Management of Digital Archives for Integrated Web Access to Scientific and Cultural Information

Michael B. Toth; R. B. Toth Associates; Oakton, Virginia/USA

Abstract

Digital imaging of the Archimedes Palimpsest poses a complex set of data archiving challenges. The thousand-year-old manuscript contains the earliest known copies of some of Archimedes unique mathematical works, overwritten with a book of prayer. Integrating the ancient Greek transcriptions of Archimedes' mathematical texts with digital images, and hosting them on the Web or in digital media for a broad set of global users offers a complex set of digital information storage and sharing challenges. Given the fragility of the manuscript, the product of this effort must be available as a digital repository over the next millennium for further research and study. Unique to document imaging efforts is the requirement for spatial metadata to allow linkages across content types to specific images and locations within the Palimpsest images. This is necessary not only to register locations on the same section of a manuscript leaf in various spectral bands, but also to link to the original images the various x-ray fluorescence images, conservation drawings, image data and transcriptions. Key issues also addressed in the metadata include rights and intellectual property management to ensure this information from a privately funded imaging effort is broadly available in the public domain. Linkage of Archimedes Palimpsest information from various content types is the ultimate goal of the program. Just as geospatial data can be linked to images from earth resource satellites, the Archimedes Palimpsest team is linking a range of "scriptospatial data" with images from a range of imaging devices. With standardized metadata, data points in these images can be linked to information derived from the images, including conservation, scientific and scholarly information, in a useful digital image archive.

Archimedes Palimpsest Program

The Archimedes Palimpsest Program is an ongoing, multi-year effort to produce a digital archive of seven treatises of Archimedes in Greek text, as originally written on parchment in the latter half of the 10th century. In the early 13th century, this text was scraped off and overwritten, or "palimpsested," with Greek text to create a prayer book [1]. Beginning in 1999, under the auspices of the Walters Art Museum in Baltimore, Maryland, USA, a team of scientists, conservators and scholars has been disbinding, conserving, imaging, analyzing, transcribing, translating and studying the 174 fragile parchment folios that make up the Archimedes Palimpsest.

Using a range of imaging techniques, the imaging team has produced over 500 Gigabytes (GB) of digital data, yielding images of unique Archimedes diagrams, and the only copies of Archimedes' treatises *The Method* and *Stomachion*, the only copy in Greek of *On Floating Bodies,* and copies of the *Planes in Equilibrium*, *On Sphere and Cylinder, Spiral Lines*, and *Measurement of the Circle* [2]. They have also imaged ten folios of text by the fourth-century B.C. Attic Greek orator Hyperides; six folios from a Neo-Platonic philosophical text; four folios from a liturgical book; and twelve folios from two other books which have yet to be deciphered [3]. The Archimedes Palimpsest team also imaged prints of original photographs of the Archimedes Palimpsest taken almost 100 years earlier in Constantinople at the direction of Johann Ludwig Heiberg. These photographs offer standardized images of text that have since been lost to mold or other damage, and of one leaf that has been lost in its entirety. They also pose an interesting metadata contextual challenge, since they refer not to the original object, but to data derived from the object many years prior to the current effort.

Key to management of a usable digital archive of images and data derived from the Archimedes Palimpsest was the architecture of an end-to-end system that would integrate conservation, digital imaging, scholarly study, and data storage. This required attention to details of the data content early in the program to ensure compatibility and integration throughout the program

Digital Imaging

Central to the collection of digital information from the original Archimedes text is the multispectral digital imaging of the ancient manuscript. To read the underlying ancient Greek text that was overwritten orthogonally with Greek prayer text called the "Euchologian" in similar iron gall ink, the imaging scientists identified different spectral signatures between the two inks, as well as the underlying parchment. The red/green/blue (RGB) images were taken under ultraviolet (UV) and visible illumination to produce the spectral differences needed to digitally reveal the underlying Archimedes text (Fig. 1) [4].

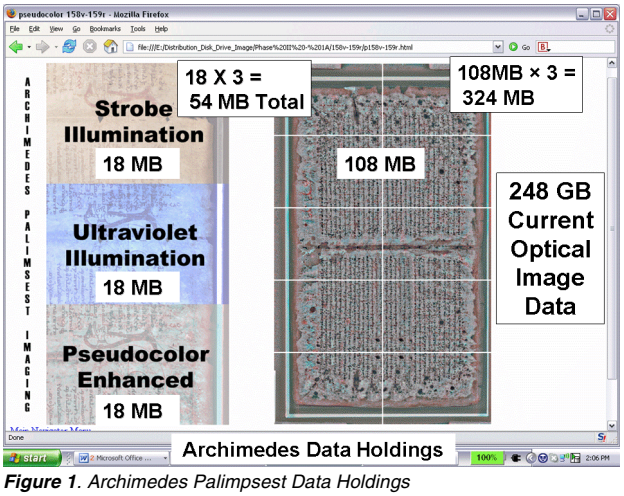


Figure 1*. Archimedes Palimpsest Data Holdings*

These digital images were collected during a series of "production" imaging sessions from 2000 to 2006, in conjunction with the disbinding and conservation of the original manuscript. These images were then digitally processed to produce false color, or "pseudocolor," images of the underlying Archimedes text. The images were taken with a spatial resolution of 600 dpi, which required ten sectional images of each page, with each image stored as an 18 Megabyte (MB) Tagged Image File Format (TIFF) file. They were then digitally stitched to create full-size images of the manuscript leaf, with a file size of about 108 MB that have been used by scholars for study and transcriptions. Specialized and "experimental" imaging techniques were developed to digitally image key enigmatic areas in response to feedback from the scholars. With advances in imaging technology, efforts are now underway to reimage all Palimpsest leaves with a large format digital camera during a single session in 2007. This will yield images that will not require stitching that will be added to the archive, easing integration and spatial registration challenges. Thanks to Moore's Law, with the growth of storage capacity the entire digital image archive has consistently fit on a single hard drive. The data is then replicated and distributed to scholars and scientists on duplicate hard drives for study and disaster recovery.

Four of the palimpsested manuscript leaves were also overlaid with twentieth-century forged paintings, similar to those in illuminated manuscripts, and the original text on other leaves was obscured by mold or excessive scraping. A joint team of imaging scientists and x-ray physicists conducted x-ray florescence (XRF) studies of these leaves. These studies resulted in almost 300 data sets with up to twelve channels of data from some collection sessions. Each scan of an area only 40 mm × 20mm produced an ASCII data file of approximately 32 MB. Each ASCII file was converted into twelve TIFF images (one from each elemental detector), each approximately 3 MB in size. Scans of adjacent areas were then stitched to form larger images of critical portions of the forgery leaves with underlying text, which were added to the digital archive (Fig. 2).

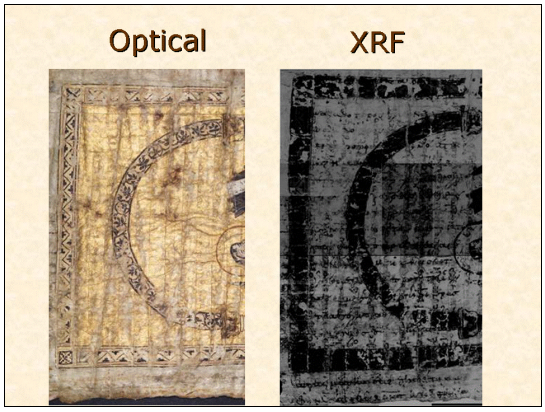


***Figure 2**. Comparison of Optical and XRF Imaging*

Data Management

The Archimedes Palimpsest is proving to be a key resource for a variety of users ranging from image scientists to classics scholars. The goal of the Archimedes team has always been to make as much digital source data as possible available to all users and the general public for study and collaboration. Management of access to original images of Archimedes text has depended on fully integrated metadata developed with input from all members of the team, especially feedback from the end users: the scholars studying the text and producing the transcriptions.

Since the beginning of the imaging effort, the Archimedes team used spatial context to link all data, with special focus on images and transcriptions of the Greek text. Just as geospatial data is linked to images from earth resource satellites, the Archimedes Palimpsest team has linked a range of what the team refers to as "scriptospatial data" with images from optical cameras and XRF detectors [5]. The optical images were all taken with the palimpsest leaf mounted on the stage of a computer controlled Velmex X:Y table, with the spatial parameters recorded for each stage (Fig. 3). A similar setup was used for the XRF scanning, with the Palimpsest leaves mounted vertically and at an angle to the x-ray beam.

With over 5,000 digital images of the Archimedes Palimpsest, the team developed robust metadata standards to ensure the data remained manageable and accessible, and that a point on any image can be registered to corresponding points on other images. The data manager is currently validating the archived image data and metadata, and integrating the image archive with the scholarly transcriptions to provide an integrated image product, which is to be released in Beta form in the spring of 2008.

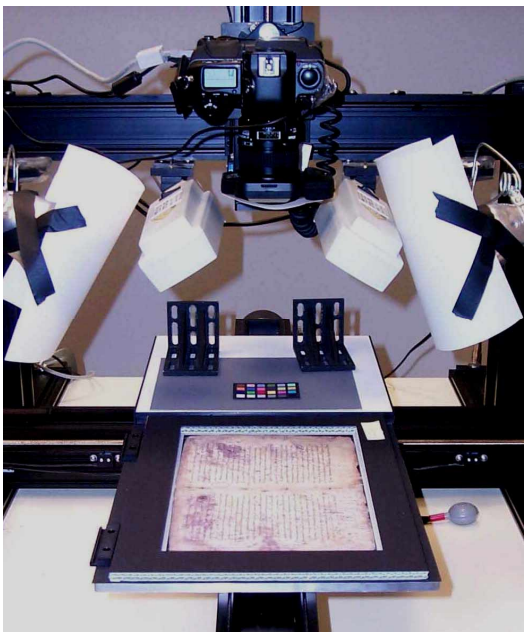
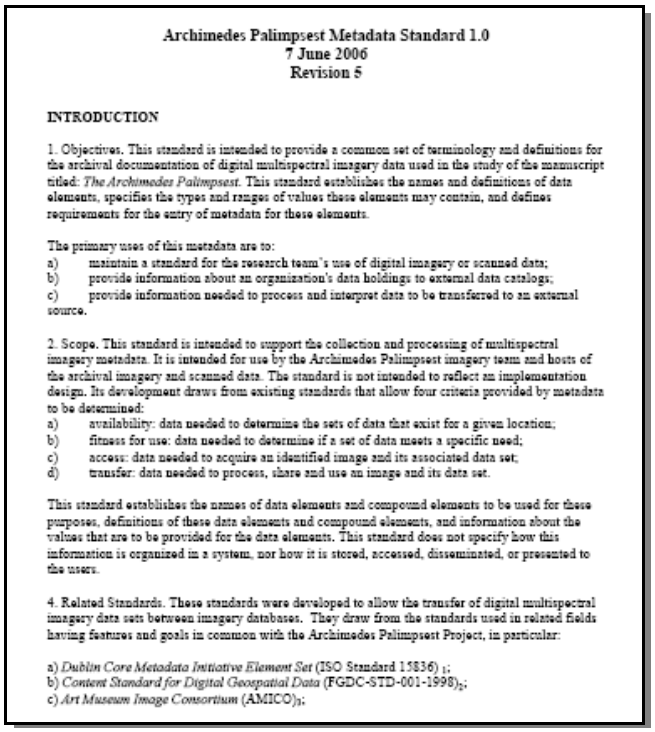


***Figure 3**. Archimedes Palimpsest Optical Imaging Setup*

Metadata Standard

The *Archimedes Palimpsest Metadata Standard* reflects the unique requirements of digital manuscript imaging [6]. Dublin Core metadata elements are used for the key identification elements required for image storage, management and retrieval, as well as to ensure all information remains in the public domain for future study. Additional spatial information was critical to the process of registering and overlaying images, stitching smaller images to yield larger, high-quality images of full manuscript leaves, and correlating image content across various media. As the program progressed, the *Archimedes Palimpsest Metadata Standard* was refined to keep

pace with new data needs, with the better understanding of both the task and the data, and to account for new imaging and processing methods. During the multiyear program, advancing technologies for data processing, storage and access have changed the priority and need for some data elements over time. The program team reviewed and revised the metadata standard at key milestones in the program based on data requirements and user needs. Metadata was captured during the imaging process by the imagers and scientists in spreadsheets and databases, and are now being integrated into the data products, including TIFF headers and ASCII files. The team is currently hosting the standard and sample integrated products on the program website at archimedespalimpsest.org, and welcomes reviewer comments from interested parties (Fig. 4) [7].



Archimedes Palimpsest Metadata Standard 1.0
7 June 2006
Revision 5

INTRODUCTION

1. Objectives. This standard is intended to provide a common set of terminology and definitions for the archival documentation of digital multispectral imagery data used in the study of the manuscript titled: *The Archimedes Palimpsest*. This standard establishes the names and definitions of data elements, specifies the types and ranges of values these elements may contain, and defines requirements for the entry of metadata for these elements.

The primary uses of this metadata are to:
a) maintain a standard for the research team's use of digital imagery or scanned data;
b) provide information about an organization's data holdings to external data catalogs;
c) provide information needed to process and interpret data to be transferred to an external source.

2. Scope. This standard is intended to support the collection and processing of multispectral imagery metadata. It is intended for use by the Archimedes Palimpsest imagery team and hosts of the archival imagery and scanned data. The standard is not intended to reflect an implementation design. Its development draws from existing standards that allow four criteria provided by metadata to be determined:
a) availability: data needed to determine the sets of data that exist for a given location;
b) fitness for use: data needed to determine if a set of data meets a specific need;
c) access: data needed to acquire an identified image and its associated data set;
d) transfer: data needed to process, store and use an image and its data set.

This standard establishes the names of data elements and compound elements to be used for these purposes, definitions of these data elements and compound elements, and information about the values that are to be provided for the data elements. This standard does not specify how this information is organized in a system, nor how it is stored, accessed, disseminated, or presented to the users.

4. Related Standards. These standards were developed to allow the transfer of digital multispectral imagery data sets between imagery databases. They draw from the standards used in related fields having features and goals in common with the Archimedes Palimpsest Project, in particular:

a) *Dublin Core Metadata Initiative Element Set* (ISO Standard 15836) [1];
b) *Content Standard for Digital Geospatial Data* (FGDC-STD-001-1998)[2];
c) *Art Museum Image Consortium* (AMICO)[3];

Figure 4*. Archimedes Palimpsest Metadata Standard*
http://www.archimedespalimpsest.org/programmanage_documents.html

The *Archimedes Palimpsest Metadata Standard* cites six types of metadata elements:
1. Identification Information
2. Spatial Data Reference Information
3. Imaging and Spectral Data Reference Information
4. Data Type Information
5. Data Content Information
6. Metadata Reference Information

The information on "Identification," "Data Type" and "Data Content" is based on Dublin Core elements [9]. The "Spatial Data Reference" and "Imaging and Spectral Data Reference" comprises metadata unique to the Archimedes Palimpsest Program, which builds on geospatial metadata elements detailed in the Federal Geographic Data Committee *Content Standard for Digital Geospatial Metadata*, building on the analogy that a camera collecting images over the manuscript is similar to a satellite collecting geospatial data over the Earth [8]. Finally, the "Metadata Reference Information" addresses metadata information about the standard itself, as well as any metadata extensions for metadata required for specific aspects of the Archimedes Program. Each of these element sets posed challenges in defining the unique data elements of the Archimedes Palimpsest.

The XRF imaging of the Archimedes Palimpsest required a unique set of metadata extensions to capture the new data elements for different imaging techniques, energy levels and data formats. The numerous variables in position around the manuscript, high energy levels and elemental information had not been anticipated during the early optical imaging phases of the program, requiring significant additional metadata in the XRF Extensions. The XRF team reviewed the metadata elements and developed the current draft of the *Archimedes Palimpsest Metadata Standard XRF Extensions* [10].

Scriptospatial Linkage

Unique to document imaging efforts is the requirement for spatial metadata to allow establishment of linkages across data domains to specific images and locations within images. This is necessary not only to register locations on the same section of a manuscript leaf in UV and RGB images, but also to link XRF images, conservation drawings, Heiberg image data (if available) and the transcriptions (and ultimately translations) to the original images. Establishing spatial reference points required the establishment of a coordinate system for the manuscript leaves, beginning with an imaging standard that the Archimedes text would be upright on the image and run horizontally from left to right. Using this coordinate system, the team collected spatial metadata for the upper left and lower right coordinates on the manuscript for each the image. Using this system, transcriptions of the text in XML format are being tagged with spatial coordinates to link lines of transcription with the respective lines of underlying text in the digital images. This will ultimately allow the management of an integrated product with links between data in various formats and from a range of sources.
With standardized metadata and XRF tagging, data points from the original manuscript in these images can be linked to derived information. Once the metadata and scriptospatial data is distributed to information providers for hosting on the Internet, the source images and related information will be accessible by a larger group than the initial team of scholars and researchers. Making the source data accessible to a broad range of interested scholars will ensure the information derived from this data – including transcriptions and translations – can be validated and further studied from the original images by a larger pool of scholars with the appropriate knowledge and expertise. With the source data, scientists from a range of disciplines will have the opportunity to apply new image processing algorithms and refine the images with new and more powerful software and hardware. The origins of modern science and mathematics can then be studied from the original text, without risking damage to the fragile manuscript from handling and exposure.

Data Integration and Archiving

Integrating the Archimedes Palimpsest data into a single integrated data archive has required focus on the end product throughout the imaging process. To ensure this product meets the

needs of users, the program managers solicited the input of scholars, users and information providers throughout this effort. Their input was used to refine not only the metadata, but the tools for display of the product and the data format. In 2008, the public data product will be made available to all information providers, ranging from Google to libraries and academic institutions. To ensure utility by as broad a range of applications as possible, ASCII format flat files will be used for the text, and standard TIFF image formats. Use of these broadly accessible formats is intended to ensure the product remains available well into the future in a range of archives, without requirements for proprietary software. This is intended to allow ubiquitous access to the integrated product with host provided GUI's as not only a single package, but also to allow users to link the data to a range of derived products from the original images and transcriptions, including translations, academic studies and mathematical proofs (see hypothetical example in Figure 5).

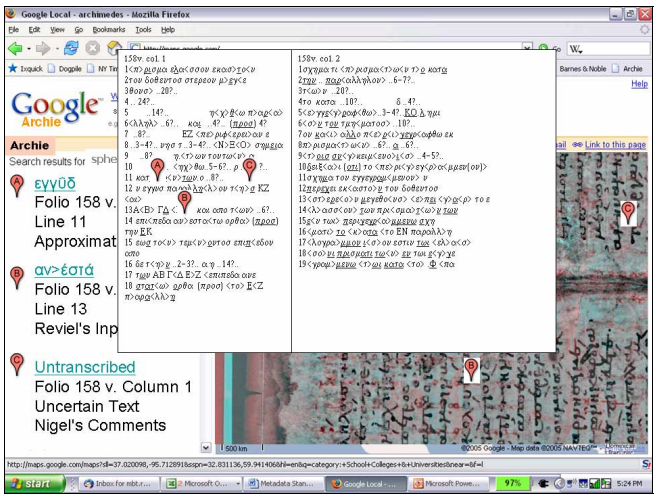


***Figure 5**. Hypothetical Example of Integrated Product*

Conclusion

Archiving digital images of the Archimedes Palimpsest Program is not only dependent on high-quality, standard images, but the management and integration of the images and derived products as a useful body of knowledge. This includes the production images of the entire manuscript and the specialized images used to study enigmatic areas, as well as transcriptions and other scholarly materials. Establishing spatial data points on the images of the two-dimensional manuscript offer a unique linking element for various digital products. The ultimate success or failure of the digital archive is dependent on the establishment of metadata standards early in the imaging process.

Integrating various types of metadata tailored to meet the needs of the users is critical to making unique data available across data domains and disciplines. The *Archimedes Palimpsest Metadata Standard* combines Dublin Core metadata elements with needed new elements from other standards, as well as unique elements not standardized previously. With standards for accurate and validated metadata, including spatial metadata, a range of derived data across multiple data domains can be spatially linked to specific locations on images to offer rich layers of information, such as transcriptions and translations. This metadata serves as a key management tool for the integration of data into a product available to users across the Internet.

With advanced information technology tools available to a broad range of information providers, a growing number of specialized metadata elements could support a broad range of applications across a wide and varied range of data domains. Digital archiving efforts utilizing a range of new techniques and technologies – such as the Archimedes Palimpsest Project, the International Dunhuang Project [11], and various papyri studies – could benefit from an on-line repository of metadata standards as an initial source for unique metadata element definitions useful for a range of digital archiving efforts.

References

[1] Walters Art Museum, *EUREKA! Archimedes Palimpsest at the Walters Art Gallery* [sic], "The Archimedes Palimpsest". <http://www.thewalters.org/archimedes/frame.html> (5 March 2006)

[2] Reviel Netz, "The Origin of Mathematical Physics: New Light on an Old Question," *Physics Today*. pgs 32-37, June (2000)

[3] Walters Art Museum (WAM), *Archimedes – The Palimpsest.* <http://www.archimedespalimpsest.org/> (2 March 2006).

[4] Keith T. Knox, Roger L. Easton, Jr., and William A. Christens-Barry, "Multispectral Imaging of the Archimedes Palimpsest," *2003 AMOS Conference*. Maui, Hawaii: Air Force Maui Optical & Supercomputing Site, Sept. (2003)

[5] Will Noel, Roger L. Easton, Jr., and Michael B. Toth, "The Archimedes Palimpsest," Mountain View, California: Google Inc., 7 March (2006)

[6] Archimedes Palimpsest Program "Archimedes Palimpsest Metadata Standard 1.0X," Revision 3. Baltimore, Maryland: Walters Art Museum, 20 March (2006)

[7] WAM, Archimedes – The Palimpsest, (4 April 2006).

[8] Federal Geospatial Data Committee (FGDC) "Content Standard for Digital Geospatial Metadata: Extensions for Remote Sensing Metadata," FGDC-STD-012-2002. October (2002)

[9] Dublin Core Metadata Initiative, Dublin Core Metadata Element Set, Version 1.1: Reference Description. <http://www.dublincore.org/documents/dces/> (2000-2006)

[10] Archimedes Palimpsest Program "Archimedes Palimpsest Metadata Standard XRF Extensions," Draft. Baltimore, Maryland: Walters Art Museum, 20 March (2006)

[11] IDP International Dunhuang Project, *Technical Infrastructure.* <http://idp.bl.uk/pages/technical_infra.a4d > (13 July 2005)

Author Biography

Michael B. Toth is an independent consultant with R.B.Toth Associates, and the Program Manager for the Archimedes Palimpsest Program since 1999. Mr. Toth brings extensive experience in program management, strategic planning and systems integration with his work on advanced information and space systems for the US Government. Since 1986, he has managed the development, integration and operation of imaging and geospatial information systems. He received his BA in History from Wake Forest University in 1979.