in the 1860s, by Tim Hitchcock and William Turkel.*

investigating the reason(s) why this may have been the case."[36]

Project leaders indicated that perhaps the chief motivation of using these tools is that they allow scholars to ask questions that would not have occurred to them otherwise: this is the power of unexpected discovery that opens paths to new thinking and further questioning. Dan Edelstein and Paula Findlen emphasize this point in their white paper for *Digging into the Enlightenment*: their geographic visualizations of historic letters "can serve a heuristic purpose, leading the user toward less known corners of the dataset."[37] "We're discovering research questions that we didn't have when we started off," echoed Peter Ainsworth, who led one of the teams on the *Digging into Image Data* project. Using an image segmentation algorithm developed during this project, Robert Markley and Michael Simeone were able to analyze digital images of 40 British and French historic maps of the Great Lakes dating from 1650 to 1800. Results showed marked inconsistencies between the depictions of some of the lakes' borders over this period; they also showed that the mapmakers' work did not become more "accurate" over time. Examining these inconsistencies, Simeone and Markley hypothesized that some "inaccuracies" reflected in the maps may actually correspond with water-level fluctuations and periods of prolonged ice cover. If they are able to collect more evidence to support this theory in their future research, Simeone and Markley "can begin to analyze maps prior to 1800 in order to provide

[36] Cohen, D., Hitchcock, T., Rockwell, G., et al. *Data Mining with Criminal Intent*. Final White Paper. August 31, 2011. Available at http://criminalintent.org/wp-content/uploads/2011/09/Data-Mining-with-Criminal-Intent-Final1.pdf.

[37] See page 7 of Edelstein's and Findlen's white paper for the NEH-funded portion of this project.

usable data for historical climate models and future projections."[38]

The promise for future revelatory explorations of our social and cultural heritage, explorations that should offer a more nuanced and expansive understanding of the human condition, is immense. The amount of work requisite to prepare and sustain the data for advanced research methodologies is nonetheless daunting. All the projects invested substantial time and effort in coaxing data corpora into reliable, "diggable" form, customizing analytical tools, or perfecting new, previously untested methodologies. The massive audio corpora at the heart of *Mining a Year of Speech* could be unlocked for investigation only after teams of researchers and students had used computer tools to align massive quantities of transcription with their corresponding audio data. *Towards Dynamic Variorum Editions* aspires to nothing less than the aggregation and morphological analysis of all extant works in classical languages; if the investigators reach this goal, they will have built the most powerful tool for philological inquiry in history. Yet for these scholars, much, much more remains to be done.

The amount of work requisite to prepare and sustain the data for advanced research methodologies is daunting.

Railroads and the Making of Modern America provides a compelling case for the potential of computationally intensive research initiatives as well as for the need for painstaking care and effort in performing them. For this project, investigators Richard Healey and William G. Thomas integrated U.S. census and railroad company data to arrive at a more accurate estimate of the population of nineteenth-century railroad workers than had been previously offered by historians (Figure 3).

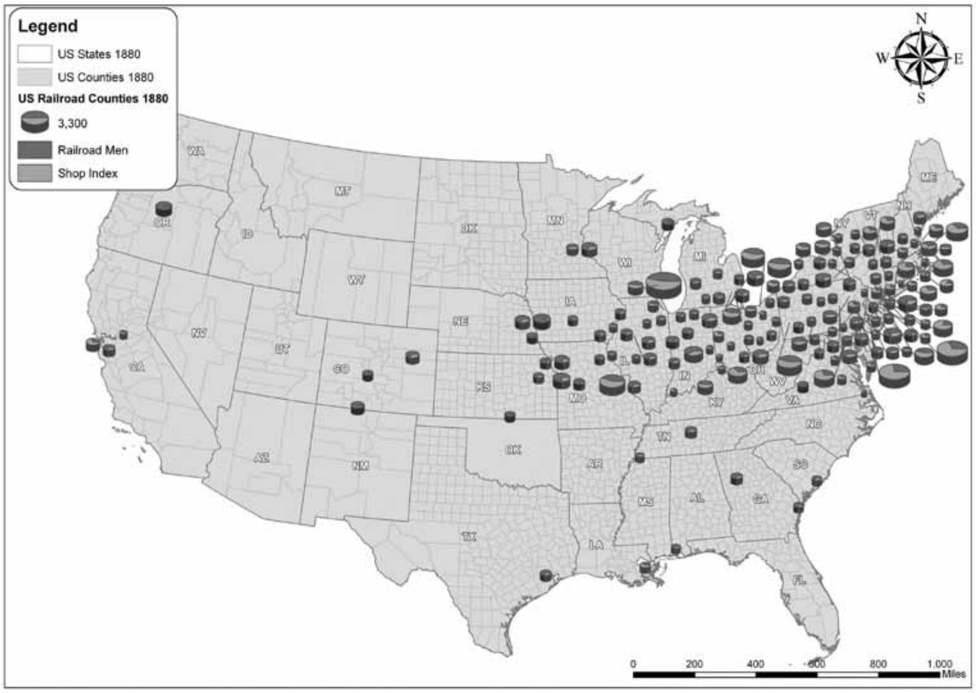

Fig.3: Integrated census ("Railroad Men") and railroad company ("Shop Index") data show the extent of railroad employment in the U.S. in 1880. Revealed are 38 "highly concentrated railroad centers."[39]

[38] E-mail from Robert Markley to Christa Williford, June 14, 2011.

[39] Thomas, William G., and Richard Healey. "Railroad Workers and Worker Mobility on the Great Plains," Western History Association, Lake Tahoe, Nevada, October 2010.

By performing this data integration, Healey and Thomas showed that if historians take 1880 census data at face value, they miss the greater part of the railroad's reach into the American workforce at that time. Only after working carefully with the data and investing significant time and hands-on effort could the scholars tell the full story.

3.2 Challenges and Concerns

Within the context of their day-to-day work, the challenges reported by investigators fell into four categories: funding issues, time issues, communication issues, and data issues.

Funding

- Funding is not always available in the amounts or for the resources most needed by investigators.
- Investigators need to continually seek external funding to sustain ongoing work.
- Many institutions lack long-term support for valuable project staff.
- Young scholars have difficulty getting travel support for meetings with collaborators.
- Computer storage infrastructure and processing cycles can be prohibitively expensive for humanists and social scientists working with large data sets.

Time

- Planning for and managing complex international, multidisciplinary collaborations takes extensive time.
- Data correction and tool development are time-consuming.
- Deep collaboration requires frequent synchronous communication, which is a major time commitment.
- Partners often have conflicting academic calendars and work schedules.

Communication

- Partners need patience and understanding to grasp perspectives of others from different backgrounds.
- Convincing technologists or computer scientists of the value of investing in humanities and social science work can be challenging.
- Managing expectations among partners with responsibilities for multiple projects can be tricky.

Data

- Data sharing requires shared tools and storage, and demands that partners trust one another.
- Making data “diggable” can be extremely labor-intensive. Error rates in data can be difficult to predict when planning a project and hard to account for in an analysis.
- Data management and analysis are iterative and cyclical, rather than sequential, activities.

The first of these categories, funding, is hardly surprising given these projects' complexity and their demand for powerful computer resources as well as for diverse kinds of expertise. In all cases, the dollar and pound limits imposed by the grant program[40] were not sufficient to achieve the researchers' full ambitions. Accordingly, the investigators framed their projects around more modest goals. They noted both positive and negative consequences of the limits on funding. The need to secure support beyond what was available through Digging into Data was a frustration (albeit not an unfamiliar one). All eight projects leveraged significant resources from their institutions or other concurrent and related grants to fill gaps in their budgets. At the same time, many investigators observed that the smaller Challenge grants seemed appropriate for the risk-taking and experimental work they most wanted to do.

More of the researchers complained about the 15-month time limit for the grants, a challenge that was complicated for some teams when agency restrictions prevented the four cooperating funders from disbursing grant funds at the same time, resulting in some partners working to significantly different project calendars. In several cases, the compressed schedule made it difficult to hire qualified students for grant positions.

Communication challenges arose both from dealing with the inevitable pressures experienced in any collaboration among busy academics as well as from managing differing expectations. Working in geographically dispersed locations, some project teams coped with extreme time differences by managing communication asynchronously. While effective, this strategy slowed the process of making collective decisions and increased chances of misunderstandings and confusion. Unforeseen technical issues, as well as legal restrictions, prevented some collaborators from sharing data and software as quickly and openly as their partners hoped. Negotiating productive solutions to these problems sometimes cost precious time.

Although the original emphasis of the Challenge was data analysis rather than management and organization, the experiences of investigators make clear that the two are deeply interdependent and that work in both is iterative and cyclical rather than sequential.

Data issues, as expected, were the thorniest of the challenges faced by investigators, though key differences between the projects surfaced, particularly the relative investment in manual manipulation versus automated "mining" and how each affected the results that were possible. Although the original emphasis of the Challenge was data analysis rather than management and organization, the experiences of investigators make clear that the two are deeply interdependent and that work in both is iterative and cyclical rather than sequential. This is particularly true for domains in which data are heterogeneous and unstructured. In these cases, manual intervention is often necessary. When the data themselves require significant scholarly effort, scholars often consider the resulting "clean" data to be just as important in potential impact as are the final research products the data make possible. As Richard Healey observed, "It has become clear that 'making data diggable' or providing 'keys' that unlock future digging potential may be just as important from a

[40] The awards were limited to 100,000 US dollars (NEH, NSF), 100,000 Canadian dollars (SSHRC), and 100,000 British pounds (JISC).

scholarly viewpoint, especially at this very early stage of the overall digging game."

Investigators coping with error-riddled data, such as the imperfect OCR applied to the digitized newspapers consulted in the *Railroads and the Making of Modern America* project, had to spend more time than they anticipated coaxing files into a usable format. Those who benefited from highly structured data sets or dealt with more homogeneous data types, such as team members on the *Data Mining with Criminal Intent* or *Digging into the Enlightenment* projects, could spend more energy on developing tools for interrogation and analysis. In some projects, such as *Mining a Year of Speech*, it was difficult for investigators to predict what data issues might arise until their projects were under way. When coping with very large-scale corpora, accounting for the potential effects of erroneous data on scholarly interpretation becomes a significant challenge that is impossible to address manually.

Two perspectives on dealing with error:

Large scale implies that there will necessarily be many errors, even if the error rate is very low. Say 1% of words in a corpus are "wrong" in some respect; e.g., wrongly annotated, classified, or indexed. 1% of a 100 million-word corpus is 1 million words with errors. Not only is it practically impossible to think of fixing 100% (even by crowd-sourcing etc.), one doesn't even know where the errors are. The inevitable consequence, it seems to me, is that humanities research in the larger scale is going to have to accept and deal with errors in a statistical fashion, just as has always been commonplace in science.

—John Coleman, *Mining a Year of Speech*

One of the long-standing but often unwritten tenets of GIS [geographic information systems] work is that if you combine/integrate multiple data sources, all of which contain different types of errors, the resulting errors in the outputs can be multiplied many-fold, sometimes to the point where the results are effectively uninterpretable. So combining heterogeneous data sources can dramatically amplify error-related problems.

—Richard Healey, *Railroads and the Making of Modern America*

The emergence of strong leadership, flexibility, and regular communication among the partners helped them resolve challenges as they arose. The partners working on *Data Mining with Criminal Intent,* for example, who brought together three relatively mature digital initiatives, attest that their weekly conference calls and frequent opportunities to meet face-to-face at scholarly conferences and workshops were key to their project's success. All the Digging into Data investigators came to their projects as seasoned researchers with a great deal of experience of grant-funded work; none seemed to have found the challenges they experienced to be out of the ordinary.

3.3 Other Challenges: The Academic Culture

In conversations at the 2011 program conference, investigators identified other concerns they deemed much more critical than insufficient resources, time, or communication. The first concern was preserving their ability to receive fair credit for their individual contributions within a collaborative effort, particularly contributions to tool development and data curation. Complicating the issue is the open question of what kinds of contributions constitute "legitimate" research for a given discipline. The extensive work required to prepare data for investigation by computer tools, including various forms of data cleanup, the addition of metadata, format conversions, etc., is not, at least in some disciplines, considered "real" research, even though such activities often call for the expertise of advanced scholars. Software development, particularly in the humanities, is often seen as a service rather than as a core research activity. Striking a balance between preparatory, experimental, and interpretive research work was an ongoing challenge for the grant partners.

Striking a balance between preparatory, experimental, and interpretive research work was an ongoing challenge for the grant partners.

Other major concerns of the investigators included a lack of effective training opportunities, difficulties faced by students and junior scholars pursuing computationally intensive research, and a lack of suitable avenues for publishing data-rich media. Each of these is discussed below.

Overall, however, project participants expressed great satisfaction with the Digging into Data program, emphasizing that it created opportunities that would not otherwise have been possible, and that its groundbreaking mission and design gave a sense of significance and urgency to their work well beyond their previous experience with grant-funded projects. Many of the partners' initiatives are expected to lead to other joint research efforts in the future.

3.4 Perspectives External to the Projects

In June 2011, participants in the Challenge convened at the National Endowment for the Humanities to present progress reports at a public symposium. Experts in each project domain contributed responses to these reports. After the symposium, the experts met with agency and CLIR representatives to discuss their impressions of the projects and their implications for the future of research. Their observations ranged widely, but many of them echoed the project participants' self-assessments. The consensus conveyed at the meeting was positive, both in response to the program and to the individual projects. Key points that arose in this discussion include the following:

1. **Level of investment:** The experts stressed that the amounts of the Digging into Data awards were not sufficient to allow researchers to achieve their mid- to long-term research goals. They agreed that major institutional commitments, in addition to further incentives from funders, would be necessary to allow this kind of research to mature.
2. **Refinements to tools developed for projects:** The experts recognized that the tools used in the projects were among their most

important deliverables. However, making these tools truly useful for other researchers, as most of the teams wish to do, will involve additional commitments that go beyond the original terms of the grants and the expertise levels represented among some of the teams. The experts suggested that institutions and agencies consider alternative programs for funding refinements to tools to make them broadly useful.

3. **Methods and technical training:**[41] The experts noted that research methods and related technical training for computationally intensive research in some disciplines, such as linguistics, is currently much more sophisticated than in other disciplines, and that this seemed to affect the rigor of the analyses employed in some of the projects. In their view, some investigators seemed too distracted by the process of learning new technologies to select the best available solutions for their research problems. More critically, the experts noted that many graduate education programs in the subject domains represented among the projects do not provide sufficient training to equip younger scholars to lead future research initiatives of this kind. They recommended that institutions consider cross-disciplinary research methods training that would introduce both undergraduates and graduates to sound data management practices; to multiple modes of data analysis, including visual, geographical, and statistical; and to skills necessary for managing collaborations with scholarly and professional experts from multiple backgrounds and departments. They noted that future projects would need even stronger expertise in relevant technologies and research methodologies to achieve their potential.
4. **Documentation and publication:**[42] Respondents noted an unfortunate incompatibility between the most important deliverables for these projects, including new tools, software, and data and their associated documentation, and peer-reviewed scholarly communication outlets available to researchers. They strongly recommended that institutions and agencies expand the range of outlets for scholars and broaden opportunities for earning credit for contributions that do not conform to traditional models for conference presentations, journal articles, and monographs. The lack of peer-reviewed publication alternatives is a significant disincentive for younger scholars to pursue computationally intensive research.

[41] Formal and informal education and training opportunities related to data-intensive research methodologies are growing more common worldwide, although formal training is still more common in the sciences (such as bioinformatics) than in the social sciences or humanities.

[42] See, for example, the MediaCommons project (http://mediacommons.futureofthebook.org/about-mediacommons); Ball, A., and M. Duke, "Data Citation and Linking," DCC Briefing Papers (Edinburgh: Digital Curation Centre, 2011), available at http://www.dcc.ac.uk/resources/briefing-papers/; Fitzpatrick, K. *Planned Obsolescence: Publishing, Technology, and the Future of the Academy* (New York: NYU Press, 2011); Withey, L., et al., "Sustaining Scholarly Publishing: New Business Models for University Presses" (2011); available at http://mediacommons.futureofthebook.org/mcpress/sustaining/.

5. **Sustainability:**[43] Respondents expressed concern about the mid- to long-term sustainability of data and tools produced by these projects, and encouraged agencies and institutions to give this issue greater attention. They observed that the low level of participation in these kinds of projects at smaller institutions may reflect an inability of such institutions to offer support for incubating and sustaining these kinds of research. They noted that it would be impractical for investigators leading these projects to assume these burdens themselves. Further, they observed that outmoded assessment practices meant that younger, non-tenured scholars do not have sufficient incentives for contributing to established digital projects, embarking upon new digital research, or experimenting with new, computationally intensive modes of analysis, putting the long-term sustainability of digital projects at even greater risk.
6. **Extensibility of methodologies:** Respondents felt that most of the methodologies demonstrated in the projects were extensible beyond the domains in which they were developed, but they cautioned against seeing them as *easily* extensible. Expertise in research methods and modes of analysis will be necessary to adapt these methods successfully to new contexts.

3.5 Moving Forward

Digging into Data recipients made the following recommendations:

- Increase incentives for engaging in collaborative and multidisciplinary research, particularly for students and junior faculty.
- Establish standards for assessing collaborative and multidisciplinary work, including work on data and tools for exploration and analysis of data.
- Facilitate cross-disciplinary research tools and methods training for students, staff, and faculty.
- Support travel for those engaged in collaborative projects, particularly for students and junior faculty.
- Facilitate sharing of hardware, software, and data across institutions.
- Use licensing agreements and memoranda of understanding to clarify partners' legal and ethical responsibilities on collaborative projects.
- Work toward multi-institutional strategies for data management that include long-term preservation.
- Increase the range of options for data-rich and multimedia scholarly publication across disciplines, particularly open-access publication.

[43] The academic library community has already taken up the challenge of supporting data-intensive research, but much work remains to be done to establish sustainable best practices for the management of research data that will work globally and across disciplines, at institutions both large and small. See Marcum, Deanna, and Gerald George, eds. *The Data Deluge: Can Libraries Cope with E-Science*? (Westport, CT: Greenwood Press, 2010).

The complexity of the Challenge projects, their diversity of approach, and the deep interdependency of their partnerships hold implications for the future—implications for the funding and staffing of academic departments, the training of academic professionals, and the education of students. The recipients of the first Digging into Data Challenge grants have offered suggestions in each of these areas.

When asked about how institutions of higher learning might better support computationally intensive research, investigators recommended increasing incentives to engage in collaborative and multidisciplinary projects, and establishing clearer, yet more flexible, standards for assessing such projects. They emphasized the need to cultivate the interest of undergraduates, graduate students, and junior faculty in participating in digital research. The involvement of graduate and undergraduate students in the Digging into Data projects was essential, most often through paid research internships and occasionally through coursework. Most investigators reported that their students found the work intellectually rewarding. As teaching faculty, the investigators were deeply committed to helping maintain this level of engagement throughout the project; they avoided relegating students solely to repetitive, low-level tasks and instead sought opportunities that offered the greatest creative challenges.

Among the more formal types of incentives mentioned by the investigators were cross-disciplinary internship or postdoctoral programs; multi-institutional, multidisciplinary, co-taught courses; and improved research methods, information architecture, and project management training for graduate students. Finally, while the assessment of digital scholarship has received increased attention in the past decade[44] and many institutions have made progress in this area, the attitudes of the Digging into Data investigators suggested that assessment practices are changing much more slowly than they would like.

Increasing the number of opportunities for interaction among project partners, especially face-to-face meetings, was by far the most frequent suggestion made by project participants. Whether or not funds for travel had been included in their Digging into Data grants, most researchers found opportunities to meet at least once, and often several times, during the grant period. Participants credited these informal meetings with advancing their projects, building trust, and ensuring overall success. In addition to meeting face to face, many investigators credited frequent teleconferences or videoconferences among partners with project success. A past history of collaboration and, in two cases, the employment of a single individual by two of the partner institutions, strengthened communication and understanding among project stakeholders.

As for what higher education at the national and international levels might do, investigators echoed their strong support for broadening the range of peer-reviewed dissemination opportunities for humanities and social science researchers, including digital

[44] See links collected at "Evaluating Digital Work for Tenure and Promotion: A Workshop for Evaluators and Candidates." Modern Languages Association. Available at http://www.mla.org/resources/documents/rep_it/dig_eval.

monographs, online conference proceedings that are edited or commented upon by conference participants, and online journals and data repositories. These should include affordances for embedding multimedia and interactive data visualizations within published work. New, stable, and secure collaborative research environments shared across institutions are also required to support this research.

Clearly this supporting cyberinfrastructure demands major, long-term investments. The ideal system will be multi-institutional, multidisciplinary, and distributed rather than centralized. Robust architecture of this kind will take careful planning and dedicated, professional maintenance—scholarly expertise in this area will be essential, yet relying on this expertise alone will not be enough. The Digging into Data researchers and their like-minded colleagues are deeply committed to advancing research in their disciplines, and understandably their priorities lie in the pursuit of new work rather than in the long-term preservation of their past and current research products. Even those scholars who are keenly interested in sustainability often lack the experience and training for planning research projects in ways that would enable repositories to accept their data or make it possible for other scholars to reuse those data in the context of other disciplines. Nevertheless, most researchers doing computationally intensive work recognize that their data should be stored long term, and stored in ways that are permanently linked to information about their provenance as well as to any future work that may rely upon them. In addition to data, algorithms, interfaces, and other tools used to explore data must be preserved and given proper context. Given these demands, it is clear that new models of publication will call for publishers, libraries, and other information suppliers and repositories to have a sophisticated understanding of evolving research practice as well as evolving technologies.

New models of publication will call for publishers, libraries, and other information suppliers and repositories to have a sophisticated understanding of evolving research practice as well as evolving technologies.

Some of the Digging into Data investigators emphasized that embracing open access, data sharing, and open-source software development would be critical to advancing research like theirs. While support for making open data a prerequisite of funding among Challenge grant recipients was not universal, there was considerable appetite for the agencies to encourage openness more forcefully.[45] The *Towards Dynamic Variorum Editions* team argues persuasively for the benefits of openness for the public at large; they write, “We are able to develop intellectual conversations that can, if we so choose, serve to advance our understanding and to reach a wider audience—and to both at once without compromise.”[46] As a practical means to this

[45] In the second Digging into Data Challenge, the eight participating agencies have been advocating increased transparency in handling intellectual property–related issues. For example, JISC Director of the Strategic Content Alliance Stuart Dempster reports “paying site visits, forwarding exemplars of good practice and giving grantees a strong steer on licensing of project outputs” (e-mail to Brett Bobley, March 23, 2012). The JISC website has helpful information and tools related to intellectual property rights and licensing. See, for example, Naomi Korn, “Embedding Creative Commons Licenses into Digital Resources.” JISC Strategic Content Alliance Briefing Paper, 2011. Available at http://www.jisc.ac.uk/publications/programmerelated/2011/scaembeddingcclicencesbp.aspx.

[46] See page 18 of the project white paper. See also Crane, G. “Give us Editors! Re-inventing the Edition and Re-thinking the Humanities.” In *Online Humanities Scholarship: The Shape of Things to Come*. (University of Virginia: Mellon Foundation, 2010). Available at http://cnx.org/content/m34316/latest/.

more inclusive, international academic culture, the *Digging into Image Data* partners advocate for the use of model partnership agreements such as memoranda of understanding that both protect researchers' rights to credit for their work and establish rules for sharing hardware, software, data, and credit among partners. Other teams stressed that making data available through application programming interfaces is not sufficient for many purposes where the ability to manipulate data is critical; they noted that the corpora listed on the Digging into Data website[47] seem to offer varying degrees of access to researchers.

While all Challenge respondents reported that their projects had met or exceeded their expectations, several factors seem to have enhanced their perceptions of success. Trust among partners, whether built formally through memoranda of understanding or informally through frequent communication about expectations and roles, was a major factor influencing these perceptions. Those working within more familiar collaborative partnerships in which they had already built bonds of trust generally found the planning and coordination of their projects easiest; for others, project management and communication required more formality and greater effort. Designing projects that provide mutual and equal benefits to partners was another important factor contributing to the satisfaction of participants: most projects set out to make significant impacts on multiple fields (i.e., computer science *and* humanities or social science). The Digging into Data teams stressed the importance of involving computer science faculty and students not in supportive or developers' roles but as active research partners; given the limited level of funding available, the additional incentive of contributing to the advancement of computer science was seen as critical. From the perspective of the computer scientists and engineers on many of these teams, the complexity, ambiguity, and unstructured nature of humanities and social science data posed intellectual and creative challenges beyond those they had encountered on other projects.

> *The researchers viewed their work as augmenting and transforming, rather than supplanting, research practice within their disciplines.*

In general, the researchers viewed their work as augmenting and transforming, rather than supplanting, research practice within their disciplines. The new methodologies they and others continue to develop will ultimately effect a much broader transformation: one that calls into question what the boundaries of those disciplines should ultimately be. It is time for international leaders and funders in higher education to take note of these changes and begin to adapt. There are exciting opportunities to incorporate computationally intensive research collaborations into curricula, staffing models, and professional development programs. This much is clear: "big data" are not just for scientists anymore. The new, bigger, and broader questions that computationally intensive research makes possible will not be simple to answer, and the tensions between the multiple, often oppositional, research traditions will not be easy to resolve. Yet we in the academy are becoming one culture. Whether we embrace and accommodate our differences is for us to decide.

[47] http://www.diggingintodata.org/.

Afterword: A Charge to Stakeholders

> *With ten billion digital objects being created every day, for those who work in the humanities, for those who deal with the human record, it cannot be a question of whether computer tools will be an important part of the humanistic disciplines—they will need to be. This will require re-imagining the humanities, rethinking and re-envisioning the way humanists go about their work.*
>
> —Dean Rehberger, *Digging into Image Data to Answer Authorship-Related Questions*

One Culture documents the promising consequences of innovative, and sometimes surprising, partnerships. These partnerships cross disciplines—most frequently humanities, computer science, and engineering; create genuine interdependencies; and provide a framework for new kinds of research and inquiry, new methods of execution, and exciting discovery that would not be possible if the partnering experts remained incurious of one another. This report does not endorse a wholesale blending of academic departments and fields, but is confident that strategically planned instances of collaboration can indeed yield compelling insight concerning both our cultural legacy and the digital tools, applications, and resources used in the search for new knowledge.

The various components of higher education, from universities to departments, centers, and support services, often define themselves by exclusivity and singularity of purpose. We compete one with one other; we measure ourselves in comparison and contrast with one another; and we hold tightly to our idiosyncrasies as defining elements of status. There is a palpable tension between these inherited conceptual notions of separate, particular, and solitary, and a tripartite, networked infrastructure of information, modes of delivery, and human expertise that has no "place."

The future successes of these and subsequent Digging into Data projects rests in part on our willingness to conceive ourselves less in traditional slots and silos and more as a flexible and imaginative cohort. Librarians, information technology specialists, computer scientists, scholars, administrators, and publishers who represent the various components of scholarly information—discovering, reconstituting, publishing, and sharing knowledge, and keeping its various manifestations securely preserved and accessible—are more interrelated and interdependent. The inherited norms, customs, traditions, and institutions that have structured research and teaching now need to be constructively challenged, redefined, and reassembled. Higher education could make enormous contributions to assure its vitality, expanding its capacity for future discovery while not compromising its exactitude and rigor; the prized idiosyncrasies and powerful identities would remain intact.

The multitude of stakeholders represented within these projects is encompassing and vital, and there are still others whose contributions are sorely needed. We observed groups of young scholars conducting completely new modes of research, usually in concert with tenured, established faculty, academic technologists, and librarians. Specialists in advanced programs in preservation and interpretation have clear roles to play, as do the creators of new digital tools and resources, data curation professionals and archivists, scholarly societies, liberal arts centers, programs in support of pedagogy, foundations, government document centers, supercomputer centers, and library and information schools. The complexity of the Digging into Data projects, despite their small number, offers an enormous opportunity for, and essentially requires, this array of stakeholders to build new bases of support, reach new constituencies, cultivate funding streams, and develop lasting, mutually sustaining connections between traditionally disparate sectors, to seek together effective and efficient means of support and continuity for the humanities and social sciences in a digital era.

For case studies describing each of the eight 2009 Digging into Data projects, visit the web-based version of this report at http://www.clir.org/pubs/reports/pub151.

Suggestions for Further Reading

Arms, William, and Ronald Larsen, eds. 2007. *The Future of Scholarly Communication: Building the Infrastructure for Cyberscholarship*. NSF/JISC Workshop, Phoenix, Arizona, April 17–19, 2007. Available at: http://www.sis.pitt.edu/~repwkshop/SIS-NSFReport2.pdf.

Bartscherer, Thomas, and Roderick Coover. 2011. *Switching Codes: Thinking Through Digital Technology in the Humanities and the Arts*. Chicago: University of Chicago Press.

Daniels, Morgan, Ixchel M. Faniel, Kathleen Fear, and Elizabeth Yakel. 2012. "Managing Fixity and Fluidity in Data Repositories." In *Proceedings of the 2012 iConference,* 2012. Toronto: ACM Press. Available at: http://www.dipir.org/publications.html.

Fitzpatrick, Kathleen. 2011. *Planned Obsolescence: Publishing, Technology, and the Future of the Academy*. New York: New York University Press. See also: http://www.plannedobsolescence.net/.

Friedlander, Amy. 2009. "Asking Questions and Building a Research Agenda for Digital Scholarship." In *Working Together or Apart: Promoting the Next Generation of Digital Scholarship*, 1-15. Washington, DC: Council on Library and Information Resources. Available at: http://www.clir.org/pubs/reports/pub145.

Gold, Matthew K., ed. 2012. *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

High Level Expert Group on Scientific Data. 2010. *Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data*. Report to the European Commission. Available at: http://cordis.europa.eu/fp7/ict/e-infrastructure/high-level-group_en.html.

Kroll, Susan, and Rick Forsman. 2010. *A Slice of Research Life: Information Support for Research in the United States*. Report commissioned by OCLC Research in support of the RLG Partnership. Available at: http://www.oclc.org/research/publications/library/2010/2010-15.pdf.

Lynch, Clifford A. (2008). "Big Data: How do Your Data Grow?" *Nature*, 455.7209. Abstract available at: http://www.nature.com/nature/journal/v455/n7209/full/455028a.html.

_____. 2010. "Imagining a University Press System to Support Scholarship in the Digital Age." *Journal of Electronic Publishing*, 13.2. DOI: http://dx.doi.org/10.3998/3336451.0013.207.

Lyon, Liz. 2009. *Open Science at Web Scale: Optimising Participation and Predictive Potential*. London: Joint Information Systems Committee. Available at: http://www.jisc.ac.uk/publications/reports/2009/opensciencerpt.aspx.

Maron, Nancy L., and K. Kirby Smith. 2008. *Current Models of Digital Scholarly Communication. Results of an Investigation Conducted by Ithaka for the Association of Research Libraries*. Washington, DC: Association of Research Libraries. Available at: http://www.arl.org/bm~doc/current-models-report.pdf.

McDonald, Diane. 2012. *Value and Benefits of Text Mining*. London: Joint Information Systems Committee. Available at: http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx.

McGann, Jerome, ed. 2010. *Online Humanities Scholarship: The Shape of Things to Come*. Houston: Connexions. Available at: http://cnx.org/content/col11199/latest/.

Miller, Kerry. 2012. *5 Steps to Research Data Readiness*. DCC Briefing Papers. Edinburgh: Digital Curation Centre. Available at: http://www.dcc.ac.uk/resources/briefing-papers/five-steps-research-data-readiness.

National Science Foundation Cyberinfrastructure Council. 2007. *Cyberinfrastructure Vision for 21st Century Discovery*. Available at: http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf.

National Science Foundation Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging. 2011. Final Report. Available at: http://www.nsf.gov/od/oci/taskforces/TaskForceReport_CampusBridging.pdf.

Pryor, Graham, ed. 2012. *Managing Research Data*. London: Facet Publishing.

Ramsay, Stephen. 2011. *Reading Machines: Toward an Algorithmic Criticism*. Champaign, IL: University of Illinois Press.

Sehat, Connie Moon, and Erika Farr. 2009. "The Future of Digital Scholarship: Preparation, Training, Curricula." Washington, DC: Council on Library and Information Resources. Available at: http://www.clir.org/pubs/resources/archives/SehatFarr2009.pdf.

Unsworth, John, Roy Rosenzweig, Paul Courant, Sarah E. Frasier, and Charles Henry. 2006. *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. New York: American Council of Learned Societies. Available at: http://www.acls.org/programs/Default.aspx?id=644.

Van den Eynden, Veerle, Libby Bishop, Laurence Horton, and Louise Corti. 2010. "Data Management Practices in the Social Sciences." Essex: UK Data Archive. Available at: http://www.data-archive.ac.uk/about/publications.

Van de Eynden, Louise Corti, Matthew Woolard, Libby Bishop, and Laurence Horton. 2011. *Managing and Sharing Data: Best Practice for Researchers*. Essex: UK Data Archive. Available at: www.data-archive.ac.uk/media/2894/managingsharing.pdf.

van der Graaf, Maurits, and Leo Waaijers. 2011. *A Surfboard for Riding the Wave: Towards a Four Country Action Programme for Research Data*. Copenhagen: The Knowledge Exchange. Available at: http://www.knowledge-exchange.info/Default.aspx?ID=469.

Waters, Donald. 2012. "Digital Humanities and the Changing Ecology of Scholarly Communications." Opening keynote presentation at the TELDAP International Conference 2012, Taipei, Taiwan, February 21. Available at: http://msc.mellon.org/staff-papers.

_____. 2008. "Open Access Publishing and the Emerging Infrastructure for 21st-Century Scholarship." *Journal of Electronic Publishing* 11:1 (Winter). DOI: http://dx.doi.org/10.3998/3336451.0011.106.

Whyte, A. and J. Tedds. 2011. *Making the Case for Research Data Management*. DCC Briefing Papers. Edinburgh: Digital Curation Centre. Available at: http://www.dcc.ac.uk/resources/briefing-papers.

Digging into Data Challenge Award Recipients, 2009: Project Participants

Using Zotero and TAPOR on the Old Bailey Proceedings: Data Mining with Criminal Intent (DMCI)

Dan Cohen (George Mason University [GMU], US) served as principal investigator for the NEH-funded portion of the project and managed the workflow and partnership at GMU.

Fred Gibbs (George Mason University, US) wrote the Zotero plug-in that extracts trial transcripts from *The Proceedings of the Old Bailey Online*, organized them, and sent their text to mining services. He also conducted research using the project's tools.

Tim Hitchcock (University of Hertfordshire, UK) served as principal investigator for the JISC-funded portion of the project as well as liaison between the Old Bailey team and other project partners, ensuring that data were available in the right form. He also worked with Turkel on detailed textual analysis and on organizing the stakeholders' engagement with the project.

Geoffrey Rockwell (University of Alberta, Canada) served as co-principal investigator for the SSHRC-funded portion of the project and worked with Sander and John Simpson to implement the data warehouse model for data from *The Proceedings of the Old Bailey Online* in preparation for the Old Bailey Application Programming Interface (OBAPI).

Jörg Sander (University of Alberta, Canada) worked with Rockwell and John Simpson to select and then implement the data warehouse model for data from *The Proceedings of the Old Bailey Online* in preparation for the Old Bailey Application Programming Interface (OBAPI).

Robert Shoemaker (University of Sheffield, UK) managed the implementation of the Old Bailey Application Programming Interface (OBAPI) at Sheffield.

Stéfan Sinclair (McGill University, Canada, previously McMaster University) served as co-principal investigator for the SSHRC-funded portion of the project and designed a new, simplified skin (a combination of tools) to optimize Voyeur/Voyant Tools' visual ease of use.

Sean Takats (George Mason University, US) worked with Cohen and Gibbs on the incorporation of the plug-in that extracts trial transcripts from *The Proceedings of the Old Bailey Online* and imports them into the Zotero research management tool.

William Turkel (University of Western Ontario, Canada) imported project data into *Mathematica* to create visualizations for the project.

Other contributors and stakeholders:
Cyril Briquet (McMaster University, Canada)
Hugh Couchman (SHARCNET, Canada)
Clive Emsley (Open University, UK)
Margaret Hunt (Amherst College, US)
Jamie McLaughlin (University of Sheffield, UK)
Michael Pidd (University of Sheffield, UK)
Milena Radzikowska (Mount Royal University, Canada)
Kevin Sienna (Trent University, Canada)
John Simpson (University of Alberta, Canada)
Kirsten C. Uszkalo (Independent Scholar)

Digging into the Enlightenment: Mapping the Republic of Letters

Nicole Coleman (Stanford University, US) provided leadership for the day-to-day work on the project, including providing collaborative research support and facilitating project documentation and communication.

Peter Damian-Grint (Electronic Enlightenment Project, University of Oxford, UK) serves as correspondence editor for the Electronic Enlightenment Project and contributed subject expertise in French language and literature.

Dan Edelstein (Stanford University, US) served as principal investigator of the NEH-funded portion of the project and provided subject expertise in European history, literature, and culture.

Paula Findlen (Stanford University, US) lent subject expertise for the project in European history and culture and coauthored a project white paper with Edelstein.

Robert McNamee (University of Oxford, UK) served as principal investigator of the JISC-funded portion of the project. He heads the Electronic Enlightenment Project, the major source of data and metadata for the initiative, and offered both technical and subject expertise.

Mark Rogerson (University of Oxford, UK) is technical editor of the Electronic Enlightenment Project and offered data expertise for the project.

Rachel Shadoan (University of Oklahoma, US) worked with Weaver on the analysis of project data using the Improvise advanced visual analytics tool, including software engineering and usability evaluation.

Chris Weaver (University of Oklahoma, US) served as principal investigator of the NSF-funded portion of the project, involving implementing the Improvise advanced visual analytics tool.

Other contributors and stakeholders:
Density Design Research Lab (Polytechnical Institute, Italy)
Keith Baker (Stanford University, US)
John Bender (Stanford University, US)
Giovanna Ceserani (Stanford University, US)
Jon Christensen (Stanford University, US)
Dario Generali (National Publication of the Works of Antonio Vallisneri, Italy)
Anthony Grafton (Princeton University, US)
Carl-Olof Jacobson (Uppsala University, Sweden)
Wijnand W. Mijnhardt (Utrecht University, the Netherlands)
Peter M. Miller (Bard Graduate Center, US)
Guliano Pancaldi (International Center for the History of Universities and Science, Italy)
Mark Peterson (University of California at Berkeley, US)
Jessica Riskin (Stanford University, US)
Jacob Soll (Rutgers University, US)
Francoise Wacquet (French National Center for Scientific Research, France)
Caroline Winterer (Stanford University, US)

Towards Dynamic Variorum Editions

Alison Babeu (Tufts University, US) is the digital librarian for the Perseus Digital Library and contributed both data and subject expertise to the project.

David Bamman (Tufts University, US) is a computational linguist who contributed both technical and subject expertise to the project.

Federico Boschetti (Institute of Computational Linguistics of the National Research Council, Italy) worked with Robertson on customizing optical character recognition engines for ancient Greek source texts.

Lisa Cerrato (Tufts University, US) is managing editor of the Perseus Project and contributed both data and subject expertise.

Gregory Crane (Tufts University, US) served as principal investigator for the NEH-funded portion of the project.

John Darlington (Imperial College London, UK) served as principal investigator for the JISC-funded portion of the project.

Brian Fuchs (Imperial College London, UK) designed and implemented a scalable computer infrastructure for processing large data sets of page images from books.

David Mimno (University of Massachusetts Amherst, US) is a computer scientist who contributed both technical and analytical expertise to the project.

Bruce Robertson (Mount Allison University, Canada) served as principal investigator for the SSHRC-funded portion of the project and worked with Boschetti and a team of undergraduate students on producing classifiers suitable for the optical character recognition of ancient Greek source texts.

Rashmi Singhal (Tufts University, US) is lead programmer for the Perseus Project and contributed technical expertise.

David Smith (University of Massachusetts Amherst, US), a computer scientist, contributed technical and analytical expertise to the project.

Mining a Year of Speech

Lou Bernard (University of Oxford, UK) is assistant director of Oxford Computing Services and has long been responsible for the distribution and maintenance of the British National Corpus, the data set at the heart of the JISC-funded portion of the project.

Christopher Cieri (University of Pennsylvania, US) is executive director of the Linguistic Data Consortium at the University of Pennsylvania and contributed administratively and substantively to the project. He is an expert in corpus-based phonetics.

John Coleman (University of Oxford, UK) is professor of phonetics and served as principal investigator for the JISC-funded portion of the project, based at Oxford's Phonetics Laboratory, which he directs.

Sergio Grau (University of Oxford, UK) is a research fellow at University of Oxford and performed most of the analysis on the British National Corpus data for the project.

Gregory Kochanski (University of Oxford, UK) is a senior research fellow at Oxford's Phonetics Laboratory and contributed subject and analytical expertise to the project.

Mark Liberman (University of Pennsylvania, US) served as principal investigator for the NSF-funded portion of the project, based at the Linguistics Data Consortium.

Ladan Ravary (University of Oxford, UK) is a research fellow at Oxford's Phonetics Laboratory and an expert in the engineering of speech recognition and alignment technologies.

Jonathan Robinson (British Library, UK) is lead content specialist in Sociolinguistics and Education at the Social Sciences Collections and Research Department of the British Library and contributed technical, managerial, and subject expertise to the project.

Joanne Sweeney (British Library, UK) is a content specialist in the Social Sciences Collections and Research Department of the British Library and contributed technical expertise and support to the project.

Jiahong Yuan (University of Pennsylvania, US) is assistant professor of linguistics and the developer of the Penn Phonetics Lab Forced Aligner, a tool that was adapted and used extensively for the project.

Harvesting Speech Data Sets for Linguistic Research on the Web

Mats Rooth (Cornell University, US) served as principal investigator for the NSF-funded portion of the project. A computational linguist, he was responsible for working with graduate and undergraduate students at Cornell to design and implement the harvesting methodology used for the project.

Michael Wagner (McGill University, Canada), a linguist, served as principal investigator for the SSHRC-funded portion of the project and was responsible for leading the analysis of data harvested during the course of the project, which included the comparison of results of computational statistical analysis with analysis using traditional formal-linguistics methodologies.

Jonathan Anthony Howell (McGill University, Canada) is a postdoctoral fellow who specializes in statistical and machine learning methodologies for phonetic analysis. His doctoral dissertation project formed the basis for the collaboration funded through the Digging into Data program.

Structural Analysis of Large Amounts of Music Information

J. Stephen Downie (University of Illinois Urbana Champaign, US) is a music information retrieval and computational musicology specialist based at the Graduate School of Library and Information Science at UIUC who led the NSF-funded portion of the project, which, once complete, will have generated hundreds of thousands of structural analysis files for musical pieces.

David De Roure (formerly University of Southampton, now University of Oxford, UK) is a computer scientist with expertise in distributed information systems, Web 2.0, and Semantic Web technologies and served as the principal investigator of the JISC-funded portion of the project, which included the development of a standardized ontology for musical structures based upon the Resource Description Framework.

Ichiro Fujinaga (McGill University, Canada), associate professor of music technology, is the principal investigator of the SSHRC-funded portion of the project. He directed the preparation of the open-source “ground truth” data against which the team measured the performance of the structural analysis algorithms.

Advisers, data contributors, and other contributors:
Mert Bay (University of Illinois Urbana Champaign, US)
John Ashley Burgoyne (McGill University, Canada)
Alan B. Craig (University of Illinois Urbana Champaign, US)
Tim Crawford (Goldsmiths University of London, UK)
Andreas Ehmann (University of Illinois Urbana Champaign, US)
Benjamin Fields (Goldsmiths University of London, UK)
Linda Frueh (Internet Archive, US: data contributor)

Eric J. Isaacson (Indiana University, US)
Lisa Kahlden (Anthology of Recorded Music, Database of Recorded American Music: data contributor)
Kevin R. Page (Oxford e-Research Centre, University of Oxford, UK)
Yves Raimond (British Broadcasting Corporation, UK)
Jordan B. L. Smith (formerly McGill University, Canada, now Queen Mary, University of London, UK)
Michael Welge (National Center for Supercomputing Applications, University of Illinois Urbana Champaign, US)

Music annotators:
Christa Emerson, David Adamcyk, Elizabeth Llewellyn, Meghan Goodchild, Michel Vallières, Mikaela Miller, Parker Bert, Rona Nadler, and Rémy Bélanger de Beauport

Digging into Image Data to Answer Authorship-Related Questions

Core participants involved in all project elements:

Peter Ainsworth (University of Sheffield, UK) served as principal investigator for the JISC-funded portion of the collaboration and contributed subject and technical expertise as director of the Online Froissart Project.

Simon Appleford (University of Illinois Urbana Champaign, US) is a cultural historian and digital humanist based at the Institute for Computing in Humanities, Arts, and Social Science at the University of Illinois. He contributed as a subject specialist to the project.

Peter Bajcsy (formerly University of Illinois Urbana Champaign, now National Institute of Standards and Technology, US) was the founder and leader of the Image Spatial Data Analysis Group at the National Center for Supercomputing Applications, University of Illinois. He led project planning and served as co-principal investigator for the NSF-funded portion of the project.

Steve Cohen (Michigan State University, US), an evaluation specialist, helped with project assessment throughout the grant.

Matthew Geimer (Michigan State University, US), a computer scientist, contributed technical and analytical expertise to the project.

Jennifer Guiliano (formerly University of Illinois Urbana Champaign, now assistant director for the Maryland Institute for Technology in the Humanities, University of Maryland) served as project manager for the NSF-funded portion of the grant and contributed subject expertise as a cultural historian and digital humanist.

Rob Kooper (University of Illinois Urbana Champaign, US) is a computer scientist and senior research programmer for the Image Spatial Data Analysis Group at the National Center for Supercomputing Applications. He served as co-principal investigator for the NSF-funded portion of the project.

Michael Meredith (University of Sheffield, UK) contributed computer science expertise and served as developer for the JISC-funded portion of the project.

Dean Rehberger (Michigan State University, US) is director of MATRIX, the Center for Humane Arts, Letters, and Social Sciences Online at Michigan State University (MSU) and history adjunct curator of the MSU Museum. He served as principal investigator for the NEH-funded portion of the project and contributed subject expertise in the digital humanities generally as well as expertise specific to his involvement with the Quilt Index.

Justine Richardson (Michigan State University, US) served as project manager for the NEH-funded portion of the project based at MATRIX, Michigan State University. She also contributed subject expertise in cultural history and digital humanities as well as expertise specific to her involvement with the Quilt Index.

Michael Simeone (University of Illinois Urbana Champaign, US) contributed as a subject expert in historical cartography and served as project manager for the NSF-funded portion of the project based at the Institute for Computing in Humanities, Arts, and Social Science, University of Illinois.

Contributing additional expertise in computer science:
Wayne Dyksen (Michigan State University, US)
Alhad Gokhale (Independent Researcher)
Zach Pepin (Michigan State University, US)
William Punch (Michigan State University, US)
Tenzing Shaw (University of Illinois Urbana Champaign, US)

Contributing additional expertise in quilt making and quilt history:
Beth Donaldson (Michigan State University Museum, US)
Amy Milne (Alliance for American Quilts, US)
Marsha MacDowell (Michigan State University and MSU Museum, US)
Amanda Silkarskie (Michigan State University, US)
Mary Worrall (Michigan State University Museum and Quilt Index Project, US)

Other consulting quilt experts:
Karen Alexander, Barbara Brackman, Janneken Smucker, Merikay Waldvogel, Jan Wass and members of the American Quilt Study Group e-mail discussion list

Contributing art historical and other expertise related to medieval manuscripts:
Heather Tennyson (University of Illinois Urbana Champaign, US)
Colin Dunn (Scriptura Limited, University of Oxford, UK)
Godfried Croenen (University of Liverpool, UK)
Caroline Prud'homme (University of Toronto, Canada)
Victoria Turner (University of Warwick, UK)

Anne D. Hedeman (University of Illinois Urbana Champaign, US)
Natalie Hanson (University of Illinois Urbana Champaign, US)

Contributing expertise in historical cartography and environmental literatures:
Robert Markley (University of Illinois Urbana Champaign, US)

Railroads and the Making of Modern America

Project Participants

William G. Thomas, III (University of Nebraska-Lincoln, US) served as Principal Investigator of the NEH-funded portion of the project and contributed as a data and subject expert in American history.

Richard Healey (University of Portsmouth, UK) served as Principal Investigator of the JISC-funded portion of the project and also contributed as a data and subject expert in American railroad history, geography, and geographic information systems (GIS).

Ian Cottingham (University of Nebraska-Lincoln, US) is Chief Software Architect in the Department of Computer Science and Engineering at UNL and contributed technical and analytical expertise, leading the team designing and building the Aurora Engine for the exploration of geographic data.

Leslie Working (University of Nebraska-Lincoln, US) is a Graduate Instructor in History based at the Center for Digital Research in the Humanities at the University of Nebraska-Lincoln and contributed project management expertise for the NEH-funded portion of the project, helping to supervise a team of students doing data checking and correction for the project.

Michael Johns (University of Portsmouth, UK) is a transportation GIS specialist who had responsibility for development and enhancement of GIS and database resources relating to the Eastern Trunk Line Railroads for use in web-based visualisations

Nathan B. Sanderson (University of Nebraska-Lincoln, US) is a Ph.D. candidate in American History at the University of Nebraska-Lincoln who contributed subject and project management expertise to the Railroads and the Making of Modern America Project based at the University of Nebraska.

Other participants and advisors
Anne Bretagnolle (Paris One University, France)
Ian Gregory (University of Lancaster, UK)
Anne Kelly Knowles (Middlebury College, US)
John Lutz (University of Victoria, Canada)
Sherry Olson (McGill University, Canada)
Ashok Samal (University of Nebraska-Lincoln, US)
Martin Schaefer (University of Portsmouth, UK)
Stephen Scott (University of Nebraska-Lincoln, US)

Emma White (University of Portsmouth)
Richard White (Stanford University, US)
Eli Katz (Stanford University, US)
Danny Towns (Stanford University, US)
Kathy Harris (Stanford University, US)