Archimedes Palimpsest Transcription Integration Plan

The Archimedes Data Product brings together all the Archimedes Palimpsest digital information available by the summer of 2008 into a single data set, with a standard structure and format. It serves the following functions:

1. Provide derived information (i.e., transcriptions, processing information) in the context of digital images of the original manuscript in a single integrated package.
2. Serve as the authoritative digital data set in a standardized format that meets the needs of users, information providers, archives and libraries.
3. Offer a standard product to which current or future contributors can add additional standardized information (e.g., alternate texts, image analyses or conservation information).

This was carried out by a team of scholars, imagers and information specialists for release of a final product on October 29, 2008. The Archimedes Data Product was built from the processed set of registered TIFF images for each bifolia, a set of scholarly transcriptions, and existing metadata. The digital production required the integration of imaging, scholarly and data products, within the context of the current program structure, schedule and processes.

The scholarly effort made use of the best efforts of the scholarly community, utilizing standards and work processes broadly used by the community. These provide a simple end product that offers best value to the program and is fully compatible with global practices and standards. The scholars and/or their collaborators (students or fellow scholars working as a team) produced transcriptions using common tools and standards to provide the following end product for integration with the Archimedes Data Product:

XML Standard

Scholars produced XML transcriptions using common tools and standards that conform to the TEI P5 guidelines. Transcriptions were validated against a RELAX NG XML schema based on TEI P5. The RELAX NG schema language has been selected for its ease of use and to conform with work processes in the TEI community where RELAX NG is preferred over DTDs, as in the TEI P5 guidelines.

This transcription standard for the Archimedes Data Product is not intended to limit dissemination and collaboration of the scholars' works, but to provide an end product that meets globally accepted standards within the schedule and budgetary limitations of the current program. If scholars find they need additional flexibility with their transcriptions which are not provided for in the TEI P5 standard, they should propose additional TEI subsets to Professor Patricia Easterling, who will then advise on the utility of the proposed subset with the support and advice of a global team of technical scholars.

Citation Scheme and other mark-ups

The palimpsest's Euchologion foliation serves as the basis for the transcription citation scheme. That is, the structure of each XML transcription is based on Euchologion folios, column (where applicable), and line (e.g. 93v-93r, col 1, line 16). Lines are indicated by the TEI <seg> element to ensure the designation physical divisions in a text. Euchologion folios are secondarily identified by their under-text foliation. For example, 93v-92r is also be identified as Arch65v (Archimedes 65 verso).

Non-linguistic components, such as diagrams and the identification of diagram features by Greek letters, are tagged according to the following standard. Figures are designated by the TEI <FIGURE> element. Non-linguistic text components are marked appropriately, using the <NUM> element for numbers.

Unicode Encoding of Greek Text

The Greek text of the transcriptions are in Unicode Greek characters taken from the Greek (U+0370-U+03FF) and Greek Extended (U+1F00-U+1FFF) code charts provided by Unicode version 5.0. Where a composed character in the Greek extended range is canonically equivalent to a composed character in the Greek basic range, the latter is used.

Combining characters (letters plus accents and breathing symbols) comply with Unicode Normalization Form C (NFC), "Canonical Decomposition, followed by Canonical Composition." This Normalization Form should ensure the broadest compatibility with existing systems and is in compliance with the recommendation of *W3C Character Model for the World Wide Web* (http://www.w3.org/TR/charmod/).

XML Editing Tools

Scholars use a range of tools and work processes as long as they provided transcriptions compatible with the standards defined for the Archimedes Palimpsest Data Product. Many tools are available for creating XML documents, both commercial and freely available. Among these are: Oxygen (commercial), jEdit (free), and Emacs (free). The Informatics team at the Faculty of Informatics, University Eotvos Lorand, Budapest, also developed a tool for the Hyperides transcriptions.

Transcription to Image Mappings

The digital product also includes XML-encoded mappings to regions on the images corresponding to the lines of the transcriptions. These mappings are intended to help readers relate difficult to discern passages and diagrams on the images to the transcriptions and vice versa. The project validated these mappings against the custom tei-analysis-figs.rng schema.

Metadata Standards

The project provides metadata standards for all file types, including transcriptions and transcription-to-image maps. As with the digital images, metadata is provided with the XML files to:

a) maintain a standard for the research team's use of supporting data files;
b) provide information about the dataset to external data catalogs;
c) provide information needed to process and interpret data to be transferred to an external source.

Identification metadata uses the Dublin Core Standard. The image and transcription products are intended to stand alone, as well as link with each other for specific applications and studies. Each is intended to meet the needs expressed by the various users, and to provide an open platform for unfunded activity by those service providers who have expressed interest in building on the distributed dataset. The development of a fully capable, web-based product with GUI's is not envisaged as part of this program.

Standards Cited

- **TEI P5** – maintained by the Text Encoding Initiative Consortium
- **EpiDoc Guidelines** – maintained by the EpiDoc Community
- **RELAX NG** –maintained by Organization for the Advancement of Structured Information Standards (OASIS)
- **Unicode 5.0** – maintained by the Unicode Consortium
- **Dublin Core** – maintained by the Dublin Core Metadata Initiative