Introduction and Project Aims

For more than 40 years, humanities scholars have used computational analysis to help resolve issues of authorship. Through stylistic and linguistic analysis, researchers have puzzled out answers to questions that range from who wrote *The Federalist Papers* and to who collaborated with Shakespeare on *Henry VIII* and *Pericles*. While determining a writer's "genetic fingerprint" is a difficult task, the wealth of scholarship and algorithms that have developed around printed textual analysis promises to help solve a number of vexing authorship issues as well as expand our knowledge of the written arts. However, in the area of visual arts, computational analysis of authorship has not made the same inroads. To do authorship studies of visual works, scholars must often do painstaking point by point analysis of small sets of 2D images of the objects. This work becomes all the more difficult when dealing with cultural artifacts such as quilts, maps and medieval manuscripts that often have corporate and anonymous authors working in community groups, guilds, artisan shops, and scriptoriums. Beyond the difficulties of authorship attribution, larger important humanities questions about the influence and migration of artistic elements and patterns become all but impossible to assess when large datasets require individual scholarly inspection of each image. To this end, we propose to address authorship and the corresponding image analyses leading to computationally scalable and accurate data-driven discoveries in image repositories.

This effort will utilize three datasets of visual works -- 15th-century manuscripts, 17th and 18th-century maps, and 19th and 20th-century quilts. Overarching humanities research questions emerge from these groups of works, such as how visual and production styles reflect regional tastes or historical moments, how traumatic historical events manifest in cultural production, and how artifacts reflect and influence relationships between cultural groups. Together these works present a range of complex authorial relationships and a strong base for advancing knowledge on the research problems inherent to scalable, automated image analyses. Open research problems are divided below into artistic, scientific and technological questions based on the specific datasets that elicit those questions. We expect these questions will be useful across the work of all three groups.

For the 15th-century manuscripts, Froissart's Chronicles, the *artistic questions* include: Where and by whom were these manuscripts created? How does a manuscript reflect the tastes of the particular region and historical moment to which it belongs? What does the codicological evidence—scribal hands, catchwords, page layouts, artistic styles in the miniatures and marginal decoration—suggest about book production in this period? *The scientific questions* for Froissart's Chronicles ask: Since these manuscripts were made during the Hundred Years' War, what was the impact of war on culture as measured by the various aspects of these manuscripts, e.g., evidence of patronage? How do they reflect contacts between the cultures of France and England? How do they reflect the ideology of chivalry or the concept of history? The questions for these medieval manuscripts are related to: (a) studying the composition and structure (codicology) of the manuscripts as cultural artifacts of the book trade in later medieval Paris; and (b) identifying the characteristic stylistic, orthographic and iconographic 'signatures' of particular scribes and artists and their collaborators who contributed to the illustration and decoration of the volumes, through the use of image recognition and data mining techniques. A further potential output from identifying scribal hands using image analysis techniques is a process that can transcribe the text from the images, a task that is currently done manually by skilled scholars. Thus not only would the content be subjected to analysis but it might also be possible to process it to allow scholars to perform further text-based mining (although not as part of this proposal) on the previously untouchable textual corpus that is locked away as pixels in an image.

The 17th- and 18th-century maps come from atlases by Joan Blaeu and Herman Moll (original atlases and digital scans held at the University of Illinois Library). The *artistic questions* for these maps include: What characteristics distinguish individual and corporate groups of artists and engravers? Criteria such as color palette, graphic representations of ships, shading of coastlines, and fonts can be considered as distinctive traits that identify both a) particular artists and engravers, b) the corporate styles developed by the Blaeu family in 17th-century Amsterdam (Joan was the son of Willem Blaeu, who founded the largest mapmaking engraving and publishing house in the world) and by Moll and his collaborators who adapted Dutch conventions of mapmaking for English audiences in the early 18th century, and c) national styles of depicting specific geographic and manmade features (cities, fortifications, trade centers, etc.). The *scientific and technological questions* are: Do specific maps show a more detailed geographical and/or climatological knowledge in representations of coastlines and harbors? Or navigable rivers? Or shoals and sand bars that pose dangers for ships? Or mountain passes that indicate potential routes for exploration and trade? The scientific and technological questions both influence and are influenced by the artistic

questions. In particular, engravers develop specific artistic techniques for representations that were essential for ships' captains, navigators, and merchants who used published maps to sail often unfamiliar and dangerous waters in South America, Asia, and the Pacific (see Appendix D). Maps therefore negotiate among art, science, trade, and politics, and determining the principles that allows researchers to distinguish among different maps and mapmakers will aid scholars working in the history of science and cartography, art, literary studies, colonial history, and economic history.

For 19th- and 20th-century quilts, *artistic questions* include: What are the distinct characteristics of an individual quiltmaker or relevant quiltmaking group's choices of pattern selection, fabric and color choices, execution of measurement, layout, needlework and craftsmanship of the pattern design, and, most interestingly, original deviations from traditional patterns? Published quilt patterns became much more common starting in the late 1800s, when certain pattern designers mass-produced their patterns and disseminated them through ladies magazines, and later in syndicated newspaper columns. Geographically dispersed quiltmakers who were exposed to this media began gaining new patterns and pattern ideas. Thus, in a large test bed of documented historic quilts, the societal rise and influence of mass media should be seen through the proliferation of quilts that execute patterns disseminated through syndicated columns. The *scientific questions* include: Can the quilts created by quiltmakers from a cloistered family, community, ethnic, or religious group at a particular time period be differentiated from those of other communities, especially those more exposed to mass media? If so, can changes in the community's participation in mass culture be found through changes in quiltmaking styles? Can a resurgence or interest in a particular historic cultural community's quiltmaking styles be found in quiltmaking a century later? To what extent are quilts made by one Amish family in the 19th century similar or dissimilar to those made by urban quilters in the same time period? Does this change over time? Or, from an even more fine-grained perspective, do we find more or less divergence in quilts from the North and from the South? To what extent are quilt patterns regional and to what extent national? Does this change over time? A major theme in American cultural history is the eclipse of regional cultural differences during the 20th-century. Can we test that hypothesis by looking at quilts? Can we use the Quilt Index dataset to measure the impact of traumatic historical events—say 9/11 or Pearl Harbor—on American culture? Do we see a measurable change in imagery, colors, or composition after such events?

While identifying distinct characteristics of artists is time-consuming, computer-assisted techniques can help humanists discover salient characteristics and increase the reliability of those findings over a large-volume corpus of digitized images. Computer-assisted techniques can provide an initial bridge from the low-level image units, such as color of pixels, to higher-level semantic concepts such as brush strokes, compositions or quilt patterns. The technological questions are related to the design of algorithms that can extract evidence at the low-level image units that could be aggregated into higher-level semantic concepts and support humanists in image understanding and authorship assignment. The further technological questions are about the statistical confidence of authorship hypotheses obtained by processing volumes of images that could not have been visually inspected with the current human resources within a reasonable time frame. How to extract knowledge about authorship and how to increase our confidence in the characteristics of authorship are the key technological questions.

Problem Description

Based on the artistic, scientific or technological questions, we formulate and address the problem of finding salient characteristics of artists from two-dimensional (2D) images of historical artifacts. Given a set of 2D images of historical artifacts with known authors, we aim to discover what salient characteristics make an artist different from others, and then to enable statistical learning about individual and collective authorship. The objective of this effort is to learn what is unique about the style of each artist, and to provide the results at a much higher level of confidence than previously has been feasible by exploring a large search space in the semantic gap of image understanding.

Motivation

Currently, humanists must look at images of historical artifacts to determine distinct characteristics of certain individual (or groups of) miniaturists and map engravers, scribes, quilters, and so on. Such visual inspection involves identifying objects in 2D images, recognizing specific types of objects, discriminating differences among those objects, classifying realizations into groups with similarities, building cumulative evidence over multiple groups of objects with similarity in realization, and assigning authorship based on temporally evolving expertise in visual inspection. For example, to assign a label of an artistic hand to an illustration in Froissart's Chronicles, we would first

identify objects such as boats, castles, crowns, faces, group of knights, horses, landscapes, skies, spears, tents, towns and water. Next, we would look for and identify the discriminating differences in all found instances of these objects and group the objects based on similarities. Finally, using meticulous book-keeping of all groups with similarities, we would build a mapping between the groups of classified objects and the potential assignment of authorship. This manual process is very labor-intensive and cannot be scaled up to large volumes of digital artifacts. In addition, the salient characteristics (a collection of discriminating differences) per artist are described at a high semantic level, which makes it difficult to automate the discovery process. Thus, there is a need to explore the semantic gap in image understanding and to establish the mappings between pixel level image properties and the higher-level abstract image descriptions.

Data Repositories and Selection Criteria

Data repositories: The primary datasets that we propose to use for this research include:

1. Nine complete Froissart manuscripts from the 15th century that have been digitized to similar standards and quality (see Appendix C). These are: Toulouse, Bibliothèque d'Etude et du Patrimoine MS 511, Besançon, Bibliothèque d'Etude et de Conservation MS 864 & MS 865, Stonyhurst College MS 1, Brussels, Bibliothèque Royale MS II 88, MS IV 251 tomes 1 & 2, and Paris, BnF MS Français 2663 and 2664. We are currently seeking funding to add two further complete manuscripts to this dataset: Pierpont Morgan Library MS M.804, and British Library MS Royal 15 E.VI. The current collection of 15th-century manuscripts consists of over 6,100 images mainly at 500 DPI, hosted on a federated Storage Resource Broker (SRB) facility between UoS and UIUC using a web-front end collaboratively developed by the two sites (see http://cbers.shef.ac.uk). The images can also be retrieved from the SRB system via an API that provides direct access to the image dataset within a programming environment such as the Image To Learn toolset (see technical methodology section).
2. Details on the 17th- and 18th-century map collections: the University of Illinois Library holds a 1664 Blaeu Atlas and over twenty of the Atlases published by Herman Moll in the early 18th century, as well as digital scans of the maps for this project (see Appendix D). These atlases include hundreds of additional maps, and the algorithms developed by this project can be applied to the thousands of pre-1800 maps that are gradually being digitized by libraries across the world. There are currently no systematic means of determining authorship for many of these maps, and the open source software developed by this project will help to encourage more digital scans of these rare and valuable but understudied resources.
3. Details on 19th- and 20th-century quilt images: the Quilt Index (a partnership of Michigan State University and the Alliance for American Quilts) contains images and detailed information on nearly 25,000 quilts, which will grow to 50,000 by the end of the grant period (see appendix E). The quilts, dating from the 1700s to the present day, are mostly American in origin though the Index will expand to include international collections in the future. Access images (550 pixel-wide JPEG files 72-150 ppi resolution) have been contributed by museums, libraries and documentation projects for education and research use. The set is hosted in MATRIX's open source digital repository, KORA, and available at www.quiltindex.org. Many thousands of styles and quilt makers are represented in this dataset as well as a range of image quality depending on original photography. For this project we have selected groupings to address four aspects of authorship: Illinois Amish family quilts from the 1800s, 1930-era Detroit quilts of Mary Schafer (who developed a very distinctive border style), typical turn of the century "crazy" quilts by Iowa quilt maker Lottie Enix, and quilts made by multiple quiltmakers using a published 1930s pattern by artist Eveline Foland. Determining salient characteristics of colors, shapes, borders, layouts and patterns with these four distinct groupings will be important to automated clustering which will uncover important new similarities and dissimilarities for a broad range of humanities analysis.

Selection of data repositories: Given the overarching goal of understanding characteristics of authorship, the proposed framework should consist of generic image analysis algorithms that could be used or adapted for use on other projects and many other datasets. We have selected datasets that represent three different historical periods and three different media but that raise analogous problems in determining authorship. The purpose of choosing such a variety of datasets is to show how seemingly different humanities research questions can share software and resources. The diversity across the three major datasets permits us to consider the computational scalability and broad applicability of the image analysis algorithms; hence we will not be producing methodologies that are only suitable to one specific type of dataset; they will have a much wider impact and use. The different datasets will further

foster integration and evaluation of algorithms so that common parts across many datasets as well as dataset specific parts of algorithms would be well understood. In addition, work developed across three or more cognate projects is certain to reinforce critical mass and to establish a creative dialogue; solutions that may seem relatively obvious to one project team may prove to be a revelation to another team in the consortium.

Project Methodology (Approach)

We propose to break down the computing problem of discovering salient characteristics into three low-level semantic components characterizing image content: (1) image representations, (2) feature descriptors, and (3) machine learning methods and similarity metrics for assignments of authorship (Appendix B illustrates a diagrammatic outline of this process).

(1) Image Representation: the image representations refer to various ways in which digital images could represent the information about physical objects. The representations include Color spaces (e.g., RGB, HSV, YUV, CIE) [1], Frequency transforms (e.g., Fourier, Hough or digital cosine transform), Special transforms (e.g., Gabor filters, co-occurrence matrices), Decomposition transforms (principal components, wavelets) and Edge transformations (Sobel, Canny, Robertson, etc. [2]). While there have been studies of what representations are close to human perception following Gestalt psychology principles [3], it has not been established how the image representations map towards discriminating artists and to higher-level semantic descriptions. We plan to explore the search space of the above image representations.

(2) Feature descriptors: Once an image representation has been selected, there is a need to describe properties of image pixels (called features) that capture local and global, deterministic and statistical, spatial and spectral image characteristics. The extraction of features can be specifically designed to focus on color, shape, texture, or motion properties. We plan to explore the search space of the most common features including 1D vector of values, color histogram, shape context [4], or texture descriptors (e.g., extracted from co-occurrence matrices) [5].

(3) Machine learning methods and similarity metrics for assignments of authorship: Given a set of features and classification labels of authorship, there exist many machine learning methods that would generate data-driven models (mappings) to convert input features into a desired set of labels. The data-driven models compare input features using similarity metrics and try to find parameters of a model that would predict the authorship labels accurately. The positive yield of models is evaluated based on the number of correctly classified instances of input features. We plan to explore the search space of multiple machine learning methods including K-nearest neighbors, support vector machine (SVM), artificial neural network (ANN), decision tree, and K-means [6].

Our approach is to explore a large dimensional space consisting of all possible combinations of image representations, feature descriptors, supervised machine learning methods and their parameters in order to select the most salient characteristics per artist. These characteristics per artist are selected based on the accuracy reported by supervised machine learning methods that compare predicted authorship assignment using the data-driven models with the provided authorship labels. The result of such extensive searches would lead to an n-tuple that provides the highest discrimination with reference to two artists. For instance, let us assume that the n-tuple found consists of (a) hue color channel in HSV image representation, (b) frequency of occurrence of each hue value – hue histogram, and (c) similarity of hue histograms measured by chi-squared error and aggregated into groups with similar features using three nearest neighbors. Then, a humanist could interpret the discriminating characteristics of two artists to be a hue component of image colors, a statistical distribution of hue variations in image pixels, and neighboring similarity of hue distributions in the space of all possible statistical distributions. Thus, visual inspections by a humanist would be assisted by a computer-driven recommendation to focus on a hue component of color images and the similarity of hue distributions in images (or the similarity of hue value frequencies across images). This would reduce the search time of a humanist and could change the role of visual inspection from searching to verification and validation. Furthermore, the images would be delivered for visual inspection in the appropriate representation (e.g., hue channel and its hue histogram) rather than leaving a humanist to recover the hue representation from another color space representation by color transforming images inside of his/her brain.

Similarly, these pair-wise (artist-to-artist) analyses would lead to a matrix of discriminating characteristics that could be summarized and presented to a humanist. The summaries provide computer-assisted input into research questions about what salient characteristics of an artist dominate within a group of artists, a school of artists

or a community of artists. Furthermore, they would be tremendously useful in forensic studies when unseen images are presented to determine authorship.

Essentially, and in respect to determining scribal hands, this scientific methodology would help scholars to identifying recurring tell-tale signs pinpointing the work of Scribe A, B or C by providing answers to our scholarly questions such as which forms, or combinations of letters (e.g., ligatures such as 'ct' or 'br'), do our algorithms reveal as being key to distinguishing between A, B and C?

Computational Requirements

The proposed approach is clearly computationally intensive due to the huge size of the search space for selecting the optimal triplet (the most salient characteristics represented by image representation, features and machine learning methods). The dimensionality can be estimated by multiplying the number of image pairs to be evaluated times the number of image representations times the number of features times the number of machine learning metrics times the number of cross validations. For example, computing the optimal triplet that discriminates between two artists, each represented by 10 images, over 2 color spaces (HSV, H, S, V, HS, HV, SV, RGB, R,G, B, RG, RB, GB), 1 feature type (histogram of each band with the number of bins varying from 100 to 255), one machine learning method (k-nearest neighbors with the k parameter taking values 1, 2,3,4,5) and 5-fold cross validation requires (10+1)*5 of image pairs to be evaluated times 14 image representations times 155 features times 5 machine learning variations times 5 cross validation evaluations. This number is equal to 2,983,750 computations with many floating point operations during accuracy evaluations of machine learning models. Clearly one feature type and one machine learning method will not be sufficient to capture the spectrum of underlying image characteristics that discriminate artists. We anticipate computations to number about 300 million after adding variables one each dimension and reducing the number of parameters based on initial optimizations (e.g., the number of histogram bins).

Development Methodology & Standards

The Image to Learn (IM2 Learn) toolset[1], developed at NCSA, provides a suite of image transformation functions that will be applied to our dataset to support the first phase of our investigations. IM2Learn additionally provides an API to facilitate the development of other image transformation algorithms that do not currently exist within the package. This standard API will be used by the project partners while developing further functions to ensure cross-compatibility across the sites.

The feature description catalogue (phase 2) will be recorded using an XML database and document type definition (DTD) that will be developed in consultation between the technical developers and scholars to meet the needs of the project from both perspectives. This will provide a structured document that is both human and machine readable that will be directly input into phase 3.

The machine learning component of our methodology (phase 3) will make use of the open source Weka software[2], developed by the University of Waikato, which provides a collection of machine learning algorithms and a platform to develop additional functionality to suit this proposal. However, unlike the image transformation phase and the use of IM2Learn, our project may not be best served by a direct uptake of the Weka interface due to an increased complexity in operations, particularly since one of the aims of this initiative is to scale-up the algorithms to work on high-performance computing grids. The technical project team will therefore meet to decide a suitable framework, also taking into account previous decisions from the earlier stages to ensure the process is both streamlined and coherent while the partner sites collaborate on aspects of functionality.

The algorithms contained within the IM2Learn and Weka packages will be enhanced by the inclusion of new and existing code from the project partners converted into the common framework/interface. This activity will be supported by bringing together the algorithms that each partner has developed in past projects which relate to this application. Furthermore, to facilitate scalability, we anticipate that the algorithms that feature in our final pipeline will need to be written and optimized to allow large-scale parallelism. The new algorithms developed within the scope of this proposal will be made available as open source in line with the exit strategy outlined below.

---

1 http://isda.ncsa.uiuc.edu/Im2Learn

2 http://www.cs.waikato.ac.nz/ml/weka

Documentation on the framework and an evaluation of the interfaces, protocols and standards used during the development stages will be published as part of the final report. This will include particular references to the team success in achieving a platform for working collaboratively on interoperable code across geographically remote sites.

Environmental Scan

Based on our knowledge, there has not been an effort in the past such as the proposed one. There are publications referring to forensic studies of individual photographs of arts investigating the authenticity [7, 8], or the use of optics during art creation [9, 10]. Several researchers have explored certain sub-sets of features, for instance, brush strokes [11, 12], for classification purposes. The closest effort to the proposed one is the work of Shen [13] where 1080 classical paintings from 25 artists were analyzed using a collection of color, texture and shape features. However, the search space in [13] is constrained to only searching over one data representation (CIE) and one machine learning method (neural network). In addition, our proposed effort is different not only in the scale of data sets and required scalability of search computations but also in posing a fundamentally different question of finding salient characteristics discriminating two authors and groups of authors at the low image level by computers and then at the higher semantic level by humanists. This approach has been initially investigated by two coPIs of this project at UIUC (Hedeman and Bajcsy) with funding from the NCSA Fellowship program [14].

Final Product and Deliverables

The final product will primarily consist of (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. The data could be viewed as evidence supporting authorship assignment based on (1) image-derived primitives including image representation, image features and machine learning model for assigning authorship, and (2) human-defined semantic descriptors of unique characteristics that map into a combination of multiple image-derived primitives.

The final product would be used immediately by scholars for book trade studies, for understanding engravers and cartographic artists, and for addressing questions about collective and individual authorship of quilts. Additional likely users of the final product would be the researchers and students at UIUC, MSU, University of Sheffield, and other universities for educational purposes. We foresee other humanists wanting to know about the authorship of all kinds of collections and using the software framework and the algorithms. The algorithms and software developed will have appropriate technical and user documentation and all content made publicly available as open source for non-commercial use. The final products will demonstrate to a wide audience how large-scale data repositories of cultural heritage materials can change the nature of research for the humanities and social sciences.

Dissemination

Websites will be created at each of the partner sites at the start of the project to introduce and publicize the work. These sites will act as primary points of contact for project information, offering up-to-date examples of results and blogs (which will be used to inform the final report). The websites will host the final deliverables and include the source code and documentation developed during this initiative.

A workshop in Sheffield is planned to take place around month 9 to bring together other expert scholars in the field of authorship from around the UK. Its primary purpose is to engage with the wider community and understand more broadly how other experts interpret authorship and how the preliminary project results might help their work. We also anticipate that the workshop will influence our choices of image representations during the machine learning phase. The engagement we accomplish during this workshop will also disseminate the work by demonstrating its potential to a wider audience, who themselves have a specific interest in authorship.

Further avenues for dissemination and publication that the UK team will target include the All Hands and Digital Resources for the Humanities and Arts conferences which showcase e-Science tools and methodologies and the Digital Humanities Quarterly journal. Other workshops and conferences will also be attended to disseminate and publish this work. Members of the UIUC team will present status reports on the project at national and international conferences in art history, cultural history, and medieval and eighteenth-century studies.

Project History

These three partners each have considerable experience with their own datasets represented, and have established working relationships on imaging issues prior to the development of this proposal. In April 2009 members of each partner site were part of an NSF workshop, “Imaging and Image Analyses Applied to Historical Objects,” at

the University of Illinois which addressed the process of going from physical historical objects to digital historical objects available via the Internet for educational and research purposes. The overarching theme of the workshop was to understand the challenges associated with imaging and image analysis that are inherent in this process.

The Sheffield project director (PD) is close to completion of the AHRC-funded "Online Froissart" project. The project will deliver an online facsimile/transcription and translation from Books I-III of Froissart's Chronicles, based on digital surrogates of original manuscripts and using XML. The "Online Froissart" will therefore help reinforce mark-up techniques and standards for the present proposal as well as provide knowledge about scribes and artistic hands; research tools emerging from the project, including image-processing algorithms, will also feed into the Digging into Data proposal. The Sheffield PD has also led on the EPSRC-funded "Pegasus" project, which considered grid-enabled interfaces [15] for sharing and displaying in real-time online image datasets and virtual exhibitions to a distributed audience; this was done in partnership with Urbana and used the "Virtual Vellum" software from Sheffield.

With the support of a Faculty Fellowship (2008-09) from NCSA, Co-PIs from the University of Illinois (Hedeman and Bajcsy) began work on developing cyber tools for analyzing the visual imagery embedded in the corpus of Froissart manuscripts (a corpus that will ultimately include the Shrewsbury Book and Morgan MS), in order to provide insights into both the artists' contributions to the construction of these specific books, and more broadly, the functioning of the medieval book trade. The Sheffield and Urbana team have also worked closely on other grant applications including a pending JISC/NEH digitization proposal. NCSA researchers have also explored preliminary automated pattern analysis of "crazy" quilts, which will be part of the dataset and questions addressed in this project. Finally, NCSA and ICHASS have been actively supporting the 18th Connect project with computational resources and expertise needed for pre-1800 optical character recognition.

Time Plan and Project Management:

Month 1: Data and initial source code sharing platform established between all partners with all relevant material uploaded. Project meetings will be held over Access Grid. Project websites launched at all three sites. Consortium agreement (JISC requirement) including any new responsibilities, data policies, finalized evaluation plan, etc. arising from first round of project meetings.

Month 2: Scholars and developers discuss base-line algorithms and selection of training data

Months 3-4: Integration of existing algorithms and alpha development of new image transformation code from initial conversations. The alpha versions of new code will be rapidly developed (i.e. not optimized) as a means of getting results quickly to determine how useful the algorithm will be long-term.

Month 5: Scholars determine appropriate feature sets based on transformations with discussions with developers and DTD for XML database decided.

Month 6: Optimization, modification and validation of image transformation code for large-scale processing.

Months 7-8: Apply machine learning algorithms to test dataset to determine suitable discriminative n-tuples suitable for authorship classification. Project exposure to classroom students at UIUC, MSU and UoS

Month 9: Sheffield workshop to discuss results and authorship issues among larger scholarly community. Analyze results to prune misleading representations. Optimize algorithms for large-scale processing.

Month 10: Re-apply machine learning algorithms to test dataset to determine validity of changes.

Months 11-12: Final data preparation, computational resource configuration and data processing of complete dataset collection. Presentation of the work at the eScience meeting.

Months 13-14: Documentation of algorithms and technical processes. Scholarly interpretation of results. Draft final report. Incorporating the project results into teaching materials.

Month 15: Completion of documentation, final reports (including financial statement) and website updates. Dissemination of project results to classrooms outside of the three partnering universities.

Note: During development of algorithms, the developers and scholars will continue to work closely together; the milestone months indicated above will be used to consolidate progress; formal meetings will be held to monitor progress across all partners.

The three project directors from each partner site (Ainsworth, Bajcsy & Rehberger) and project managers (Guiliano & Richardson) will meet formally on a quarterly basis to monitor and report overall progress and respond to unexpected issues that risk causing deviation from the time plan. These meetings will be conducted over access grid/teleconferencing. Each site project director has responsibility for overall management of his/her research teams

(as per funding bodies) and reporting back to the funding councils. The research teams primarily include scholars and technical staff – the project directors also fall into one of these categories – and are responsible for undertaking the work as outlined above.

The project team has been selected across the partner sites to bring together the best cross-disciplinary levels of expertise calculated to deliver the collective goals (see Résumés section for team and individual areas of expertise). Expertise in the different areas will be shared throughout the project to meet the objectives; all technical staff across the sites will work closely together to develop the algorithms, as opposed to each site focusing on a specific dataset. This has the additional benefit of ensuring that work is not duplicated, while cross-site compatibility is assured by the common framework and interfaces that will be used: the Image2Learn API will enforce the standards used to develop the image transformation algorithms; however the machine learning algorithms and optimization steps are more technically involved. Thus to ensure consistency between sites, the Sheffield developer will spend a few weeks at NCSA working with their developers at this stage of the project (approximately month 6/7). During this time, any APIs necessary and methodologies required to run the code across the different sites on different grid infrastructures will be established along with optimization strategies. The coordination of work across the technical staff will be the responsibility of the head technical developer and NCSA project director, Peter Bajcsy. All technical staff will liaise informally via electronic communications.

Dissemination activities will be undertaken by either the research team or specialized personnel depending on how the research teams at each site are organized (see list of participants). Dissemination personnel will be managed by the site PD and overall strategic dissemination policies will be realized across the whole team. Throughout the project all team members will communicate with each other and outcomes will be disseminated throughout the team via the blogs. Formal meeting minutes will be made available on the websites.

As one part of project management, the teams have also discussed budget allocations devoted to the three key components of the project, such as computer science (CS), humanities (H) and dissemination (D). The teams have aligned their resource allocations with the project requirements in order to contribute with complementary expertise. The approximate allocations are (a) JISC funding 80% (CS): 10% (H): 10% (D); NSF funding 73% (CS): 20% (H):7% (D), and NEH funding 40.1% (CS): 44.6% (H): 15% (D).

Exit/Sustainability Plan

The project deliverables will be subject to the licensing policies outlined by the Creative Commons initiative, and Illinois open source license[3], and made publicly available for non-commercial use as open-source, thus fostering adoption within the wider community and use for different datasets. The deliverables will be deposited with JorumOpen (in line with JISC's funding policy) and our project websites for download.

We believe that the scholarly results obtained from the data analysis performed will reach beyond this application with only preliminary findings published in the project lifespan due to the development program and time constraints. The methodologies developed will therefore continue to be used by the team members as new discoveries are found and validated.

From a development perspective, at this stage we do not expect the computer to determine authorship indisputably, but rather to cluster similarities together to allow the scholar to focus on these without trawling through large datasets. The work carried out within this proposal will therefore be used as a platform to bid for further funds from agencies, to allow us and others to continue to explore research directions that this initiative uncovers.

Risk Management and Intellectual Property Rights

We do not consider there to be any significant risks associated with the proposal. Key staff members are completely committed to the project and we have full rights and permission to use the image datasets in line with the purposes outlined. All existing software we plan to use is open-source. In the event of unexpected changes in staffing, we anticipate that our collective research team and further recruitment will be able to compensate due to the complementary skills that each team member brings to the project.

In terms of the technical validity of this application, a precursor to this work has already been undertaken at NCSA [14], demonstrating promising findings, and thus proof-of-concept, which support the research methodologies set forth in this application. This application will, however, broaden and extend beyond the previous work.

---

[3] http://www.otm.illinois.edu/node/396