Ensuring a Future for the Past

Long-term Preservation Strategies for Digital Archaeological Data

Adam Rabinowitz,[1] Maria Esteva,[2] and Jessica Trelogan[1]

The University of Texas at Austin, [1]Institute of Classical Archaeology and [2]Texas Advanced Computing Center

Abstract

As the documentation of archaeological research is increasingly born digital, the preservation of archaeological knowledge is more and more dependent on the documentation and long-term curation of those digital files themselves. Because archaeological investigation is destructive, excavation records provide the only source of evidence for contextual relationships. Digital archaeological records are threatened not only by technical change and equipment failure, but by insufficient metadata. This paper describes the collaborative efforts of the Institute of Classical Archaeology and the Texas Advanced Computing Center to develop strategies to ensure the preservation and accessibility of the digital archaeological record in the long term.

Authors

Adam Rabinowitz is Assistant Professor of Classics and Assistant Director of the Institute of Classical Archaeology at the University of Texas at Austin. He holds a PhD (2004) from the Interdepartmental Program in Classical Art and Archaeology at the University of Michigan and is a 2002 Fellow of the American Academy in Rome. He has more than twenty years of archaeological experience at Greek, Roman, and Byzantine sites in Italy, England, Israel, Tunisia, and Ukraine. His current projects include the publication of recent excavations at the site of Chersonesos (Ukraine), which will involve an online, GIS-enabled database of primary documentation.

Maria Esteva has been the Data Curator at the Texas Advanced Computing Center in the Data and Collections Management Group since receiving her PhD in Information Science from the School of Information at the University of Texas at Austin in 2008. Her work includes the design of metadata architectures for evolving scientific and humanities data collections, the implementation of data collections management and archiving workflows, and the training of researchers in diverse aspects of data management. Esteva is the PI of an ongoing research collaboration with the National Archives and Records Administration (NARA) that investigates the use of computational and visual analytics methods for archival processing. She has presented her work at the Digital Humanities, Digital Curation, ASIST, IS &T Archiving, SAA and VIS/VAST conferences.

Jessica Trelogan is a Research Associate at ICA specializing in GIS and remote sensing technologies for archaeological applications. She has an MA in Classics (1994), and has been involved in all stages of archaeological research -- from the field to publication to digital archiving -- at ICA's excavation, conservation, and survey projects in Italy and Ukraine.

1. The nature of the problem

As the UNESCO Memory of the World program recognizes, digital tools have exponentially expanded our ability to create and share documents of all forms. At the same time, this vast collection of digital data is disturbingly fragile, as digital files and systems themselves require active curation if they are to survive in a usable form even a few decades into the future. The tension between the possibilities and risks of digital documentation is nowhere more clear than in the field of archaeology.

In most humanities research, digitization or digital recording provides an additional layer of documentation for texts or objects that exist in the physical world. If that layer of digital information disappears, the physical items it represented can still be examined. The use of digital documentation strategies, then, only offers advantages: if digital versions are preserved, they provide an additional mechanism for the long-term survival of the information contained in the item, and if they are not preserved, the sum of knowledge is at least no less than it was before the digital version was created. This fact has been recognized by those in charge of archaeological archives consisting of paper documents and film negatives, and both major and minor projects focused on the digitization of such archives have emerged in recent years.[1] The challenges in these projects have been mainly related to the modelling of the data collections and the development of methods to share them effectively. Digitization in this case is a low-risk, high-reward strategy for preservation and dissemination.

The situation is quite different, however, for the use of digital records in the primary documentation of current archaeological research, especially when that research involves excavation. The excavation of an archaeological site, by its very nature, destroys the object of investigation. After a dig, the contextual relationships between artefacts, soil deposits, buildings, and other elements of the human environment only survive in the project's documentation. Perhaps even more than other humanistic disciplines, archaeology has been quick to adopt sophisticated digital tools, and the use of everything from digital photography to computer databases to three-dimensional modelling is now widespread in the field. As a result, the contextual record for many archaeological sites now exists primarily in digital form.

On one level, this has been a boon: digital tools have led to unprecedented levels of detail in field recording, better means of visualization of the process of excavation, and a dramatically increased ability to share digital and digitized data in all their original contextual splendour. Digital photography has freed projects from the financial constraints of film and development costs, and excavations that were once documented with a few hundred photographs are now documented with thousands, if not tens of thousands, of digital images. Digital frameworks composed of relational databases and Geographic Information Systems (GIS) allow information to be queried and filtered in ways that promote new understanding and interpretations. Moreover, the integration of digital datasets from different projects allows new questions to be asked and inspires new observations, and web-based interfaces have opened the results of research to the public on a global level.

---

[1] One of the most visible of these efforts in the field of Classical Archaeology is the digitization of the archives of the excavations carried out by the American School of Classical Studies at Athens in the Athenian Agora and at the site of ancient Corinth ("Digital Collections, The American School of Classical Studies at Athens," accessed September 14, 2012, http://www.ascsa.net/research?v=default). Two of the authors of this paper have been involved in such digitization projects both at our home institution, the Institute of Classical Archaeology, and at the site of Chersonesos in Crimea, where substantial portions of more than a century of archival records have been digitized and presented online ("K. K. Kostsyushko-Valyuzhinich and his Reports for the Imperial Archaeological Commission," accessed September 14, 2012, http://kostsyushko.chersonesos.org/; "Discovering Chersonesos," accessed September 14, 2012, http://www.discovering.chersonesos.org).

In order for data to be shared and preserved, however, they must also be documented in ways that allow them to be compared and reconstructed in the future. Over the last 15 years, several centralized repositories for archaeological data have emerged, with their primary task the creation of metadata and the curation and migration of original datasets.[2] At the same time, various groups have worked on metadata ontologies and schemata that allow highly heterogeneous archaeological data to be described in ways that are mutually comprehensible between archaeological communities and across national boundaries.[3] These efforts in the archaeological sphere are similar to efforts in other spheres of digital documentation, which, guided to a large extent by library practice, have focused on metadata standards and formal state-level or institutional digital repositories.

Digital archaeological documentation presents much greater barriers to standardization, however, due to its heterogeneity, its level of relational complexity, its use of a wide variety of proprietary or custom-built software platforms to manage contextual relationships, and the extremely varied and idiosyncratic circumstances of its production.[4] The general archival schemata most frequently used by libraries are ill-fitted to the contextual relationships embedded in archaeological documentation, while custom-built schemata or ontologies designed to reflect the full complexity of archaeological data are inevitably so complicated themselves that they are very difficult for either non-information-scientists or non-domain-experts to understand. The varied environments in which archaeological data are produced also work against centralizing efforts. Where there is a relatively small, homogeneous community of practice and a strong incentive (e.g. a legal requirement) for the submission of digital data to a repository, centralization has been an effective strategy: in the UK, for example, the ADS, by governmental mandate, is the repository of record for all information produced by contract archaeologists. Where the community is large, diverse, fractious, and lacking in incentives, on the other hand, centralization tends to fail. In Mediterranean archaeology, projects are still more likely to request funds to build their own unique databases, in which they use their own idiosyncratic terminologies and metadata structures, than they are to request funds to allow them to deposit their data in an existing repository.

The issue of decentralization is compounded by national and cultural factors: not only do practitioners in different countries use different languages and describe and organize their material differently, but many archaeologists are still intensely suspicious of digital repositories and deeply reluctant to share the results of their research in any form other than traditional paper publications. Since the professional incentives for most academic archaeologists still centre on publications "branded" by individuals or at least by individual projects, even those scholars who embrace the potential of digital

[2] The most prominent of these are the UK's Archaeology Data Service ("Archaeology Data Service Homepage," accessed September 14, 2012, http://archaeologydataservice.ac.uk/), established in 1996, and the US-based Digital Archaeological Record ("tDAR," accessed September 14, 2012, http://www.tdar.org/).

[3] These efforts have taken several forms, one of which focuses on the creation of semantic frameworks and formal ontologies for cultural heritage material (e.g. the CIDOC-CRM: "The CIDOC Conceptual Reference Model," accessed September 14, 2012, http://www.cidoc-crm.org/), another of which combines a domain-specific ontology with a specific metadata schema (e.g. OCHRE and ArchaeoML: "Online Cultural Heritage Research Environment," accessed September 14, 2012, http://ochre.lib.uchicago.edu/), and yet another of which focuses on the adaptation of existing metadata standards like Dublin Core.

[4] In the field of Mediterranean archaeology alone, one finds everything from small-scale academic projects run by one individual to large-scale academic projects carried out by several institutions, long-term excavation projects with 100-year histories controlled by one country's research bases in another country, contract archaeology carried out by for-profit companies, rescue archaeology carried out by the staff of national archaeological services, museum research, etc.

media largely focus on short-term goals like attractive, customized websites rather than less exciting issues of long-term preservation.

This introduces the dark side of digital documentation in archaeology: few or no provisions have been made for the long-term preservation of, and access to, the vast majority of digital archaeological data. Many, if not most, digital archaeological collections are at risk on two major levels, even if we leave aside the question of the longevity of storage media. The first threat lies in the obsolescence of the proprietary applications that manage contextual relationships. Most archaeologists who were already working in the field during the advent of digital methods wince when they remember databases in DBIII that are now unrecoverable. The second threat lies in insufficient metadata for individual files associated with complex digital records, either because the intermediary program that managed the metadata is no longer accessible or because metadata was never created for these files in the first place.[5] In the latter case, the metadata exists only in the heads of the excavators, and without access to those personal memories, much, if not all, of the context surrounding a given piece of documentation is lost. The loss of that context means the loss of some or all of the information contained in that file -- and when that file documented a feature that was destroyed in the course of archaeological investigation, this means the permanent and irrevocable loss of that part of the memory of the world.

This is no longer primarily a technical problem for archaeology, for many of the technical issues have already been addressed by existing domain-specific initiatives. The archiving community has established standards and best practices for the curation and migration of many types of digital files; the community of information scientists associated with archaeology and cultural heritage has created ontologies and metadata schemata that are suitable for almost all the types of information archaeologists currently produce; and several effective, sustainable repository infrastructures are in place.[6] It continues, however, to be a human problem. The creation of extensive metadata and the curation of files and formats is time-consuming and unrewarded, and therefore a low priority. Metadata standards are hard to understand and harder to apply to an existing dataset without the help of a specialist in information or library science. The submission of a dataset to a repository usually involves both the loss of control and diminished functionality, and requires that the dataset remain static and unchanging, all drawbacks that are anathema to many directors of archaeological projects. And all of these things cost money that few archaeological projects, especially outside the first world, can spare.

To escape the cloud of digital amnesia now looming over the field of archaeology, we must address not only the technical side of the question, but the human side as well. It is crucial to recognize the barriers -- cultural, psychological, financial, and disciplinary -- that prevent the creators of archaeological data from taking effective measures for their long-term preservation. Any solution has to involve not only centralized repositories and universal standards, but also tools that will work for individual data owners on a cellular and distributed level. Such tools should make it fast, easy, and cheap for those data owners to provide metadata for their collections. They should bridge the knowledge gap that deters many domain specialists from dealing with metadata standards and archival practices. They should allow data owners to preserve the idiosyncratic conceptual and organizational principles that structure their datasets. Finally,

[5] Examples are digital photographs taken of a stratigraphic layer or find during excavation, which in many databases are embedded or referred to without being given metadata of their own on a file level, or the individual shapefiles or feature classes in a GIS, which are difficult to understand without additional documentation.

[6] See, for example, the extensive guides to best practice prepared by ADS and tDAR: "Archaeology Data Service/Digital Antiquity Guides to Good Practice," accessed September 14, 2012, http://guides.archaeologydataservice.ac.uk/g2gp/Contents.

they should offer clear short-term rewards, either making it easier to share and publish data online or lowering the costs in time and money of the eventual transfer of a dataset to a repository.

The collaboration between the Institute of Classical Archaeology (ICA) and the Texas Advanced Computing Center (TACC) at The University of Texas at Austin, now in its fourth year, has been focused on these issues. It developed because ICA found itself in exactly the sort of situation described above: it possessed a large quantity of digitized archaeological data and a growing set of born-digital data, much of which was in proprietary formats or managed by proprietary applications, and much of which lacked metadata on the file level that could, in the absence of a database or the excavator, explain what was represented, for example, in a digital image. Together with this increasingly unmanageable digital dataset, a series of equipment failures and file corruptions provided compelling reasons to seek a long-term solution. The dataset was too large, too complex, and too much in flux for it to be handled by the digital repository of the UT library system, however, and the centralized archaeological repositories that existed at the time could not offer to preserve the full database functionality and spatial tools that ICA saw as essential to the management and publication of its data.

TACC, on the other hand, was very familiar with the management of complex, dynamic datasets, and its services are available at low or no cost to UT research projects.[7] Corral, its data facility, had the technological infrastructure to support the full functionality of ICA's existing data management solutions: this resource consists of six petabytes of online disk space and supports databases, web-based access, a high-performance parallel file system, and other network protocols for storage and retrieval of data. For database collections, Corral provides DB server nodes running MySQL, PostgreSQL, and SQL Server, as well as support for open source domain specific databases; a GIS server to allow web and desktop access to spatial datasets is also being implemented. Most importantly, TACC had recently deployed a new Data and Collections Management Group (DCM) to support data intensive research activities. The DCM group designs, builds, and maintains the data applications facilities and consults with researchers in aspects of their collections lifecycle, from creation to long-term preservation and access. Group members are specialized in software development, Relational Database Management Systems (RDBMS), Geographical Information Systems (GIS), scientific data formats, metadata, large storage architecture, systems administration, digital archiving and long-term preservation. At the same time, TACC was also seeking collaborations in the digital humanities, and thus a joint project was mutually beneficial.

2. The initial ICA-TACC collaboration: excavation data from Chersonesos, Ukraine

2.1. The dataset

This project began with a primarily born-digital dataset created in the course of excavations at the Greek, Roman, and Byzantine site of Chersonesos in Crimea, Ukraine. These excavations, carried out between 2001 and 2006, had focused on a residential block in the urban centre of the site with a continuous 2000-year record of occupation and a wide range of archaeological material, from ceramics and metal objects to human remains, charred seeds and metalworking waste. Contextual records were kept in a digital database, first in Microsoft Access ®, then in a SQL database with an Access front-end, and then in ARK

[7] Maria Esteva et al., "Cyberinfrastructure supporting evolving data collections," in *Proceedings of the 8th International Conference on Preservation of Digital Objects, iPRES 2011, Singapore, November 1-4 2011*, eds. Borbinha et al. (Singapore: National Library of Singapore and Nanyang Technical University, 2011), 93-96.

(the Archaeological Recording Kit), a SQL database with a web-based front end.[8] This database also managed a large number of digital photographs of excavation and objects, and it was linked to a GIS that included two- and three-dimensional data in both vector and raster form. Other born-digital data from Chersonesos included specialist spreadsheets, 3D models of objects, and a set of interactively-lighted image files (reflectance transformation images), each of which was derived through the processing of a large number of related individual digital photographs.[9] In addition to these born-digital records, the digital dataset also contained digitized versions of paper recording sheets, hand-drawn plans, and hand-inked object illustrations, as well as both scans and electronic transcriptions of excavation notebooks.

ARK serves as a tool for the management and presentation of archaeological data, but it is not intended as a long-term preservation system, nor does it provide a complete representation of all the objects produced through the research process. Furthermore, some files in the dataset were in proprietary formats, while even files in open formats were often managed by proprietary systems such as ESRI's ArcGIS. The problem of organizing and preserving the original raw data, as well as the data produced off-site and after field seasons, still had to be resolved. Increasing numbers of DVDs, hard-drives, two servers, and personal computers filled with unlabelled or inconsistently labelled data were overwhelming the researchers and undermining the potential for the future reuse of those data.

2.2. Evaluation and triage

During initial conversations and interviews between the DCM group and the ICA team, it became clear that the project had complex requirements. ICA needed to archive data and it also needed an integrated digital environment in which different team members could organize and keep track of active and archival data within a dataset that was both changing and growing as research progressed. As a first step, the team carried out a triage of the Chersonesos collection. This triage was an extensive preliminary examination of the collection that aimed to define its structure, identify the types and locations of its various components, and describe the research workflows that produce those components. To approach this task, we used Records Management concepts and practices, including functional analysis and collections inventorying.[10] Our analysis was able to identify four broad research functions with associated data types and workflows: on-site data collection in the course of excavation, post-collection processing of both excavated objects and digital information (both during and after the field season), analysis and interpretation, and publication. We then designed a system to inventory groups of objects in relation to these research functions. For practical reasons, this system focused not on the documentation of individual items, but on the description of groups of digital objects, for which type, location, and relations

[8] For a description of the evolution of the recording system, see Adam Rabinowitz, Stuart Eve, and Jessica Trelogan. "Precision, accuracy, and the fate of the data: experiments in site recording at Chersonesos, Ukraine," in *Digital Discovery: Exploring New Frontiers in Human Heritage, CAA 2006*, eds. Jeffrey Clark and Emily Hagemeister (Budapest: Archeolingua, 2007), 243-256. For discussion of ARK, see Stuart Eve and Guy Hunt, "ARK: a developmental framework for archaeological recording," in *Layers of Perception: Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), Berlin, Germany, April 2-6, 2007*, eds. Axel Poluschny and Karsten Lambers (Bonn: Deutsches Archäologisches Institut; Dr. Rudolph Hablet Gmbh, 2008).

[9] Adam Rabinowitz, Carla Schroer and Mark Mudge, "Grass-roots imaging: a case study in sustainable heritage documentation at Chersonesos, Ukraine," in *Making History Interactive: Proceedings of CAA 2009*, eds. B. Frischer and L. Fisher (Budapest: Archaeolingua, 2010).

[10] William Saffady, *Managing Electronic Records*, 4th ed. (Lenexa, KS:ARMA International, 2009).

to other groups of objects or records -- such as "component of", "derivative of", "description of", "complementary documentation of" -- were defined.

This initial evaluation and description of the collection allowed both TACC and ICA collaborators to understand it more thoroughly.[11] The process revealed the variety of objects included in the collection and highlighted the need to impose a more logical structure to identify systematically their types, roles, and content, but it was also limited and inefficient. It required the efforts of TACC information scientists, a graduate student from the UT School of Information, and several undergraduate students, in addition to those of ICA and TACC staff members, and the inventory and organization process alone took nearly two years. Even after this process, a substantial number of digital objects lacked full identifying information. Attaching metadata by hand to all of the files would have involved prohibitive costs in time and person-hours. We also wanted to be able to apply the large quantity of information about data objects and their contextual relationships that were already encoded in the ARK database. The answer lay in the creation of a semi-automated solution for the generation of object-level metadata.

2.3. Metadata extraction

The solution centered on the use of iRODS,[12] a rules-based storage infrastructure deployed on Corral. iRODS acts as a management platform for archival collections, while the preservation of the data it contains is ensured by replication at TACC's Ranch tape archive and at other HPC sites in Texas and across the country. Within such an environment, research teams can seamlessly integrate data management activities throughout the various stages of research. In the case of the Chersonesos dataset, our study of the collection and its workflows led to a decision to implement a file-management strategy that would also produce metadata automatically as data were ingested into iRODS for storage.[13] This strategy involves two basic components: a standardized file naming convention, and a hierarchically labelled directory structure to classify and group related data. Both components were designed to provide descriptive, structural and contextual metadata that can be parsed and mapped to standards and that reflect the project's data lifecycle. An ingest script on the iRODS side uses the naming convention and the directory hierarchy to extract such contextual metadata, together with preservation metadata, and encode them as XML Dublin Core (DC), Preservation Metadata Maintenance Activity (PREMIS), and Metadata Encoding and Transmission Schema (METS) metadata schemata.[14]

The file naming convention contains four information elements: a) an alphabetic code that indicates the ARK database module to which the digital object should be attached; b) the alphanumeric code assigned in the field and in the ARK database to the object, context, or record represented by the data object; c) the stage of the research process in which the data object was produced; and d) the designation of the file as either a master (an unmodified original) or a version (a derivative of a master file). The research stage designations were developed by the researchers; they apply primarily to images of objects

[11] Maria Esteva et al., "From the site to long-term preservation: a reflexive system to manage and archive digital archaeological data," in *Archiving 2010. Proceedings of the Archiving Conference, vol. 7* (Society for Imaging Science and Technology, 2010), 1-6.
[12] "IRODS Data Grids, Digital Libraries, Persistent Archives and Real-time Data Systems," accessed September 14, 2012, https://www.irods.org/index.php/IRODS:Data_Grids,_Digital_Libraries,_Persistent_Archives,_and_Real-time_Data_Systems.
[13] David Walling and Maria Esteva, "Automating the extraction of metadata from archaeological data using iRods rules," *International Journal of Digital Curation* 6:2 (2011), accessed September 14, 2012, doi:10.2218/ijdc.v6i2.20.
[14] "Library of Congress Metadata Schemas," accessed September 14, 2012, http://www.loc.gov/standards.

and include: 'b'=before conservation,'d'=during conservation, 'a'=after conservation, 'l'=lifting, 'm'=microscope, and 's'=studio. Multiple data objects may be produced for a given archaeological object at a given research stage, so a sequence is recorded numerically following the stage designation. For example, for the data object (in this case, an image) named 'sfi_CH05SR_3065_a1_m.JPG', 'sfi' identifies the file as associated with the "special finds" module in ARK; 'CH05SR_3065' identifies the object represented as registry number 3065 from Chersonesos excavations in the city's South Region in 2005 and matches the item key for this object in ARK, 'a1' indicates that the file is the first of several created after the object's conservation, and 'm' indicates that it is the original version of the file (in this case, the original jpg generated by the camera when the photo was taken). Naming an object in accordance with this convention allows for the automatic extraction of metadata about that object, both from the stage codes embedded directly in the file name and from an external data management system like ARK through the inclusion of item keys that match records in the database. In the latter case, the matching of the item key also allows the automatic extraction of related metadata, so that, for example, information about the stratigraphic context of the find can also be included in the metadata record for the image file.

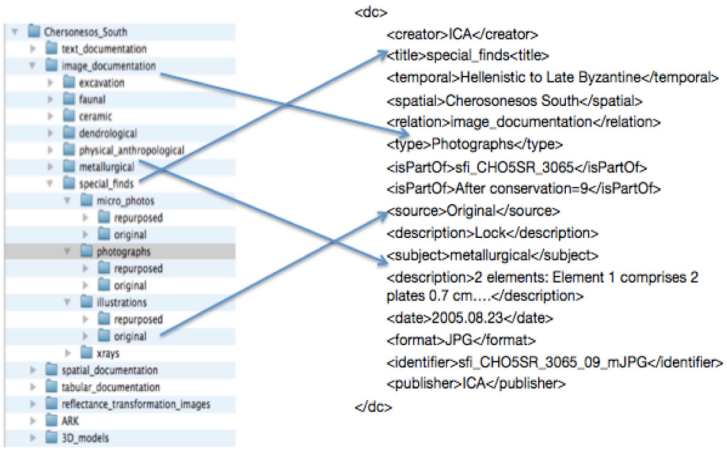

Figure 1. Hierarchical directory structure devised to classify ICA data.

The hierarchical directory structure (**Figure 1**) serves to categorize the data as it is gathered and produced during the different research stages. Top-level directories are labelled according to documentation type. For each type, the sub-directories within reflect the materials to which that type of documentation is applied, and then the different kinds of documentation that are generated during the analysis, interpretation and publication of these materials. When a given data object is placed in a particular folder and accessioned into iRODS, metadata reflecting all the other classifications implicit in the directory path leading to that folder is extracted and mapped automatically to the DC "subject" element.

Technical metadata from the files is extracted using FITS and mapped to PREMIS. A METS document is generated to contain all the metadata schemata. As a result, each object stored in the recordkeeping system on iRODS has a METS metadata record. The descriptive metadata is also registered on the iRODS metadata catalogue to enable search and retrieval through different available iRODS interfaces. By virtue of the object code, and the relationships established by the "isPartOf"

element in DC, the resultant metadata acts as a glue that maintains the relationships between files and the components of the archaeological record and tracks changes to these files across different research stages. Most importantly, these metadata records ensure that the contextual relationships between records, and records and files, are not dependent on the survival of the original database for long-term preservation.

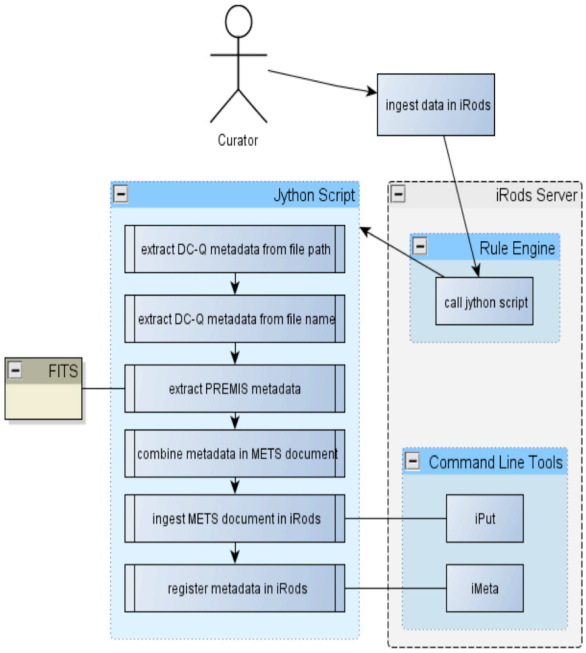

Figure 2. Workflow for automated metadata extraction using iRODS rules.

The implementation of the processes for metadata extraction and mapping involved the creation of the metadata extractor script and the integration of this script into the iRODS rule engine, which manages the entire process. The workflow is exemplified in **Figure 2**. As objects are ingested to the collection, a Jython script is called to manage numerous sub-tasks for metadata manipulation. In turn, other rules enable controlled manipulation of data objects including removal, renaming, or relocation in the context of active collection management or in cases of misclassification of a file, while at the same time enforcing administrative approval for certain tasks. As a whole, the system constitutes a semi-automated platform that allows both the active management of the data objects in the collection and the generation of rich, contextual, continually updated metadata for each object.[15]

2.4. Problems and unresolved issues

After two years of using and testing this system, we have been able to evaluate its benefits and identify what needs to be improved from both human and systems perspectives. While the record-keeping system still requires archaeologists to manually name files and place them appropriately within the directory hierarchy, it automates the rest of the metadata generation process and provides, in conjunction with information retrieved from an external database, more extensive descriptive documentation for each file. Since naming and classification are manual steps, however, the extraction process must account for human error. Inconsistencies in the application of a file naming convention will cause the ingest script, which can only document relationships on the basis of perfectly-formatted file names, to omit descriptive

[15] Walling and Esteva, "Automating the extraction of metadata."

DC metadata. Poorly named files will thus be provided with a METS document containing only a PREMIS record. Furthermore, the system has a limited capacity to deal with files that lack 1:1 relationships with records in the database. For example, an image of a group of pottery fragments from several different stratigraphic contexts can only be associated by its file name with one of those contexts. Also, because the metadata is constructed automatically, the script must be manually customized to include collection-level metadata that applies to all digital objects, like project title, creator, or spatial and temporal coverage. All these issues create some barriers for the general application of this strategy.

3. Toward more broadly applicable solutions: excavation and survey data from Metaponto, Italy

3.1. The challenge of a new, more heterogeneous dataset

Our team spent a large amount of time designing a recordkeeping system tailored specifically to one excavation project, but the Chersonesos dataset represents only a small fraction of ICA's total digital collection. In fact, the ICA archives contain multiple datasets representing a series of projects carried out over more than three decades of research, each with its own characteristics and range of diverse data types. The bulk of ICA's collection is composed of information produced in the course of both excavation and archaeological survey in South Italy, especially in the rural territory of the ancient Greek site of Metapontion (modern Metaponto), from the early 1970s to the present. In the last ten years, many of the original analogue records produced by these projects, including maps and plans, negatives, slides, and paper recording sheets, have been digitized on an expedient and non-systematic basis. As a result, a large proportion of this analogue documentation now also exists in a variety of digital formats, from GIS shapefiles to scanned image files to word-processing documents. A significant number of these data lack sufficient metadata and organization, making it extremely difficult for researchers to work efficiently with them and almost impossible to share them with the larger public. While several small databases have been created by various specialists or eager students over the years to manage specific subsets of this material, especially images, most of the files are not associated with a database of any kind and almost none of them are documented with reference to their original provenance in the same detail as the Chersonesos data in ARK. While the Chersonesos excavation and documentation methodology relied explicitly on the individual stratigraphic context as its central relational unit, ICA's earlier excavations were organized around much more vaguely defined excavation areas. Furthermore, in the case of its survey datasets, which reflect entire ancient landscapes as opposed to structures, the central unit is the "archaeological site". This very different body of digitized data provided an excellent test-case for the expansion of the metadata-extraction strategies we deployed for the Chersonesos collection to datasets with different, looser structures and less extensive documentation.

The challenges to the integration of metadata from the different systems in which the Chersonesos dataset had been created and stored throughout the research process were, as we suspected, exponentially greater for the Metaponto datasets. Our solution for Chersonesos involved a high degree of automation, but it also required rigid and consistent rules for naming and for classifying the digital objects. This has made it very difficult to replicate the solution for the Metaponto collection, much of which was generated over a long period of time prior to the involvement of the current ICA research team and therefore presents little consistency in naming or classification. It is likely that a rules-based automated solution would pose even more significant barriers to adoption by archaeological projects at other institutions,

which would need not only to develop and adhere to their own strict naming conventions, but also to be able to implement an iRODS storage infrastructure and customize ingest scripts to match their own conventions and collection details. All of this would require a team of highly specialized software developers and metadata specialists that most institutions are unlikely to have in place.

3.2 Visualization-based strategies for triage

Our current goal, then, is to explore methods to capture standardized and complete metadata from heterogeneous and idiosyncratic archaeological collections without having to impose on them a rigid top-down system, and without depending on the ingest of datasets into a particular storage infrastructure. The first step in this exploration has been the identification of better methods for the initial triage of complex collections. As we discussed above, inventorying and organizing the Chersonesos collection took a significant amount of time and effort, even though the team was already intimately familiar with its contents, having been involved in the data production and curation from the beginning of the project. To process larger, more disorganized, and less familiar collections at ICA more efficiently, therefore, we are using and extending a visualization tool developed through a National Archives and Records Administration (NARA) research collaboration.[16]

The visualization application was developed for purposes of helping curators explore large and heterogeneous datasets with which they are not necessarily familiar. The tool allows for the visual exploration of a collection's structure according to several different organizing principles: its directories and sub-directories, its main data types, how many files are duplicated or contain errors and where these are located. It also makes it possible to visualize directory labels and file names as tag clouds, so that the user can quickly identify the most frequent names or labels in the dataset. This information in turn allows users to make inferences about the collection's contents and how they were generated and organized, and to learn about its characterization information -- specifically, file format information and the corresponding preservation risks.[17] Ultimately, this makes it possible to approach diverse data in the form of more manageable aggregated groups and facilitates long-term preservation decisions about what needs to be re-organized, labelled, re-named, or deleted if exact duplicates exist.

This phase of our collaboration began with the transfer to a single server of 1,370,000 files from the Metaponto collection, originally dispersed over several separate personal computers and external storage devices. The data, haphazardly organized depending on the computer or storage device of origin, were kept in their original order with reference to the original directories.

After the files were consolidated on a single server, DROID (the Digital Record Object IDentification tool developed by the UK National Archives)[18] was run over the entire collection to identify file formats and checksums for the enormous number of digital objects it contains. Through a complementary script, the directory path to each file, the date of its last modification, and its file size were also obtained for all objects in the collection. These metadata could then be pre-processed and loaded into the visualization platform, allowing large amounts of information to be aggregated at more manageable levels. Different

[16] Weijia Xu et al., "Analysis of large digital collections with interactive visualization," in *Proceedings of the IEE Conference on Visual Analytics Science and Technology, Providence Rhode Island, U.S.A. October 23 – 28* (2011), accessed September 14, 2012, doi: 10.1109/VAST.2011.6102462.
[17] Esteva et al., "Assessing the preservation condition of large and heterogeneous electronic records collections with visualization," *International Journal of Digital Curation*, 6:1 (2011), doi:10.2218/ijdc.v6i1.171.
[18] "DROID," accessed September 14, 2012, http://droid.sourceforge.net/.

file formats, for example, are classified according to PRONOM's[19] file format classification criteria: .jpg and.tif/.tiff, together with other known vector and raster file types, are classified as images, while .rtf, .doc and .docx files are classified as word-processor documents. In turn, checksums are processed to identify duplicates, and dates are aggregated per a number of years specified by the user. Different interactive functionalities make it possible to aggregate or select different metadata values, which can then be visualized in a way that allows the viewer to make inferences about the collection's contents and organization.

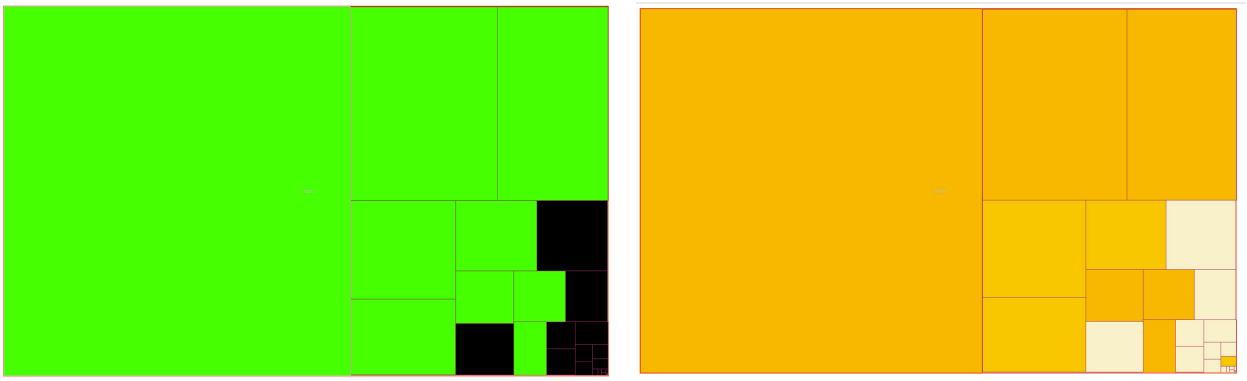

Figure 3. Location of administrative vs. research data using visualization. The image on the left shows ICA's GIS data identified via the aggregator function in the visualization. On the right is a view of directories containing a combination of GIS, image, and database data using the selector function.

The pre-processed metadata were rendered visually as a treemap to show the collection's overall structure and the distribution of different file classes across the structure. The most obvious example of this is the use of the visualization tool to separate general administrative documents from research data (**Figure 3**). This was an important step towards identifying sections of the collection that are of highest priority for reorganizing. Using the aggregator function of the visualization (on the right in **Figure 3**), the user can find all the directories containing GIS data, which -- since GIS files and administrative documents almost never occur together -- allows in turn the implicit identification of directories that might be primarily administrative, and thus of low priority for reorganization. In this visualization, the user can see that GIS files are distributed in most directories, with the exception of the ones highlighted in white. By the same token, using the selector function (on the left in **Figure 3**) the user can visualize directories containing a very low proportion of image, database, and GIS files, which are the main file classes generated by staff whose primary function is research. That the directories in white in the first visualization appear again as directories with few image or database files (in black) in the second confirms the impression that those directories are not focused on research data. Conversely, the directories in which the largest proportion of files were pdf, email, word-processing and other text documents were deemed most likely to be administrative. The use of the visualization tool therefore allows the user to identify quickly and efficiently those parts of a collection than can safely be assigned a low priority for further intervention (in this case, administrative files as opposed to research data).

The visualization tool was deployed in this case to allow researchers involved in the Metaponto data triage to explore and make sense of a dataset generated over a long period of time by a large and

[19] "PRONOM," accessed September 14, 2012, http://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

revolving number of staff, specialists, and students. This situation is not uncommon for large academic archaeological projects that are carried out over long periods of time and with a high turnover in personnel. Visualization tools that provide an efficient and comprehensive way to understand what a disorganized collection is composed of, and how it might be reorganized, are therefore likely to find an eager user base among archaeologists. Using as our starting point the characteristics of ICA's data and the priorities identified in the initial triage of the Metaponto collection, we are currently developing new visual analysis functions that we hope will be applicable to other humanities collections as well. The new functionalities will allow us to re-establish control over the data, discard data that are redundant, and derive new information about the collection that can be integrated into subsequent documentation and preservation strategies.

4. Automated documentation, semantic web, and complex humanities collections: goals for the future

The Metaponto test case is itself a preliminary stage for a larger project with broader applications for other humanistic disciplines. It has become increasingly evident to us, as we have worked on this material, that most scholars are generating large quantities of digital information, much of which might turn out to be another scholar's research data in the future, and most have neither the time nor the skill-sets that will allow them to document their collections for preservation and reuse. In order to address this problem we are planning to re-engineer the tools that we used in our project to make them more flexible so they can be integrated to different standards and infrastructures. Semantic web principles will be an important part of this effort, since they offer the flexibility and potential for interoperability that are crucial in current efforts to ensure that digital data remain accessible and reusable.

The TACC-ICA team is therefore now concentrating its efforts on the development of plans for a project to create a desktop toolkit for the generation, extraction, and management of metadata for both structured and unstructured digital collections. This project will allow us to generalize our own experiences with archaeological datasets to the broader world of digital humanities collections. The basic functionality of the toolkit is a direct offshoot of our collaboration, on both a human and a technical level: simply put, it is meant to act as a translator between the disciplinary language of the collection creator and that of the digital archivist. It will allow the user to describe a collection intuitively, according to the intellectual framework and concepts with which he is familiar. At the same time, the software will encode those concepts, together with technical metadata extracted from the files, as structured metadata in formats such as XML that will be compatible with existing schemata and standards used by digital archives. Moreover, because the toolkit will allow researchers to manage their files and easily create standardized metadata both during and after the research process, it will ensure that data are more thoroughly and consistently documented from the beginning.

The toolkit will be built on the open-source ontology editor Protégé-OWL,[20] which will act as the back-end of a graphic user interface that will allow users to create organizational structures using the visual metaphor of the "bucket" [21]to replace the formal ontological concepts of classes and instances. To the user, the toolkit will present a seamless Graphic User Interface (GUI), in which he can extract

[20] "The Protégé Ontology Editor and Knowledge Acquisition System," accessed September 14, 2012, http://protege.stanford.edu/.

[21] Susan Cisco, "Big buckets for simplifying records retention schedules," ARMA International's Hot Topics. Trimming your Bucket List. Supplement to the Information Management Journal (2008), 3-6.

metadata from his collections and drag and drop his files into user-defined categorical "buckets". For example, an oral history institute could organize its collection by individual oral history cases or according to broader classes such as regions, localities, themes, etc. If a user wants to capture the process history of individual files, on the other hand, he could begin by creating buckets that reflect research stages, and within those, other buckets that represent units of observation. In the background the tool captures the bucket designations, and the relationships between buckets, in a standardized format within a formal ontology.

Behind the scenes, a mapping framework will allow a bucket labelled by a user as a unit of observation (e.g. stratigraphic context, oral history project, or work of art) or as a research stage (data gathering, analysis, interpretation) to be related to a standard metadata element in the ontology. In this way, the specialized concepts of ontology building are disguised for the user, who instead simply creates an intuitive organizational structure that is consistent with his own research workflow. To address the authenticity and integrity of the managed collection in compliance with archival and preservation requirements, the toolkit will create notations in the underlying ontology that record the presence, location and deletion of or change to the files in the collection as PREMIS OWL[22] events when the user arranges and rearranges those files in the buckets. This document will also include information such as checksums, file format identification and dates of most recent modification, which will be extracted automatically from the files themselves. Technical metadata extracted from the files will be recorded in PREMIS OWL to provide information necessary for their long-term preservation. As end products, the ontologies built with the software will be exportable as XML/RDF and other metadata standards. In this way, at the end of the research project, both data and metadata can be integrated to digital repository infrastructures such as Fedora, which also uses OWL to establish relationships between properties of objects and other objects.[23]

Such a toolkit will help to bridge the gap between data producers and repository managers by making it cheaper and easier for the producers to meet the curators halfway. Putting automated metadata creation tools in the hands of those who are responsible for digital datasets of any scale right now, without imposing rigid standardization or insisting on ingest into a specific repository, will make it simpler to apply repository-based preservation and dissemination solutions in the future. These tools will be especially important for digital collections in developing countries, where more permanent long-term preservation infrastructure does not yet exist. In the case of many digital humanities and cultural-heritage collections, the digital "memory of the world" is actually composed of the very human memories of individual researchers, many of whom have not yet had time to encode the knowledge in their heads as formal metadata for their digital collections. Without tools to facilitate the creation of such metadata, the world's digital memory will diminish, little by little, with the inevitable loss of the human memories of each of those individuals. Where the digital record is the primary record of research, as it is increasingly in field archaeology, the resulting amnesia will be absolute and permanent. It is critical, then, that the digital preservation community address not only the preservation and dissemination of documents through digitization, but the preservation of meaning in the digital documents themselves.

---

[22] "Public workspace for PREMIS OWL ontology," accessed September 14, 2012, http://premisontologypublic.pbworks.com/w/page/45987067/FrontPage.

[23] For example, the discussion of the description of relations using OWL in the Fedora ontologies wiki: "Ontologies - Fedora Repository Development," accessed September 14, 2012, https://wiki.duraspace.org/display/FCREPO/Ontologies.

References

American School of Classical Studies at Athens. "Digital Collections, The American School of Classical Studies at Athens." Accessed September 14, 2012. http://www.ascsa.net/research?v=default.

Archaeology Data Service and Digital Antiquity. "Archaeology Data Service/Digital Antiquity Guides to Good Practice." Accessed September 14, 2012. http://guides.archaeologydataservice.ac.uk/g2gp/Contents.

Archaeology Data Service. "Archaeology Data Service Homepage." Accessed September 14, 2012. http://archaeologydataservice.ac.uk/.

International Council of Museums. "The CIDOC Conceptual Reference Model." Accessed September 14, 2012. http://www.cidoc-crm.org/.

Cisco, Susan. "Big buckets for simplifying records retention schedules." *ARMA International's Hot Topics. Trimming your Bucket List. Supplement to the Information Management Journal* (2008): 3-6.

Digital Antiquity. "tDAR." Accessed September 14, 2012. http://www.tdar.org/.

Esteva, Maria, Christopher Jordan, Tomislav Urban, and David Walling. "Cyberinfrastructure supporting evolving data collections." In *Proceedings of the 8th International Conference on Preservation of Digital Objects, iPRES 2011, Singapore, November 1-4 2011*, edited by J. Borbinha, A. Jatowt, S. Foo, S. Sugimoto, C. Khoo, and R. Buddharaju, 93-96. Singapore: National Library of Singapore and Nanyang Technical University, 2011.

Esteva, Maria, Jessica Trelogan, Adam Rabinowitz, David Walling, and Stephen Pipkin. "From the site to long-term preservation: a reflexive system to manage and archive digital archaeological data." In *Archiving 2010. Proceedings of the Archiving Conference, vol. 7*, 1-6. Society for Imaging Science and Technology, 2010.

Esteva, Maria, Weijia Xu, Suyog Jain Dott, Jennifer Lee, and Wendy K. Martin. "Assessing the preservation condition of large and heterogeneous electronic records collections with visualization." *International Journal of Digital Curation* 6:1 (2011). Accessed September 14, 2012. doi:10.2218/ijdc.v6i1.171.

Eve, Stuart and Guy Hunt. "ARK: a developmental framework for archaeological recording." In *Layers of Perception: Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), Berlin, Germany, April 2-6, 2007*, edited by Axel Poluschny and Karsten Lambers. Bonn: Deutsches Archäologisches Institut; Dr. Rudolph Hablet Gmbh, 2008.

Fedora. "Ontologies - Fedora Repository Development." Accessed September 14, 2012. https://wiki.duraspace.org/display/FCREPO/Ontologies.

iRODS. "IRODS Data Grids, Digital Libraries, Persistent Archives and Real-time Data Systems." Accessed September 14, 2012. https://www.irods.org/index.php/IRODS:Data_Grids,_Digital_Libraries,_Persistent_Archives,_and_Real-time_Data_Systems.

Library of Congress. "Library of Congress Metadata Schemas." Accessed September 14, 2012. http://www.loc.gov/standards.

National Archives (UK). "PRONOM." Accessed September 14, 2012. http://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

National Preserve of Tauric Chersonesos. "Discovering Chersonesos." Accessed September 14, 2012. http://www.discovering.chersonesos.org.

National Preserve of Tauric Chersonesos. "K. K. Kostsyushko-Valyuzhinich and his Reports for the Imperial Archaeological Commission." Accessed September 14, 2012. http://kostsyushko.chersonesos.org/.

University of Chicago. "Online Cultural Heritage Research Environment." Accessed September 14, 2012. http://ochre.lib.uchicago.edu/.

PREMIS Ontology. "Public workspace for PREMIS OWL ontology." Accessed September 14, 2012. http://premisontologypublic.pbworks.com/w/page/45987067/FrontPage.

Protégé. "The Protégé Ontology Editor and Knowledge Acquisition System." Accessed September 14, 2012. http://protege.stanford.edu/.

Rabinowitz, Adam, Stuart Eve, and Jessica Trelogan. "Precision, accuracy, and the fate of the data: experiments in site recording at Chersonesos, Ukraine." In *Digital Discovery: Exploring New Frontiers in Human Heritage, CAA 2006*, edited by Jeffrey Clark and Emily Hagemeister, 243-256. Budapest: Archeolingua, 2007.

Rabinowitz, Adam, Carla Schroer and Mark Mudge. "Grass-roots imaging: a case study in sustainable heritage documentation at Chersonesos, Ukraine." In *Making History Interactive: Proceedings of CAA 2009*, edited by B. Frischer and L. Fisher. Budapest: Archaeolingua, 2010.

Saffady, William. *Managing Electronic Records*, 4th ed. Lenexa, KS:ARMA International, 2009.

SourceForge. "DROID." Accessed September 14, 2012. http://droid.sourceforge.net/.

Walling, David and Maria Esteva. "Automating the extraction of metadata from archaeological data using iRODS rules." *International Journal of Digital Curation* 6:2 (2011). Accessed September 14, 2012. doi:10.2218/ijdc.v6i2.20.

Xu, Weijia, Maria Esteva, Suyog Dutt Jain, and Varun Jain. "Analysis of Large Digital Collections with Interactive Visualization." In *Proceedings of the IEE Conference on Visual Analytics Science and Technology, Providence Rhode Island, U.S.A. October 23 – 28* (2011). Accessed September 14, 2012. doi: 10.1109/VAST.2011.6102462.