Archive Ingest Handling Test (AIHT) Proposal

From Johns Hopkins University

Existing Infrastructure

JHU intends to develop solutions based on DSpace, Fedora, WebWare, and locally-developed programs. Of the three systems, only WebWare is being run in a production mode at JHU. Johns Hopkins News and Information, with consultation from the Sheridan Libraries, coordinated a pilot evaluation of WebWare between April 2003 and October 2003. The Sheridan Libraries tested WebWare with several locally digitized collections. At the end of the evaluation period, News and Information decided to continue using WebWare, while the Libraries felt additional testing would be required. WebWare will require most probably require additional programming to accommodate large-scale ingestion and digital preservation needs. During the initial pilot, WebWare ran the server remotely. In the current production implementation, the server is being managed locally by JHU's own IT department. The Sheridan Libraries have contributed funds to this effort, since we wish to lead the evaluation of WebWare. Based on a conversion with and documentation provided by WebWare's CTO, we believe that WebWare provides the requisite facilities vis-a-vis web services interfaces to support the activities of the AIHT.

We have been running a test instance of DSpace for more than a year. During this time we have performed bulk ingestion of collection with size on the order of 10,000 records. We are currently running version 1.1.1 of DSpace. We have run some instance of Fedora since version 0.9, just before it was made generally available in April 2003. We have not worked with a large number of objects in this environment, but feel that it will be a critical piece of this architecture because of its ability to associate behaviors with objects. Neither DSpace nor Fedora has been deployed outside of the Digital Knowledge Center.

For the purposes of the AIHT, most importantly so that we can guarantee destruction all of provided content at the end of the test and to avoid interfering with our other development activities, we plan to install new instances of DSpace and Fedora on a separate server (or workstation acting as a server).

Preservation Philosophy

The Sheridan Libraries is currently in the process of developing its digital preservation strategy, with collaboration between the Digital Knowledge Center and the Preservation Librarian. For the purposes of this test, however, we will follow these guiding principles:

- Original documents should be preserved, even if they go out of scope.
- Original documents should not be altered in any way
- Possibility of data loss will be mitigated by backups and export to multiple sites.
- Because different organizations may wish to handle the migration/emulation choice in different ways (or a single institution may wish to handle it differently based on collection or format), we should attempt to create an interface that supports both, at least for simple object types.
- Because some formats require server-side processing (e.g., shtml, php, asp), a plug-in architecture supporting at least some of these should be developed.

Execution of Phases I-IV

During the course of the project, JHU will analyze data, develop software and processes, perform tests, and record results to support the activities and reporting requirements laid out in the SOW.

* *Phase I*

Initial ingestion work will depend on format and state of contents on distribution media. Therefore, our first task will be to analyze the distribution media.

The next step will be to develop small suite of tools to process and validate the provided transfer metadata and to generate new transfer metadata. New transfer metadata will be validated against the provided transfer metadata for consistency. The newly generated transfer metadata must be consistent with, but not necessarily identical to that provided. We will transmit to the Library a copy of our locally-generated transfer metadata in the format specified in Appendix A of the SOW.

As mentioned in previous documentation, none of the systems provide mechanisms to track dependencies amongst the datastreams in an archive. This functionality is necessary to guarantee completeness (at least within an archive) when preserving chunks smaller than the entire archive. For example, if one wanted to make a preservation copy of a single web page, it would be useful to store -- in the same package -- copies of image datastreams displayed on that page. More importantly, these structures will be useful for tracking dependencies for elements affected by a format going out of scope.

Because we will be doing research to determine how best to handle this, it is difficult to state with certainty how this will ultimately be accomplished. Because we are interested in leveraging existing tools and ongoing efforts, we plan to initially evaluate the feasibility of using the Metadata Encoding and Transmission Standard (METS) -- specifically, the <structMap> and <smLink> sections -- to determine if it can meet these needs. If this is the case, then we will

develop a METS profile to support this reference tracking.

* *Phase II*

At the beginning of Phase II, we will develop the interfaces necessary to support import into and export from each of our archive implementations. Our assumption, at this point, is that the import/export format will closely mimic the transfer metadata format delivered with the archive medium. Most of this work will involve integration with interface tools to support the handshake between archives. We will work primarily with our assigned import and export partners to determine the exact formats used; but ideally this work will be coordinated amongst all AIHT participants. We expect that, as we work with our import and export partners, we will need to revise various components of the Phase I work to correct incompatibilities.

* *Phase III*

In order to support archive-wide format migration/emulation issues, we need to be able to determine which datastreams have a given format. We cannot rely solely on the semantics of datastream identifiers (often the file extension) to determine format. Nor is the MIME-type sufficiently granular in all cases. Therefore, it will be necessary to develop tools that can read a datastream and associate it with one or more formats. For example, an HTML 4.0 document might be considered all of the following: text, tagged text, html, html 4.0. This processing would occur as datastreams are imported into the archive. Our energy will not be directed at creating new utilities to process individual formats -- that expertise lies in other domains. We will focus on creating a framework into which we can plug utilities created by others.

We will develop a plug-in to support the migration task assigned in Phase III. The system will associate provenance information with each datastream produced by a plug-in (objects imported into the archive will have provenance metadata from the transfer metadata).

We will also attempt to align this work with the Global Digital Format Registry work being coordinated at Harvard and with the efforts of Caroline Arms and Carl Fleischhauer at the Library of Congress.

* *Phase IV*

Our activity during this phase will depend on the outcomes of our research and development efforts and those of the other participants. Therefore, it is not possible to layout a specific technical plan for
Phase IV.

We will work with the Library and other project participants, as appropriate, to complete the activities of this phase.

End of Project

At the end of the project, JHU will provide the Library with a copy of the archive on new media, but as it existed at ingestion time, along with associated transfer metadata. JHU will work with the Library to determine if the contents must be in the same order as in the original archive or if it is sufficient to provide bit-identical datastreams within the archive. JHU will also work with the Library to determine the type of medium on which the copy should be delivered.

JHU will also provide to the Library a snapshot of the local versions of the archive, as they exist at the end of the project, along with their associated transfer metadata.

We anticipate that the archive will be installed onto three machines during the course of this project. In order to guarantee that all remnants of the archive are removed, we will format the disk(s) and reinstall the operating system on each of these machines.

Roles and Responsibilities of Key Personnel

Choudhury, Project Manager, will provide overall administrative oversight of the project, including personnel issues, project reporting, and resolution of any issues that arise during the term of the contract.

DiLauro, Technical Lead, will design the system architecture, implement some prototypes, provide overall technical oversight for the project, and supervise the programmers.

Gourley, Project Coordinator, will track project activity and work with Choudhury to satisfy both project update and financial reporting requirements.

Reynolds, Metadata Specialist, will support the metadata development aspects of the project.

Deliverables

We will subscribe to, and monitor a testwide mailing list, and provide documentation during and after each phase of the test. We will also deliver monthly, quarterly, and end of project reports on the status of the project as well as budget and financial data, using forms or templates according to ISS guidelines. These reports will outline the various actions, findings, and results from each phase, including overall ease or difficulty, best practices, actual effort and cost vs. expected ones, and fruitful areas of further work. The reports will be specific enough to describe actions undertaken by individuals and for other institutions to undertake similar actions.

The quarterly financial report will be based on OMB Standard Form 269, as indicated in the SOW. These reports will be submitted no later than 15 days after the end of each quarter. At the end of the test, we will submit a cumulative financial status report. We will attend meetings

during Phase I and II of the test, and for the post-mortem at the end of the project, as scheduled and planned by the Library of Congress.

Contract Type

Johns Hopkins University understands that the contract type for this award is Time and Materials. This contract type is acceptable with the understanding that the University will not report labor hours, but rather as a percent of effort of base salary as is federally regulated. We agree to work with the sponsor to informally convert such reported percents of effort into hourly rates for comparative purposes.