



## Active Preservation Plug-In

**A problem associated with archiving digital information is that file formats become obsolete as software changes. A practical solution to this is to migrate files to new, supported formats as the old ones become obsolete. The ideal way to do this is to store documents in a digital archive that is designed for such “active preservation”, such as Tessella’s SDB. However, many organizations have existing digital repositories and the Active Preservation Plug-In works with these to characterize files, determine which ones are at risk and migrate them to new formats.**

### Target Systems

The Active Preservation Plug-In (APPI) is aimed at third party document management systems and repositories. These include open source repositories such as Fedora, and commercial EDMS systems such as Documentum and Trim. All have the need to manage data in the very long term and need to take steps to avoid digital decay.

### File Format Obsolescence

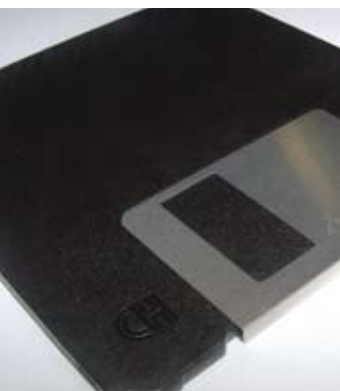
Many of today’s files are tightly linked to today’s software. As the software version moves on, the file reading becomes at first unreliable and then impossible. The problem is compounded by the increasingly complex ways that we interact with the files – the paper paradigm of static content is being replaced by interactive information making simple solutions like copying to PDF incomplete. Also, files must be managed in combination, for example web pages where the image files and HTML together

provide the digital object and must be managed in combination.

Much research is being conducted into this area and the current thinking is that files should be migrated to newer versions enabling them to be read by today’s software. The APPI is built on top of this leading research and uses state of the art techniques built in a flexible framework to combine a library of migration tools with support tools to help select the correct policy and to validate any migration activity.

### Technical Solution

The APPI has a number of components to connect to the repository, read its contents and extract the information it needs, to decide on preservation policy and to conduct preservation activities, reloading new files back into the repository as they are created.



The components are:

- **Crawler:** This finds new content in the repository, characterizes the corresponding files and stores the technical metadata describing the file formats and properties. It will also attempt to find groups of files forming a component that should be migrated as a unit, for example a web page.
- **Preservation database:** This stores the file formats and properties together with information on the structure of records to describe which files form a manifestation of a record in a particular format or set of formats. Additional record hierarchy information may be stored.
- **Technical registry (PRONOM):** This combines a store of file format information with a directory of tools that can migrate between formats. The content is taken from public data generated by leading national archives.
- **Migration workflow system:** This provides a workflow system and user interface to determine which files are at risk and to set off migration jobs. The files resulting from the migration are stored back in the digital repository, with the preservation database also being updated. As this may be a huge job for popular formats this contains throttling and start-stop facilities.

### Repository Interface

In order to extract metadata and files from the repository and to re-ingest them, adaptors are needed for each repository. The interfaces required are:

- **Metadata Harvesting:** Many repositories support OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting); this provides a common interface for crawling repositories for metadata. The APPI will use OAI-PMH where ever possible.
- **File Download:** The repository must supply

an interface to allow the APPI to extract each file from the system for characterization and possibly migration.

- **File Upload:** The repository must provide an interface for the APPI to re-load a new manifestation of the file (still linked to the original metadata) and to be able to mark the previous version as superseded.

This solution makes the active preservation add-on independent of the digital repository – it does not require the repository content model to be adapted to hold preservation metadata. It is able to make full use of the active preservation functionality of SDB while retaining the existing repository.

### Current Status

The Active Preservation system is available as a component of the SDB Digital Archive system and is in use at 7 archives and libraries. The PRONOM Technical Registry is publically available from the UK National Archives and development programmes are in place as part of the Planets research project to enhance this further.

The APPI interface is being developed ready for release later in 2009. We expect to deliver a Fedora interface first but are willing to prioritise other repositories should demand dictate.

### Tessella SDB

Tessella SDB is a complete Digital Archiving system currently in use at many archives and libraries around the world. It includes all the OAIS components: ingest, storage, data management, access and administration as well as an enhanced active preservation system that provides preservation action as well as planning.

**Tessella plc** 26 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire OX14 3YS, UK  
T: +44 (0)1235 555511 | F: +44 (0)1235 553301 | E: info@tessella.com

**Tessella Inc** 233 Needham Street, Suite 300, Newton, MA 02464, USA  
T: 1 617 454 1220 | E: info@tessella.com

**Tessella – successfully delivering IT and consulting services to world leaders in R&D, science and engineering.**

For decades, Tessella has been successfully delivering IT and consulting services to world leaders in R&D, science, and engineering. Through the application of scientific methods and rigorous quality procedures, we enable clients in life sciences, energy, the public sector, and consumer industries to achieve a wide range of objectives, including, forecasting floods, developing fusion power, enhancing military sensor capability, improving drug discovery and development efficiency, and reducing risk to health and the environment in the extraction and production of oil and gas. With offices in Europe and North America, global companies rely on Tessella for business critical assignments.

Copyright © Tessella plc 2009, all trademarks acknowledged. Issue: V1.R0.M0 | Jul-09



[www.tessella.com](http://www.tessella.com)

