PRONOM 4 User Requirements

Author: Adrian Brown
Version: 0f
Date: 29 June 2004

Document Control

Author:	Adrian Brown, Services Manager, Digital Preservation
Document Reference:	PRONOM 4-UR-0f
Issue:	0f
Issue Date:	29 June 2004

Approved By:

"I have read this document and agree that it addresses adequately the business needs and technical requirements of the National Archives".

David Ryan, Head of Archive Services.

Document History

Issue	Author	Date	Comments
0a	Adrian Brown	9 March 2004	First draft
0b	Adrian Brown	16 March 2004	Second draft
0c	Adrian Brown	4 April 2004	Third draft
0d	Adrian Brown	30 April 2004	Fourth draft
0e	Adrian Brown	25 May 2004	Incorporating internal comment
0f	Adrian Brown	29 June 2004	Incorporating PAG comments

Contents

1   Purpose of Document

The purpose of this document is to define the end-user, reporting and technical environment requirements of PRONOM 4. PRONOM is a database designed to store information on the technical dependencies of digital objects, including file formats, software applications, operating systems, and hardware.

A full review and understanding of these requirements will result in a Functional Specification document including a database definition description.

1.1   Background

PRONOM is an application for managing information about the technical dependencies of electronic records. This will be used to support the long-term preservation of records which have been transferred to the National Archives (TNA) for permanent storage, including automation of the digital preservation process, and as a resource for the wider community.

PRONOM was produced by Tessella Support Service Ltd in March 2002. The current version is Version 3, released in January 2004. This incorporates functionality and interface changes required for release on the TNA website.

1.2   Business Processes

PRONOM is a software tool which is required to support the following TNA business processes:

- Accession: PRONOM will be used to support automation of the accession process for electronic records into the digital archive. It will do this by:
    - Supporting the automated identification of file formats.
    - Supporting the validation of file formats.
    - Supporting automated metadata extraction at the file level.
    - Providing a central repository for information about file formats and their technical dependencies, thereby eliminating the need to duplicate this information within the digital archive.

PRONOM will be fully integrated into the accession workflow, identifying formats or reporting exceptions requiring manual identification, extracting metadata, and exporting that information into the digital archive.

- Preservation: PRONOM will support long-term preservation by:
    - Automated generation of technology watch alerts to trigger preservation actions.
    - Automated generation of migration pathways.

- Delivery: PRONOM will support the delivery of electronic records to users, by enabling the automatic configuration of the appropriate software environments.

- In addition, through its availability on the internet, PRONOM provides an important resource to the broader digital preservation and IT communities, by enabling the

dissemination of authoritative information on the technical dependencies of electronic records.

1.3   Known Risks

- Failure to link successfully to other systems and processes.
- Information gathering does not provide sufficient quantity or quality of data.
- Products already exist which can fulfil the primary functions.

1.4   Customer's Quality Expectations

- The technical considerations for development work in phase 4 will be governed by UK government technical standards issued by the Office of E-envoy as E-gif (www.e-envoy.gov.uk/publications/frameworks/egif).
- Development to conform with the TNA Application Development Standard.
- Project Management to PRINCE 2 standards.
- Testing completed to BS 7925-2.

1.5   Acceptance Criteria

- Conformity to above standards
- Pass all critical user tests as defined in User Test Acceptance Criteria Report.
- Produce meaningful individual and statistical reports for management of service.

2 Project Objectives

The overall project objectives are to develop a fully operational product that:

- Contains a repository of information about the technical dependencies of electronic records, including file formats, software products, operating systems, and hardware components.
- Supports automated tools for file format identification and validation, metadata extraction, and migration.
- Supports technology watch through the automatic reporting of obsolete or unsupported products.
- Supports migration through the automatic generation of migration pathways between formats.
- Supports interoperability with other systems for the purposes of data exchange and the provision of automated services.
- Is accessible on the World Wide Web.

The objectives of Phase 4 of the project are to develop a fully operational product that:

- Has been migrated to the new TNA website style.
- Provides an enhanced database structure which includes provision for the recording of detailed technical file format information.
- Includes the development of the first associated PRONOM services, which will support increased automation of the digital preservation process.

3 Requirements

3.1 Non-functional requirements

All existing system requirements still apply unless specifically excluded or superseded below.

3.1.1 Interface

1	User Interface must be web browser-based.
2	User Interface must be intuitive and easy to use without training.
3	Two installations of the system will be required - a master system located on an intranet server, linked to the administrative screens, and a public system located on a web server. Updates will be carried out on the master system, and replicated to the public system by TNA staff.

3.1.2 Security

4	Access to administrative screens must be authenticated through username and password control.

3.1.3 Performance

5	The master and public systems must be able to handle one hundred concurrent users with no detrimental effects on performance.

3.2 Functional requirements

3.2.1 Environment

6	The system must conform to the current version of the TNA Application Development Guidelines.
7	The server-side elements of the system must be compatible with Windows 2003 server and SQL Server 2000.
8	The client-side elements of the system must be compatible with the following browsers: Internet Explorer 5 and later, Netscape Navigator 4 and later, Safari 1.0 and later, Firefox 0.9 and later, and Mozilla 1.0 and later.
9	The system must be year 2000 compliant.
10	The system must comply with e-GIF version 6 and the Guidelines for UK Government Websites.

3.2.2 Application setup

11	The application must be supplied as an installation set, with comprehensive installation instructions, to allow installation by TNA staff.

3.2.3 Data Structure

12	The data must be stored in a recognized database structure.
13	The database must be capable of export to a commonly agreed format in accordance with e-gif e.g. XML or CSV.

4 Required Enhancements

4.1 Interface

14	The look and feel of the interface must conform to the new TNA website style. Style guidance and templates will be provided by TNA.

4.2 Database

15a	The database model will be extended to allow the recording of detailed information on file formats, in accordance with the PRONOM 4 file format information model.
15b	The database will be remodelled in accordance with the revised PRONOM 4 information model.
16	A costed feasibility report to investigate the following: • The potential for moving to a more flexible, object-oriented database design, which removes the hard-coding of specific types of technical object and their interrelationships, is required. This will take into account recent work on the modelling of software systems to support automated configuration management (e.g. Adele, PCL, Software Dock, and the Open Software Description format). • The potential for developing a more generic search mechanism which functions seamlessly across all technical components. • The potential for exposing PRONOM through a Web Services interface. • The potential for removing specific database and other technical dependencies, and moving to a more loosely-coupled system design.

4.3 Search Screens

17	The **File Format** search will be modified to also allow searching by format name. All searches will be assumed to be wildcarded. The search results will be displayed using the **File Formats** search results report (see 4.4 below).
18	The **Compatible Products** search will be modified to also allow searching by format name. All searches will be assumed to be wildcarded. The search results will be displayed using the **File Formats** search results report (see 4.4 below).
19	A new **PUID** search tab will be added to allow searching by PUID. All searches will be assumed to be wildcarded. The search results will be displayed using the **File Formats** search results report (see 4.4 below). The PUID tab will also allow the generation of the current PUID list (PUID, format name, format version) in XML, CSV and printer-friendly format.
20	An **Advanced** Search tab will be added, allowing advanced Boolean searches to be performed on all permissible database fields (i.e. excepting system and maintenance fields).
21	A **Storage Media** tab will be added to allow searching by storage media name and type. The search results will be displayed using the **Storage Media** search results report (see 4.4 below).
22	The **Product** tab will be renamed **Software**.
23	The **Vendor** tab will be modified to allow searching for developers and maintainers of file formats, software products, hardware components, and storage media, through the provision of a "Technical Component" list box.
24	The **Support Period** and **Release Date** tabs will be combined into a single **Lifecycles** tab.

25	The **Supported** and **Unsupported Products** searches will be combined through the provision of a "Supported/Unsupported" list box.
26	The **Release Date** and **Released within Range** searches will be combined.
27	The **Lifecycles** tab will allow searching on lifecycle information about file formats, software products, hardware components, and storage media, through the provision of a "Technical Component" list box.
28	All search screens will be modified in order to conform to new TNA standards for search screens. TNA will provide relevant standards.

4.4   Reports

29	The **file formats** search results report will be modified to display PUID, format name, and version, each column being hyperlinked to the relevant detailed file format report.
30	A new **file format** report will be created to display the detailed record for a file format.
31	The **file formats written and read** tabs of the **product** report will be modified so that each format name listed is hyperlinked to the relevant detailed file format report.
32	File format search results and detailed reports can be saved in XML, CSV and printer-friendly format.
33	A new **Storage Media** search results report will be created to display media name and version, and vendor name, each column being hyperlinked to the relevant detailed report.
34	A new **Storage Media** detailed report will be created, to display the detailed record for each storage medium.
35	Storage media search results and detailed reports can be saved in XML, CSV and printer-friendly format.

4.5   Help System

36	The help system will be modified to comprise a single HTML page. Context sensitive help will be provided by opening the page at a given list target location.
37	The help system will be updated to include all new search and report screens.

4.6   Technology Watch Tool

38	A new technology watch tool will be developed to report on file format obsolescence. This will provide a graduated series of alerts when elements of the technical environment required to access a file format (e.g. software products) cease to be supported.
39	TNA will provide a technology watch status ranking system for describing the current status of a file format. TNA will also provide a rule set for determining the current technology watch status of a file format.
40	The tool will automatically generate and send an email message to a defined list of email addresses whenever the technology watch status of a file format changes.
41	The tool will generate on demand a report on the current technology watch status of all file formats. This report will be available in XML, CSV and printer-friendly format.
42	The tool will generate on demand a report of all file formats with a specified technology watch status. This report will be available in XML, CSV and printer-friendly format.
43	The master system will allow the automatic generation of a report in XML format,

	whenever the technology watch status of a file format changes.
44	The master system will allow the automatic generation of a report on the current technology watch status of all file formats in XML format, whenever the technology watch status of a file format changes.

4.7   Identification Tool

45	A standalone Java (JRE 1.4 or later) application will be developed to perform automatic file format identification, using signatures recorded in PRONOM.
46	The tool will allow a user to browse any file system accessible from the computer on which the application is installed, and select single or multiple files for identification, including the contents of entire folders, to create a list of files to be identified. The user must also be able to subsequently remove files from the list.
47	The tool will display the resultant file list in a text box. The list will display the full filename, with the option to toggle display of the fully-qualified pathname on or off.
48	The tool will perform automatic file format identification of all files on the selected list in a single operation, using an XML signature file generated from PRONOM. TNA will work with the developer to provide an appropriate algorithm.
49	The result of the identification will be displayed against each file in the list (e.g. "Identified", "Tentative" or "Not Identified")
50	Full details of each identification (PUID, Format, Version) will be displayed in a separate text box, dynamically linked to the currently highlighted file in the file list text box. The tool will support multiple tentative identifications for a file, where the precise format version cannot be ascertained.
51	The tool will allow the results of the identification to be saved as an XML, CSV or printer-friendly report.
52	The tool will automatically check the PRONOM website at user-defined intervals for updated signature files and download them.
53	The update mechanism will be user-configurable in terms of timing, proxy server settings and authentication.
54	TNA will generate updated XML signature files as required, using the maintenance system, and make them available on the PRONOM website for download.
55	The tool will be deployed as a single JAR package, to be downloaded from the PRONOM website.
56	A separate functional specification will be required for this tool.

4.8   Maintenance

57	TNA wishes to move from the current, web-based administration screens to using a client-side administration tool based on Access 2000/XP, and the OLE DB component architecture. A proposal for allowing the import of images and reference files using such a system is required.

4.9   Submission

58	A simple workflow for automating the submission of new information will be developed. This will allow for the automatic import of submission emails into PRONOM, flagged as unverified. Unverified records will not be displayable on the public system. Flagged records will be viewable using the master system administrative tool, for verification.

4.10 Associated Documentation

59	The system will support linking from hypertext links in detailed reports to a library of supporting documents (such as file format specifications). These documents will be stored in RTF or PDF format.