First Monday, Volume 16, Number 5 - 2 May 2011

f i ® s t m ¤ ñ d @ ¥
PEER-REVIEWED JOURNAL ON THE INTERNET

Digging into data using new collaborative infrastructures supporting humanities-based computer science research

by Michael Simeone, Jennifer Guiliano, Rob Kooper, and Peter Bajcsy

Abstract

This paper explores infrastructure supporting humanities–computer science research in large–scale image data by asking: Why is collaboration a requirement for work within digital humanities projects? What is required for fruitful interdisciplinary collaboration? What are the technical and intellectual approaches to constructing such an infrastructure? What are the challenges associated with digital humanities collaborative work? We reveal that digital humanities collaboration requires the creation and deployment of tools for sharing that function to improve collaboration involving large–scale data repository analysis among multiple sites, academic disciplines, and participants through data sharing, software sharing, and knowledge sharing practices.

Contents

1. Introduction

At the core of humanistic research is the search for primary source information. Whether these sources include manuscripts, material culture objects, maps, or other forms like video and digital media, successfully navigating corpora of material has long been the substance of valuable research contributions. While scholars have traditionally worked by personally extracting data from archives and forming their own collections of information surrounding thematic or historical periods, the advent of the digital archive has posed its own unique challenges. Although they are promising in the resources they make available as virtual content, digital image archives are also potentially daunting. Digital archives can house so much data that it becomes impossible for scholars to evaluate the archive manually, and organizing such data becomes a paramount challenge. While individual scholars often commit the hours of labor needed to visually examine hundreds of images, the enormous size of some archives precludes the ability of an individual or even a team of individuals to examine each data element. Additionally, the kinds of questions that can be asked and answered are constrained when the

research data is not well organized and synthesized.

With computational tools, digital archives can reveal more than they obscure by providing organizational frameworks and tools for analysis. However, these tools — in the guise of metadata organization, indexing, searching, and analytics — are not self–generated. They require the combined work of humanists with their interdisciplinary questions and computer scientists with their disciplinary approaches to partner with one another to produce viable research methodologies and pedagogies. As interdisciplinary collaborations are becoming more common, aligning the interests of computer scientists and humanities scholars requires the formulation of a collaborative infrastructure for research where the approaches, methodologies, pedagogies, and intellectual innovations merge. While the concept of shared resources in a "cloud" is gaining popularity for office and document–based collaborations, we maintain that constructing a "cloud" of collective resources for use in researching large image archives across multiple disciplines and institutions requires a specific design that accommodates the communicative demands of multi–disciplinary academic research, the intricacies of intellectual property and publication, the needs of software developers working from remote sites, and the difficulties of serving large amounts of data for collective examination. In other words, the model we have developed tightly integrates practices and technologies for data sharing, software sharing, and knowledge sharing/communication.

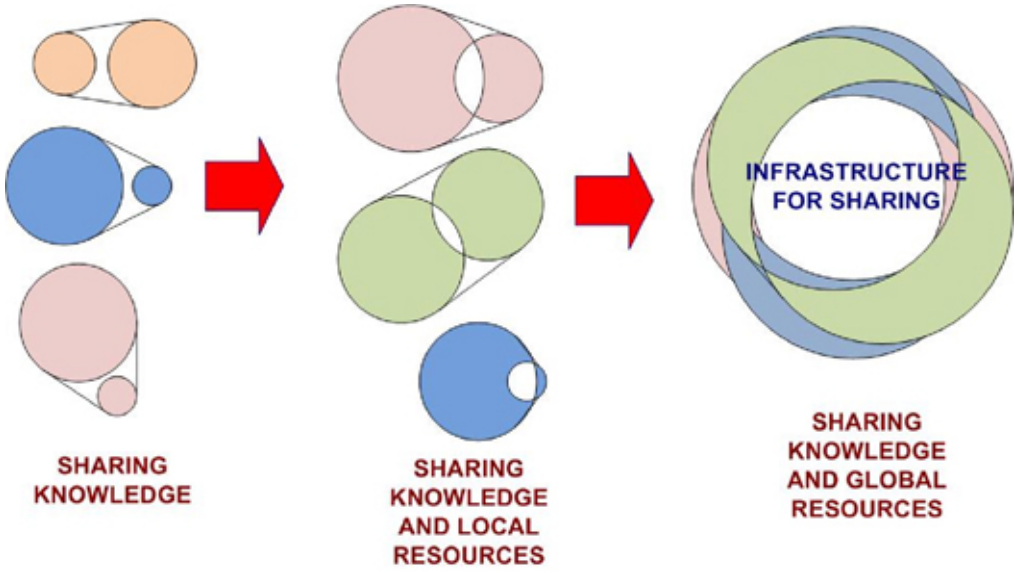

Figure 1: Paradigms for computer science and humanities collaborations. Circles of different sizes represent research and resources of each collaborator with lines connecting them to represent project collaboration. The scenario on the left indicates the most basic kind of sharing, the scenario in the middle shows a series of partnerships with some shared resources, while the scenario on the right demonstrates a shared and generative infrastructure for circulating data, knowledge, and hardware among multiple collaborators and institutions.

This article explores the construction of one such infrastructure for sharing in digital humanities and computer science collaborations. We collectively ask: Why is collaboration a requirement for work within digital humanities projects? What is required for fruitful interdisciplinary collaboration? What are the technical and intellectual approaches to constructing such an infrastructure? What are the challenges associated with digital humanities collaborative work? In doing so, we reveal that digital humanities collaboration requires the creation and deployment of tools for sharing that function to improve collaboration involving large–scale data repository analysis among multiple sites, academic disciplines, and participants through data sharing, software sharing, and knowledge sharing practices. Increasing the number of participants from diverse disciplinary fields makes the logistics for that collaboration of crucial importance. The suite of tools developed and employed under the Digging into Image Data to Answer Authorship–Related Questions (DID–ARQ) Project to manage this communication and

collaboration and help reduce the start–up and time and costs is instructive for future large–scale, multi–institutional projects. It is our goal to not only demonstrate the challenges and solutions that emerged throughout our work, but also to provide a template for future collaborations that involve multiple datasets in geographically distributed locations.

As a case study, we use our work in DID–ARQ awarded as part of the Digging into Data Challenge Competition (DID Challenge Web site). Funded by the National Science Foundation (NSF) and National Endowment for the Humanities (NEH) from the United States, the Joint Information Systems Committee (JISC) from the United Kingdom and the Social Sciences and Humanities Research Council (SSHRC) from Canada, DID-ARQ consists of an international, multi–disciplinary team of researchers from the University of Illinois (U.S.), National Center for Supercomputing Applications (U.S.), Michigan State University (U.S.), and the University of Sheffield (U.K.). The DID–ARQ team is conducting work to formulate and address the problem of identifying the salient characteristics of artists from two–dimensional (2D) images of historical artifacts. Given a set of 2D images of historical artifacts with identified and unidentified authors, our project teams aim to discover what salient characteristics make an artist different from others, and then to enable statistical learning about individual and collective authorship.

Our collaboration infrastructure produces an innovative research environment because of its ability to integrate multiple collaborators (knowledge), content types (data), and research tools (software and hardware) across multiple locations. Figure 2 shows an example of an infrastructure for sharing that was deployed for our DID–ARQ Project. The three sites involved used a cloud of technologies for sharing images, software, hardware, and knowledge with the sharing policies defined by the memorandum of understanding. Because we employed these enhanced tools for sharing, our collaboration has produced a uniquely comprehensive exchange between computing and humanities stakeholders, one that lets individual researchers or teams of researchers dynamically form their research goals the better to cooperate with other researchers within the Project. Indeed, these tools have enabled, archived, and published every phase of this multi–campus, multi–archive, and multi–disciplinary endeavor. This infrastructure allows us to add additional sites in the future that can connect themselves to this cloud, become familiar with previous collaboration work, see the data currently available, and quickly get integrated into the whole project. Providing this structure allows new partners quickly to start adding to the cloud of technologies thereby increasing the speed of deployment.

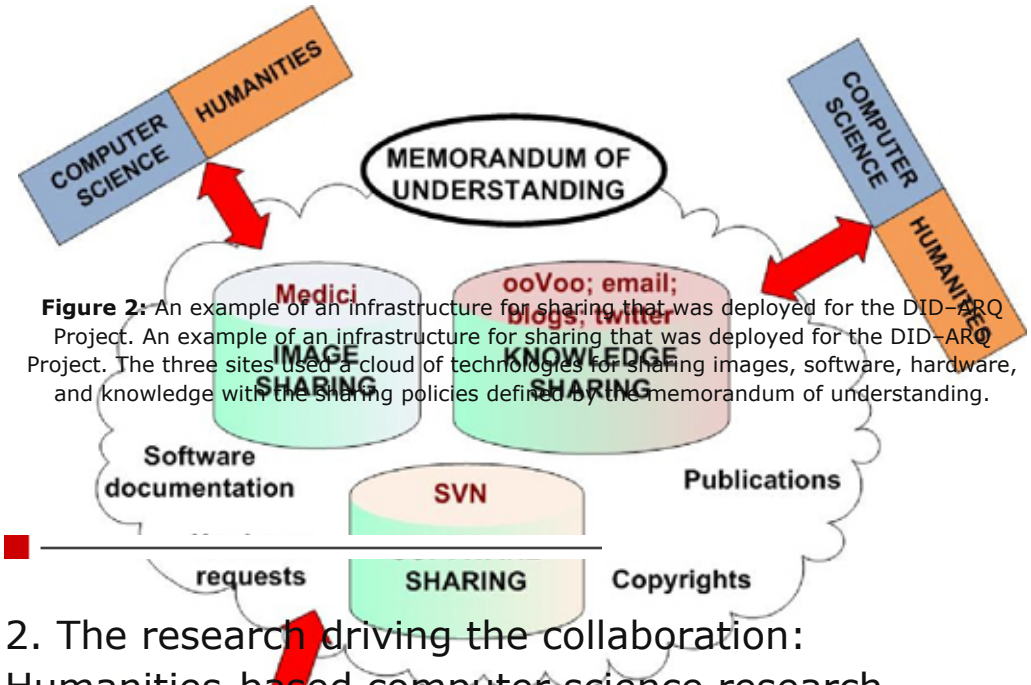

Figure 2: An example of an infrastructure for sharing that was deployed for the DID–ARQ Project. An example of an infrastructure for sharing that was deployed for the DID–ARQ Project. The three sites used a cloud of technologies for sharing images, software, hardware, and knowledge with the sharing policies defined by the memorandum of understanding.

2. The research driving the collaboration: Humanities-based computer science research

Each of the individual projects of the DID–ARQ collaboration starts with a common research question: how does one examine an archive of images when that archive is too large to examine manually? In framing that question through collaborations with geographically remote scholars, however, a second important question arises: how to share these archives and communicate their salient characteristics with research partners? Historical maps of the Great Lakes region, The Quilt Index, and Froissart Manuscripts all represent hundreds to thousands of digitized images, and every digitized item, no matter its historical context or intended purpose, is a potentially rich store of information for scholars of history, literature, and culture. Because of the sheer volume of labor required for research across the archives in question, "when" the researcher might complete his or her study of the entire corpus becomes a question of "if" that research is even possible. In this section, we will briefly explore the humanities questions from each of our humanities partners and then examine how remote collaborations have affected the questions being posed.

2.1. Pattern driven classification of quilts

The Quilt Index is a growing research and reference tool designed to provide unprecedented access to information and images about quilts held in private and public hands. A joint project of The Alliance for American Quilts, MATRIX: Center for Humane Arts, Letters and Social Sciences Online at Michigan State University and the Michigan State University Museum, the Quilt Index provides an opportunity for in–depth exploration of quilts through the lens of thousands of digital quilt images and text drawn from public and private collections. Two main issues, the sheer volume of data and variation in the text documentation, limit the capacity for any individual to examine the entire corpus of the Quilt Index thoroughly. The variation in the quality and format of the documentations lends itself to challenges of organization. While each quilt has a set of descriptive fields that contain necessary information and controlled vocabularies specific to quilt research and study, for each quilt the original textual documentation must be matched to particular descriptive fields. While differences in existing text strings can be accounted for through the search programming, entailing, for example, algorithms for effecting the simplest reconciliation of spacing and capitalization, or more complicated cross–references for known spelling variation and alternate published pattern names [1], cross–reference lexicons are not effective to account fully for unique names or to cover every local variation. Moreover, in some cases key information may be unknown or even intentionally left out during the documentation process, such as a quiltmaker's affiliation to an organization like the Ku Klux Klan. These challenges when parsing the text documentation create intensive labor for processing quilts. The challenge as Gregory Crane famously asked about books, is what can a humanities researcher do with a million quilts? (Crane, 2006) In the first place, the physical collections of quilts are geographically distributed, which rules out most side–by–side comparison. Moreover, even if

institutions allowed physical access to the collections, a scholar could not possibly unfold and physically look at more than a few quilts at a time and then only at an oblique angle, as most quilts are the size of double beds. Even in a digital environment designed for users to frontally see and compare images, close analysis of every digital quilt image is time–prohibitive when the number of digital images reaches into the tens of thousands.

2.2. Knowledge of cartographers of early French and British maps

Early mapping of the North American continent (1640–1820) generated sizable archive of ornate atlases and sheet maps. Studying seventeenth and eighteenth century maps in the context of modern understandings of cartography reveals distortions that were the inevitable results of comparatively primitive technologies, difficulties of travel by sea, rivers, and overland, and inaccurate and conflicting data from explorers and traders. Early mapmakers relied on various methods of recording distance that recorded other information as well, as techniques ranged from dropping a log overboard and watching its movement, to examining the eclipses of Jupiter's moons. Changes in coastal currents, however, could warp the apparent distance traveled by passing ships, cloud cover could block out astronomical reference point, and geographical features like bays, rivers, and mountain passes were both consciously and unconsciously exaggerated to appear more appealing to shipping and commercial interests. Furthermore, European sailors could not calculate longitude until the later part of the eighteenth century, and before then, climate affected primitive timepieces and thus any measure of east–west distance.

Yet if plotting the regions around the Great Lakes without error was impossible until the later development of more accurate compasses, timepieces, and recording equipment, early maps still contain a vast amount of valuable data. There is a rich history to what seem now to be mere mistakes and miscalculations — a history that discloses both the development of technologies for measuring distances and the ways in which French and English mapmakers interpreted the geographical data that they had at their disposal.

In the study of maps, evaluating each map's depiction of geographic regions and waterways, from their relative size to their purported location on the Earth, increases in difficulty as the maps become more intricate and detailed. Given that there are hundreds of available maps of North America produced by European and Native mapmakers, hand examination, though useful in limited qualitative studies, makes for a nearly prohibitive workload if scholars wish to pose any research questions across not only 40 maps at a time, but larger portions of the archive as well. This archive contains valuable data needed to address fundamental research problems in colonial history, the history of science and technology, historical climatology, and the history of print culture.

2.3. Artistic and scribal hands contributing to manuscripts

The Froissart Manuscripts corpus consists of more than 6,000 high–resolution image files captured photographically from 12 manuscript volumes (~2TB of data). These manuscripts were all produced between ca. 1408 and ca. 1418 under the supervision of Parisian bookseller Pierre de Liffol, who seems to have glimpsed a market opportunity for the production of luxury copies of the Middle French Chronicles of John Froissart (ca. 1337–ca. 1404), now a key source for the study of the Hundred Years' War between France and England. De Liffol produced these volumes for clients in the service of king Charles VI of France, though the illustrations to at least one of them testify to a client with pro–English sympathies. The text of the Chronicles was copied from exemplars by at least two teams of scribes. The miniatures were entrusted to two artists' workshops: those of the Giac and Boethius Masters. Secondary decoration (the initial letters and borders) may have been painted by still other artists. Once all of the copying and illustrating was done, the quires were reassembled, sewn together and provided with oaken boards covered with leather or velvet.

The virtual manuscripts allow the scholar to consult the content at their leisure and manually compare across different collections. Furthermore, the digital surrogates offer the opportunity to perform more complex analytical comparisons across the expanding collection than a human is capable of undertaking. By considering the detail and aspect of each of the manuscript folios, similar scribal hands and artistic objects can be clustered together for scholarly comment and thus help construct a pattern of work across them with the aim of determining authorship and work patterns. The algorithms will also help scholars to adduce more discriminating features than hitherto arrived at for characterising the distinctive handiwork of the Giac and Boethius Masters as well as reveal information about the hands responsible for the primary and secondary

decoration of these manuscripts. More specifically, can we discern the presence behind the labels 'Giac Master' and 'Boethius Master' of more than one individual (*e.g.*, a group of artists working in a common style)?

2.4. Need for a new collaborative infrastructure

Scholars in the humanities, including those above, are faced with a set of specific research questions that pertain to their individual disciplines and an archive that tests the limitations of their disciplinary methodology in scope and size. For each humanities project, the questions of managing and sharing data across these repositories become a key challenge. Each project has its own metadata, organizational structure, and image repository. Although each project focuses on its particular issue, yet, the need to explore authorship unites them all. Furthermore, it is not only the authorship question that unites the smaller projects but also the software that can be reused. For example, as illustrated in Figure 3, the same software algorithm has been used to find objects of interest in Froissart's manuscripts (*e.g.*, armor, faces) and lakes in historical maps. Thus, sharing data and software with hardware is critical for resource limited research in humanities.

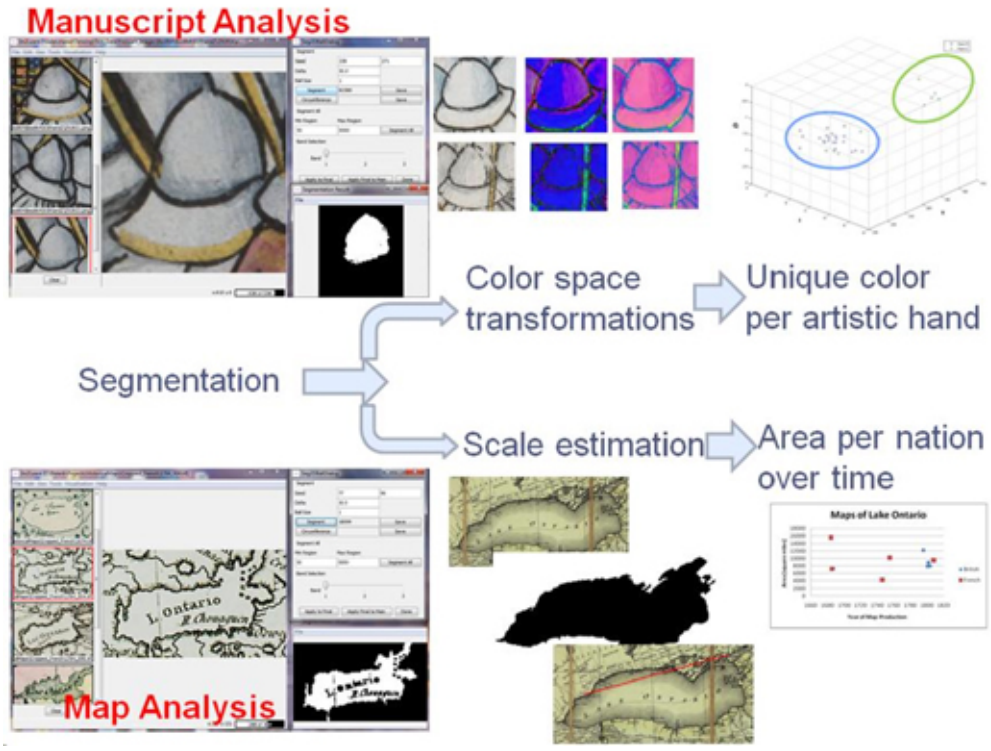

Figure 3: Illustration of the need to share software that can perform model–based segmentation of images. The same segmentation algorithm is applied to finding armor in Froissart's manuscripts and lakes in historical maps.

The challenges faced by both the individual projects and the overall DID–ARQ group for collaborating among those individual projects necessitated a new collaborative infrastructure. Nonetheless, the problem of designing such a resource poses several questions. With such large–scale archives, how is it possible to share information effectively and purposefully with project partners? If hundreds of images press the capabilities of manual examination, how are other contributors able to craft interpretations of the respective corpuses? How can a computational scientist design a solution to an image–processing problem for a group of images

that is this vast and also geographically remote? As we initiated our collaboratory, these questions began to structure our enterprise: how could scholars at three different sites working with collections of visually disparate materials and with three disparate methods of image storage and management work together to reveal salient characteristics making one artist different from another regardless of the source material?

3. The logisitics of cooperation: Collaborative infrastructure to address specific challenges of collaboration

As outlined above, in formulating a collaborative humanistic project, we had to recognize the individual questions and the methodological and theoretical approaches of each humanities scholar. Yet, a single humanist exploring an image archives on their own is not the primary audience for this project. Rather, it is the intersection of teams of scholars all unified in their technical challenges of collaboration that drives this project forward. In this section, we outline how financial, legal, and research considerations resulted in the construction of a collaborative infrastructure that not only facilitated the interchange between scholars and scientists but also enriched the very nature of information exchange. With challenges ranging from country–specific and institution–specific policies for collaborations to technical problems related to sharing data, software and computational resources, and building communication channels for knowledge sharing and project coordination, we begin by noting an underlying issue with the framing of all Digging into Data projects, including the DID–AQR Project.

3.1. DID institutions and funding

Announced in January of 2009, the Digging into Data grant competition was sponsored by the NEH, NSF, JISC, and SSRC. In releasing the request for proposals (RFP), these organizations created one main RFP with addenda for each of the sponsoring organizations. Each institution within a DID team was required to field an international partnership involving at least two of the three participating countries: Canada, England or Wales in the United Kingdom, and the United States. While our complete team would create a single application that would be sent to the agencies, each individual institution was required to create their own budget and follow the rules of their assigned funder. In the case of DID–ARQ, UIUC/NCSA applied to NSF, Michigan State to NEH, and the University of Sheffield to JISC with each team completing its own budget and sections of materials with consultations from their targeted partner institutions. A single named primary investigator was charged with completing the application and submission process, yet each institution needed its own primary investigator to complete their local grant process and deal with their assigned sponsoring agency. In effect, although the project intent was to promote international collaboration through collaborative research, the application structure immediately necessitated sub–division through the required one–to–one relationship with each sponsoring agency and the adherence to the grant submission and management systems of each applicant organization.

This format extended throughout the funding mechanisms associated with the Project. Each institution created its own budget and, when awarded, received its own contract from its assigned sponsoring agency with no mechanism for sharing funds or for collaborative oversight. DID awardees had to design their own mechanisms to leverage geographically distributed expertise and resources, as well as independent sources of funding. The DID awardees did not need to outline an in–depth plan for how they would collaborate, just that they would collaborate internationally and in some discrete way. This underlying assumption of having all collaboration infrastructure and methodology in place at all sites resulted in proposals that largely ignored the challenges, approaches, methods and technologies needed to address a distributed collaboration, as this component of the proposal was not critical to obtain the funding. As a consequence, our awarded DID–ARQ Project immediately faced challenges of how to structure and fund the expense of the collaborative activities that were not institutionally based. This included start–up costs associated with the infrastructure technologies for sharing, and with the preparation of agreements governing the collaboration. For example, in our proposal for DID–ARQ, we noted that we would utilize AccessGrid, a peer–to–peer technology

for large–format information sharing and videoconferencing. However, the costs for AccessGrid (US$136 per hour) were not figured into the initial budgets. With two and a half to three hours of technical meetings a month, costs for AccessGrid would exceed US$6,120 for the Project. With the lack of budgeting, these costs would have reduced available funds for other portions of the Project. Due to the lack of a formal mechanism for parsing start–up costs between teams or a shared pool of funds, these types of costs would have been a burden on one institution rather than on the team as a whole. As a result, the assigning of start–up costs became a *de facto* decision of convenience with the institution shepherding particular aspects of the collaboration bearing those costs.

3.2. Legal and ethical aspects of scholarly collaborations

A noteworthy issue in initiating the DID–ARQ Project involved navigating various policies specific to countries and institutes that apply to collaborations and sharing. Before project work could begin, the DID–ARQ team had to negotiate the legal aspects of collaborations regarding the conduct of the Project, publicity, publications, copyright, intellectual properties and liabilities. In accommodating U.K. and U.S. government–specific policies, as well as the institutional policies of the University of Sheffield, MSU and UIUC, it was readily apparent that formal agreements would have to be set in place. In part, this was necessitated by team concerns about project liability associated with data access. Images associated with the projects could be lifted from project Web sites and publications and used in violation of copyright laws. As such, while work on the project began while the agreement was being negotiated, all due consideration was given to limiting access to data and images by the general public. Initially drafted by the University of Sheffield, and reviewed by MSU and UIUC Grants and Contracts offices, the memorandum of understanding associated with DID–ARQ documents how all project principal investigators, their graduate students, postdoctoral scholars, and faculty staff will follow a prescribed set of behaviors in documenting the work of the Project, the publications associated with Project results, and intellectual property contributed by project participants.

One important issue for the institutions involved was ascribing ownership of images to their respective archives. The DID–ARQ team discussed extensively the rules for publishing materials with the images that would be shared across all sites for the software development and testing purposes. The digitization of Froissart's manuscripts was paid for by several private donors and by sponsors of prior projects in the U.K. [2]. The acquisition of quilt photographs was subsidized by MSU and by funding from a variety of public and private sources [3] just as the digitization of historical maps from UIUC was sponsored by the University of Illinois Libraries. The data stewards on the DID project had to decide how to select images for sharing and how to formulate the acknowledgement if an image was included and therefore referenced in a publication or presentation. As a consequence, the collaborative technology for image sharing had to provide a mechanism for entering author, copyright information, credit, and acknowledgement information that would stay affiliated with an image including any derived images or datasets. Agreement on the legal aspects of this collaborative project would have been much more difficult if the partner institutions had not all been educational ones. While working with private collectors, companies or philanthropists might bring new images to the research arena, the objectives of partners and the legal implications of different objectives might require devoting more time to this legal aspect than was the case with our effort. Construction, negotiation, and signing of the DID–ARQ MOU was completed in roughly three months as each institution was able to utilize its local legal and contract staff in creating the agreement. Significantly, institutions should plan at least four–five months to take care of the legal aspects of collaborations particularly when dealing with non–educational partners.

We organized the legalities of this effort, in part, by our observation that the grantors impose different criteria on measuring success of a project with respect to the expected project outcomes, and reporting methods. The Joint Information Systems Committee (JISC) outlined an expectation for more applied research than did the National Science Foundation (NSF). Given the unwritten rules about sponsors' evaluations — evaluations driven by the reviewers coming from the communities being supported by these sponsors — we ascertained that the National Endowment for the Humanities (NEH) desired an emphasis on academic monographs about discoveries of new facts from historical content while NSF would rank higher journal papers about the computational and statistical methods needed to obtain those discoveries. The differing interests of each funding agency underscore not only the difficulty of bringing together multiple funding agencies in providing funding but also the difficulty in meeting the expectations by the proposers and awardees. While individual team members clearly understood these differences, they had to be taken into account when legal agreements were drafted and the

division of work outlined. Graduate students focused more on basic research (*e.g.*, how to design computer algorithms to automate authorship discovery) while full–time staff pursued applied research (*e.g.*, how to support tagging and commenting when images were shared across sites). Expanding to publications, the team agreed that computer science journal papers can use images to demonstrate the algorithmic performance and computational scalability with appropriate acknowledgments, while books written by humanists about the interpretation of historical content can use computer–generated results and references to the images used with appropriate acknowledgments.

Another important question is that of authorship of articles, papers, journals, and books written by members of the DID–ARQ Project team. In an academic climate where "publish or perish" is still a measure of success for both humanists and scientists, determining the intellectual ownership of research publications has challenged our own parameters of authorship. The awarded DID–ARQ proposal included 18 participants, and the current project involves an additional four students and three full–time engineers. Multiple authors is more common in the sciences than in the humanities, but the number of individuals contributing to this project is unusually large. A question arises about publishing an article with 25 co–authors or a subset of co–authors, or in other words, about individual and collective contributions. While we do not know a fair answer to this question, a policy about co–authorship has to be discussed among the team members. The DID–ARQ Project team proposed guidelines that acknowledged that significant intellectual contributions merited authorship and co–authorship credit; the contribution of content materials related to the individual projects drawn from the original proposal project materials without additional intellectual investment in the resultant publication did not merit co–authorship. In effect, each scholar and scientist could receive acknowledgement of their content contributions (generally framing questions or summaries of work undertaken), but this acknowledgment stands apart from authorship credit. Authorship credit comes only with significant investment in the writing and constructing of interpretations and arguments in new publications.

Generally, while a member or members of the team initiate(s) a publication or public presentation, DID–ARQ relies on notification of the initiation of a resultant publication/presentation and the sharing of all works in progress in order to give all members of the team the opportunity to contribute. In this manner, DID–ARQ ensures an equal–work, equal–credit scenario where all members of the team, be they graduate students, staff or PI, can take ownership of the publication. This structural element ensures that all legalities of intellectual property and publication are met while still recognizing the need for individual autonomy by the teams making up the full collaboration. A variety of public materials have been generated through this apparatus. There has been a paper authored by a subset of team members on preliminary materials that became part of the DID–ARQ partnership, three presentations on two different continents that were authored singly but include DID–ARQ acknowledgements, and two publications in process that include acknowledgements, co–authors, and authors of differing degrees.

3.3. Data sharing

Given the challenges posed to our research and collaboration by the kind and scale of our image data, it was crucial for our DID–ARQ team to establish a common means to collect, share, annotate, and examine large amounts of image data. To address these challenges, we utilized a previously developed prototype of a multimedia content repository called Medici [4]. Medici is unique as a shared project workspace: it can accommodate the upload of all file formats in a common environment with evolving rendering and capabilities based on an expanding list of project requirements.

From a number of possible choices for applications supporting cloud computing and file sharing, the DID–ARQ team selected Medici because it allowed for distributed access and management of large-scale private image collections. Commercial systems such as Flickr (Flickr, 2010), Panoramio (Panoramio, 2010), or TinyPic® (TinyPic®, 2010) were not suitable because they contain no ways to protect copyrighted content from copying or download, a crucial feature given the constraints imposed by each individual image archive. In comparison with open source distributed data management solutions such as Storage Resource Broker (SRB) or iRODS (iRODS, 2010), Medici provided extensive visualization capabilities for viewing multimedia content including very large size images which have not been available in iRODS but have been critical for working with DID–ARQ images. Finally, the social networking features present in many commercial solutions, such as personal tagging and commenting, were very helpful during

data sharing but unavailable in the open source solutions. In the future, we foresee the need to add image–based search (see TinEye, 2010) in addition to text–based search.

In Medici's Web–based file browser, research papers and images appear side by side as thumbnails, and custom collections showcase images and documents that elucidate specific research problems. Furthermore, Medici lets its users examine shared files in their full resolution without downloading them. Built into the file viewing system is a preview window that uses Seadragon, a library built by Microsoft Labs that lets users smoothly zoom in and out when viewing large image files.

Thus, where sharing gigabytes of data was an obstacle on other platforms, it is now through Medici that there is an opportunity to provide a rich multimedia experience for collaborators. For example, large amounts of quilt images, instead of requiring individual project participants to commit significant time and resources to downloading and viewing the full–sized images, can be examined now in their full resolution through a Web browser. Unlike existing systems, including Flickr and YouTube, with some of the same functionalities as Medici, one of the draws of the Medici system is the ability to create a private data collection. The data that is stored in the system is only accessible by those people who have been given user access to the system and once the project is finished the data can be safely removed without a trace. For cross–university projects such as DID–ARQ, the ability to control access to data and the ability to remove it at the end is very important so that all legal requirements are fully met.

This ability to share image data at full resolution is crucial for our computational scientists and humanities scholars. Because these are photographs of physical artifacts like maps, quilts, and manuscripts, their study requires as much detail about their materiality as possible to be made available. The content is *digitized* not digital, and thus research on the digitized archives is not possible without maintaining fidelity as data moves from artifact to digital image to remote research location. For instance, scientists at Illinois cannot evaluate how to adapt the shape segmentation algorithms they have developed for the study of historical maps to help with the study of medieval manuscripts without a detailed replica of the object they will be analyzing, complete with the details, textures, and imperfections of the original. All collaborators were able to upload, annotate, and examine their own collections of images as developed, hosted, and deployed by NCSA/UIUC, and to do the same with collections built by other participants.

To use Medici, scholars upload files without having to specify much more than the file itself by using the standard file browser interface, or by clicking and dragging folders into a drop box. Single files are added to the default file collection where folders become a new collection of files based on the folder name. Once users add a file to a Medici repository, the system extracts basic file information such as the filename that becomes the title by default. The system assigns a unique URI, and it executes extraction or analysis services automatically. Afterward, the user can include any statement about the data, such as adding a tag, leaving a comment, or adding a specific metadata field to help describe or categorize the digitized artifacts, such as publisher, mapping organization, or region of production. Ultimately, then, Medici furnishes collaborators not only with the means to share and easily access large amounts of image data, but also to showcase proposed logics of organizing and approaching the archives in question.

3.4. Software sharing

In addition to data sharing, the DID–ARQ team had to share pre–existing software and manage concurrent software contributions from the three institutions. The pre–existing software included not only the Content Management Repository called Medici but also the Im2Learn library of basic image processing and visualization algorithms that can be applied to various image analyses, the Versus (Versus, 2010) library for content–based image comparison, the Cyberintegrator (Cyberintegrator, 2010) workflow for managing computations on distributed computational resources. The analytical capabilities come from the Im2learn library that provides a plug–and–play interface for adding new algorithms and tools. Due to the fact that the authorship questions are frequently based on a comparison operation, we have designed additional application programming interfaces (API) called Versus which allows everyone to contribute with comparison methods. Once the algorithms for image analyses and comparisons have been developed, they can be integrated into workflows (a sequence of algorithmic operations to reach the analytical goal) in Cyberintegrator workflow environment. Cyberintegrator is a user friendly editor to several middleware software components that enable users to (1) easily include tools and data sets into a software/data unifying environment, (2) annotate data, tools and workflows with metadata, (3) visualize data and metadata, (4) share data and tools using local and remote context repository, (5) execute step–by–step workflows during scientific explorations, and, (6)

gather provenance information about tool executions and data creations. Unlike other existing open source workflows used in eScience, such as Taverna (Taverna, 2010) or Kepler (Kepler, 2010), Cyberintegrator has a reconfigurable user interface by using the Eclipse Rich Client Platform (RCP), and captures all provenance information automatically using semantic representation (Resource Descriptor Framework representation). These characteristics allow a user to customize Cyberintegrator's interface to his/her preference and to mine the provenance information graph to reproduce trial–and–error research results or to gain knowledge about the use of data sets and software.

We have managed concurrent software contributions from the three institutions by setting up a subversion source control system [5]. This system allows for software sharing, remote access, software merging, version tracking and permission management.

3.5. Hardware sharing

The computational power of a single desktop has risen dramatically in the past few years. It is not uncommon to have a machine as desktop that has four computational units, each of which is many times more powerful than those found in desktop computers previously. The same growth seen in computational power can be seen in storage. Hard disk storage is now measured in terabytes instead of gigabytes. This growth in both computational power as well as storage has led scientists to tackle larger and larger problems in common computing environments. Still, running many processing steps (computational algorithms) on the same data set can take a few days. Similarly, running one fast processing step on a very large collection of data sets will also make the total time to process a few days or even months. Having access to large amounts of shared data and running the processing steps on the data will only make the computational time expand. Scientists purchasing modern desktops today expect to obtain the computational results faster than ever before. To match the expectations of scientists with the computational requirements, one needs not only high–performance computing hardware but also the software that can run in parallel and take advantage of the hardware.

High–performance computing hardware is becoming more mainstream in universities and companies. A cluster can be created using off–the–shelf personal computers that are connected using free software such a Linux and Eucalyptus, or by using solutions from commercial software vendors including Microsoft. If a cluster is not immediately available, the computation can be outsourced to places such as Amazon. Companies like Amazon sell computational resources to users for short periods, enabling them to process large amounts of data quickly. For the DID–ARQ Project, we have the advantage of having access to the high–performance computers at the National Center for Supercomputing Applications (NCSA). In addition, NCSA is currently in the process of creating the fastest public computer, while both MSU's High Performance Computing Center and Sheffield have smaller clusters that can be used as well. If any of the processing is projected to take a large amount of time, NCSA has offered to run the algorithms on its clusters assuming that the algorithms have been parallelized.

3.6. Knowledge sharing and project coordination

With partners geographically distributed across three different time zones, verbal communication posed the first major challenge for knowledge sharing and project coordination. In our 18–month long collaborative, the University of Sheffield is six hours ahead of our western most partner, the University of Illinois (except for a few weeks when daylight savings did not match up). With bi–monthly teleconferences, finding a time that took this time difference into account and would accommodate all the partner institutions was paramount. In effect, Illinois was just beginning its workday as Sheffield was closing down its efforts. Further, many of the DID–ARQ team members were "targets in motion" simultaneously juggling other project schedules, travel opportunities, and professional contingencies that made knowledge sharing difficult.

At the core of the challenge of verbal communication were the disparate languages used when discussing DID–ARQ. The fourteen to 25 participating team members in the DID–ARQ Project are drawn from diverse backgrounds including computer science, history, English, art history, folklore, and interdisciplinary project management. Some members had computational backgrounds; others had little understanding of data management and organization. In the same way, while all team members were generally familiar with each other's materials due to previous collaborations, collaborators possessed little understanding of the complexities driving each of these humanities projects. Our challenge in communicating with one another was best illustrated by one of our initial issues in our grant application: the materials produced by our

computer science colleagues referred to data while the materials produced by our humanities scholars referred to primary sources or source material. As we educated one another about the underlying meanings of data (which can be singular or plural and content–agnostic) and primary source (which is generally singular and always associated with a specific format), the realization was made that we would need to devote attention to defining not just these terms but also outlining the process by which we conducted our scholarship. Thus, it became essential to uncover not just the questions outlined above but also what the context was that lent these questions with reference to the issue of authorship. Highlighting the effort to uncover terminology and meaning was the negotiation of what constituted "progress" and "success" within the Project. The final deliverables for the project were outlined as (a) data about salient characteristics of an artist with respect to another artist and with respect to a group of artists, and (b) software for obtaining salient characteristics. Yet, as we began the Project, it was apparent that negotiating three differing projects would also necessitate explicating how each institution measured success. Was success participating in the required meetings? Meeting assigned deadlines for each person's individual task? Was it aiding one's collaborators in their tasks while simultaneously completing one's own? As this was experimental research, could "success" with the Project also be achieved by realizing avenues of research that did not yield concrete results?

Our solution to "targets in motion" and evaluating progress takes into account the need to document the project work and continually discuss our ongoing efforts. First, MSU initiated a DID–ARQ listserv that would be our primary method of internal asynchronous connection. All DID-ARQ members utilize that listserv for Project–related communication that related to the complete teams. Distribution of meeting agendas, papers/publications in progress, and informational dispatches are the primary contents of the listserv. That listserv served as an internal archive of written communications and allowed for observation by evaluators at MSU. Secondly, for public consumption, MSU authored a Web site outlining the project and its progress including a list of all presentations and publications (DID–ARQ, 2010). This site is available as a central aggregator for the project and can serve a mirror site for all Project institutions to disseminate for future presentations of the DID–ARQ Project. Accompanying the MSU sites is a DID–ARQ RSS subscription feed that delivers updates to interested parties as DID–ARQ progresses. Lastly, DID–ARQ established a video conferencing apparatus for meetings that would allow for remote participation from multi–country sites.

DID–ARQ staff quickly recognized the value of video conferencing software that could record audio and video conversations. This allows team members with scheduling conflicts to watch the meeting asynchronously and with the aid of the DID–ARQ listserv, contribute to the agenda either prior to or following the live discussions. Additionally, the "target in motion" nature of DID team members necessitated having software that could accommodate a minimum of four video streams: Sheffield, MSU, UIUC, and NCSA. The popular Skype, the free video conference technology of choice of most team members, could not accommodate either recording or multiparty calls at this point nor could the popular iChat. The communications solution was offered by ooVoo. Video conferencing software in the same vein as Skype, ooVoo accommodates up to six video streams, recording of audio and video, additional telephone participants, and desktop sharing at a cost of US$440 per year. As outlined above, this is a significant cost reduction from AccessGrid which was initially considered by the Project. All DID–ARQ meetings are initiated by NCSA, who controls ooVoo access. NCSA logs in, thereby enabling the multi–party business features, and then calls the other DID–ARQ participants, who have downloaded the free, publicly available version of ooVoo. Using an agenda distributed through the DID–ARQ listserv, each call commences with the announcement that recording will begin. Following the meeting, these recordings are post–processed and uploaded to the Medici repository and tagged as technical meetings. This allows evaluators and participants to have a full catalogue of calls at their disposal to refer to at their leisure. Most significantly, this technology allows for continual re–consultation of discussions and agreements to confirm decisions and processes undertaken by DID–ARQ staff. It is an asynchronous evaluation tool for reviewers to establish the progress made bi–monthly. All aforementioned communication technologies provide a means for knowledge sharing and project coordination as illustrated in Figure 2.

4. Summary

The paper covers a range of research problems encountered in humanities endeavors that require computer science research and a collaborative infrastructure. We documented specific challenges as encountered in the Digging into Data to Answer Authorship–Related Questions project and described collaborative technologies deployed for the DID project. As the title indicates, there has been a growing number of Digging into Data problems in humanities that need to bring together geographically distributed expertise and resources, and therefore need new collaborative infrastructures supporting humanities–based computer science research. We outlined specific technologies for sharing data, software and knowledge, and emphasized the importance of legal and team/project internal agreements (*e.g.*, about sharing hardware resources, image publications, authorship). Finally, the paper also conveyed the considerations for grantors and grantees about the legal aspects, sharing agreements, time commitments and personnel resources needed for successful collaborative efforts which should be reflected in efforts planned for the future.

About the authors

Michael Simeone earned his Ph.D. in English from the University of Illinois at Urbana–Champaign. He is currently the assistant editor for *American Literary History*, as well as a project manager for the Institute for Computing in the Humanities Arts and Social Sciences. His research interests include consumer electronics and global digital culture, postmodern fiction, visual media, research databases, and practices of mapping and geospatial representation.

Jennifer Guiliano recieved a Masters of Arts in History from Miami University (2002) and a Masters of Arts (2004) in American History from the University of Illinois before completing her Ph.D. in History at the University of Illinois (2010). She currently serves as the Associate Director of the Center for Digital Humanities, a Research Assistant Professor in the Department of History at the University of South Carolina, and a Center Affiliate of the National Center for Supercomputing Applications.
Direct comments to jenguiliano [at] gmail [dot] com

Rob Kooper received his B.S. in computer science from the University of Delft in the Netherlands in 1996 and his M.S. from Georgia Institute of Technology in 2001. He currently serves as a research programmer within the Image Spatial Analysis Group at the National Center for Supercomputing Applications. His research interests are in human–computer interaction and graphics.

Peter Bajcsy has earned his Ph.D. degree from the Electrical and Computer Engineering Department, University of Illinois at Urbana–Champaign, Ill. in 1997, and M.S. degree from the Electrical Engineering Department, University of Pennsylvania, Philadelphia, Pa. in 1994. He is currently with the National Center for Supercomputing Applications at the University of Illinois at Urbana–Champaign, working as a research scientist on problems related to automatic transfer of image content to knowledge. Dr. Bajcsy's scientific interests include image processing, preservation, novel sensor technology, and computer and machine vision.

Acknowledgements

The project gratefully acknowledges the support of the National Science Foundation (NSF) and National Endowment for the Humanities (NEH) from the United States, and the Joint Information Systems Committee (JISC) from the United Kingdom in the form of a Digging into Data Challenge Grant Award.

The authors would like to thank Michael Meredith, University of Sheffield, Anne D. Hedeman, University of Illinois, Justine Richardson, Michigan State University, and Marsha MacDowell, Michigan State University of the DID–ARQ Project for their editorial contributions to this article. We also extend our gratitude to the complete DID–ARQ team: Peter Ainsworth, University of Sheffield; Steve Cohen, Michigan State University; Wayne Dyksen, Michigan State University; Kevin Franklin, University of Illinois; Karen Fresco, University of Illinois; Matt Geimer, Michigan State University; Anil K. Jain, Michigan State University; Robert Markley, University of Illinois;

Amy Milne, Alliance of American Quilts; Dean Rehberger, Michigan State University; and, Tenzing Shaw, University of Illinois.

Notes

1. For example, the popular "Log Cabin" pattern has numerous common names depending on the arrangement, such as Streak of Lightning, Barn Raising, or Sunshine and Shadows.

2. For a complete list of sponsors of the Online Froissart, please visit http://www.hrionline.ac.uk/onlinefroissart/.

3. For a complete list of sponsors of the Quilt Index, please visit http://www.quiltindex.org/about.php#funding.

4. Medici, 2010; see Figure 2.

5. Apache™, 2010; labeled as svn in Figure 2.

References

Apache™ Subversion®, at http://subversion.apache.org, accessed 12 January 2011.

Gregory Crane, 2006. "What do you do with a million books?" *D–Lib Magazine*, volume 12, number 3, at http://www.dlib.org/dlib/march06/crane/03crane.html, accessed 16 December 2010.

Cyberintegrator, at http://isda.ncsa.uiuc.edu/cyberintegrator/, accessed 12 January 2011.

DID–ARQ Web site, at http://projects.matrix.msu.edu/did/, accessed 4 October 2010.

Digging into Data Challenge Competition Web site, at http://www.diggingintodata.org, accessed 24 January 2011.

Flickr, at http://www.flickr.com/, accessed 12 November 2010.

iRODS, at https://www.irods.org/index.php, accessed 14 November 2010.

Kepler, at https://kepler-project.org/, accessed 10 October 2010.

Medici, at http://medici.ncsa.illinois.edu/, accessed 12 November 2010.

Panoramio, at http://www.panoramio.com/, accessed 10 October 2010.

Taverna, at http://www.taverna.org.uk/, accessed 11 October 2010.

TinEye, at http://www.tineye.com/, accessed 14 December 2010.

TinyPic®, at http://tinypic.com/, accessed 10 December 2010.

Versus, at http://isda.ncsa.illinois.edu/versus/, accessed 12 January 2011.

Editorial history

Received 25 January 2011; accepted 15 April 2011.

Digging into data using new collaborative infrastructures supporting humanities–based computer science research
by Michael Simeone, Jennifer Guiliano, Rob Kooper, and Peter Bajcsy.
First Monday, Volume 16, Number 5 - 2 May 2011
http://www.firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/rt/printerFriendly/3372/2950