Digging into Data: Is the Future Now?
Alex H. Poole
School of Information and Library Science at the University of North Carolina at Chapel Hill
Archival Education and Research Institute
12 July 2012 University of California-Los Angeles

Funders:
1) Office of Digital Humanities, National Endowment for the Humanities (NEH-OHH)
2) National Science Foundation
3) Joint Information Systems Committee (JISC)
4) Canadian Social Sciences and Humanities Research Council (SSHRC)

Chronology
NOV 2007: Gestation
JAN 2009: Digging into Data Challenge announced by four funding agencies
DEC 2009: First round awardees announced (8 of 90 applicants)
MAR 2011: Eight funding agencies announce second round of Digging into Data Challenge
JUN 2011: Digging into Data inaugural conference held
DEC 2011: Second round awardees announced

The Projects
1) Digging into Image Data to Answer Authorship Related Questions
2) Digging into the Enlightenment: Mapping the Republic of Letters
3) Mining a Year of Speech
4) Structural Analysis of Large Amounts of Musical Information
5) Harvesting Speech Datasets for Linguistic Research on the Web
6) Using Zotero and TAPOR on the Old Bailey Proceedings: Data Mining with Criminal Intent
7) Towards Dynamic Variorum Editions
9) Data, Data, Data
10) Spatio-Temporal Correlation, Analysis and Visualization

Interdisciplinarity
History (4); Computer Science (3); Computing and Information Science (2); French (2); Linguistics (2)
Academic Technology; Art; Art History; Artificial Intelligence; Assessment; Classics; Communication Studies and Multimedia; Computational Linguistics; Computational Musicology; Computer Science and Engineering; Cultural History; English Literature; Environmental Literature; Geography; History of Cartography, Humanities Computing; Library and Information Science; Library Science; Museum Science; Music Technology; Philology; Philosophy; Phonetics; Speech and Language Processing

Methods of Analysis
Text mining and multiple types of visualization; geographic analysis; text mining; visual analytics; Optical Character Recognition; morphological analysis; forced alignment of transcription to audio; acoustic extraction; machine learning classification; computer-aided analysis of musical structures; adaptive image analysis; machine learning; hand-corrected data and metadata on numerous topics made available for geospatial and temporal exploration in web-based "apps" and applications

The White Papers
• "With so much virtual content to examine and with such subtle nuances to isolate and compare, the ability to cluster materials that appear to contain similar characteristics is proving to be of great benefit." (Ainsworth, Bajes ey & Rehberger)
• "[V]ery large scale is essential." (Coleman et al.)
• "Humanistic inquiry…is freeform, fluid, and exploratory, not easily translatable into a computationally reproducible set of actions." (Ede dstein and Findlen)
• "All of human life is here." (Cohen et al.)
• "The ultimate goals for our work—and arguably for all work in the humanities—is to advance the intellectual life of humanity as broadly as possible." (Almas et al.)
• "The approach adopted here is to accept neither the assumption that a complex reality can be compressed into a very small number of categories, nor that it is so infinitely complex that any attempt to standardize data or search for regularities is fruitless. Once again, a critical middle ground is a more appropriate stance in our view." (Thomas et al)

Commonalities
1) Engage with data corpora much larger than what could be read, seen, heard, or experienced by one individual
2) Apply some form of computational analysis (tool, application, or algorithm) to the corpora
3) Require continual refinements to tools and data; thus necessitates collaboration and coordination of multiple project participants with variegated backgrounds and skill sets
4) Elected a common research process: hypothesis; selection; exploration; querying and correcting, modifying, and amending; observing; drawing conclusions

Differences
1) Differing disciplinary traditions
2) Choice of collaborators most suitable for the media, scale, and organization of data sets
3) Proportion of manual to automated work
4) Need for continual adaptation of analytical tools
5) Likelihood of attaining major outcomes in fifteen months

Long Story Short...
"This much is clear: ‘big data’ are not just for scientists anymore." (Willford and Henry 2012)

Future Goals
1) Expand notion of "research"
2) Expand concept of "research data" and embrace the challenge of scale
3) Continue to foster interdisciplinarity
4) Further embrace collaboration in across disciplines, departments, and institutions
5) Develop models for sharing credit
6) Develop models for inter-institutional resource sharing
7) Expand our notion of scholarly publication
8) Invest more heavily in human infrastructure and cyberinfrastructure
9) Address major gaps in training programs

Commonalities
1) Engage with data corpora much larger than what could be read, seen, heard, or experienced by one individual
2) Apply some form of computational analysis (tool, application, or algorithm) to the corpora
3) Require continual refinements to tools and data; thus necessitates collaboration and coordination of multiple project participants with variegated backgrounds and skill sets
4) Elected a common research process: hypothesis; selection; exploration; querying and correcting, modifying, and amending; observing; drawing conclusions

Differences
1) Differing disciplinary traditions
2) Choice of collaborators most suitable for the media, scale, and organization of data sets
3) Proportion of manual to automated work
4) Need for continual adaptation of analytical tools
5) Likelihood of attaining major outcomes in fifteen months

Long Story Short...
"This much is clear: ‘big data’ are not just for scientists anymore." (Willford and Henry 2012)

Future Goals
1) Expand notion of "research"
2) Expand concept of "research data" and embrace the challenge of scale
3) Continue to foster interdisciplinarity
4) Further embrace collaboration in across disciplines, departments, and institutions
5) Develop models for sharing credit
6) Develop models for inter-institutional resource sharing
7) Expand our notion of scholarly publication
8) Invest more heavily in human infrastructure and cyberinfrastructure
9) Address major gaps in training programs

A Last Word
"Go forth and do great things…" (Borgman 2009)