National Library of New Zealand

Metadata Extraction Tool

Open Source Release Information

Introduction

The Metadata Extraction Tool was developed by the National Library of New Zealand (Te Puna Mātauranga o Aotearoa) to programmatically extract preservation metadata from a range of file formats like PDF documents, image files, sound files Microsoft office documents, and many others.

It is now available as open-source software from http://meta-extractor.sf.net/.

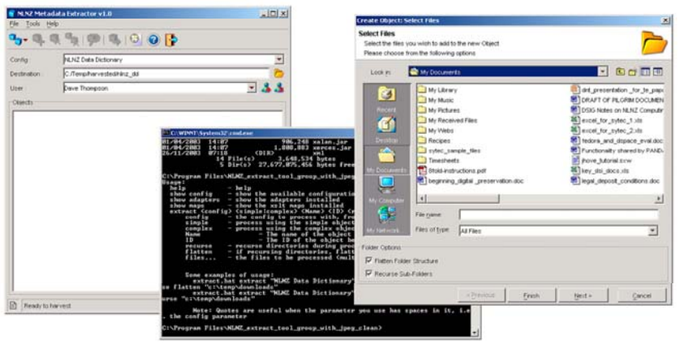
The Tool builds on the Library's work on digital preservation, and its logical preservation metadata schema. It is designed to:

- Automatically extract preservation-related metadata from digital files.
- Output that metadata in XML formats for use in preservation activities.

The Tool was designed for preservation processes and activities, but can be used to for other tasks, such as the extraction of metadata for resource discovery.

Capabilities

The Tool has both a Microsoft Windows interface and a UNIX command line interface. This enables work to be automated through batch processing or processed on an individual basis as required.



The Windows graphical user interface and UNIX command line interface

Supported file formats

The Metadata Extraction Tool includes a number of 'adapters' that extract metadata from specific types of file. Extractors are currently provided for:

- Images: BMP, GIF, JPEG and TIFF.
- Office documents: MS Word (version 2, 6), Word Perfect, Open Office (version 1), MS Works, MS Excel, MS PowerPoint, and PDF.
- Audio and Video: WAV and MP3.
- Markup languages: HTML and XML.

If a file type is unknown the Tool applies a generic adapter, which extracts data that the host system knows about any given file (such as size, filename and date created).

How the Tool works

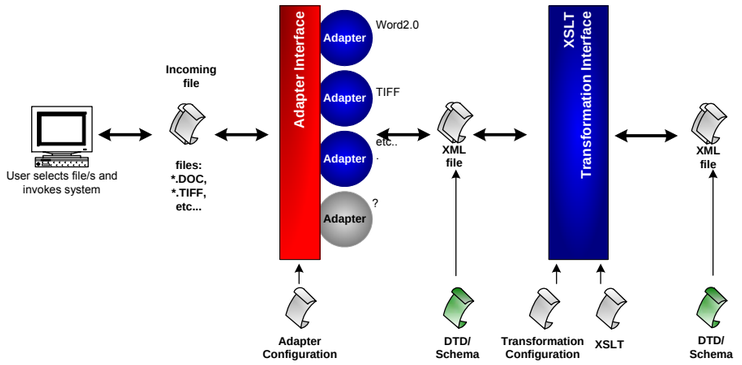
The Metadata Extraction Tool is based on a library of adapters. Each adapter knows how to recognise and extract metadata from a different type of file.

Adapters can handle dependencies within and between objects of varying levels of complexity, ranging from single, simple objects like TIFF files through to complex web sites or databases.

Architecture

Extracting preservation metadata is a two-stage process. In the first phase each incoming file is processed by the adapters until one of the adapters recognises the file type. That adapter extracts data from the header fields of the file and generates an Extensible Markup Language (XML) file.

In the second phase an Extensible Stylesheet Language (XSL) transformation converts the internal XML file into an XML file in a useful format. The Tool currently outputs the XML file using the NLNZ preservation metadata data model schema.



Solution Architecture

Preservation constraints

The Tool operates under these constraints:

- It must not change the files it processes.
- It must process many thousands of files.
- It must be consistent.
- It must process simple objects (single files) and complex objects (many dependant files).

The Tool opens all files as read-only, ensuring the integrity of original files. It usually only reads header information so the extraction process is fast. Consistency is vital because decisions about preservation will be based upon the extracted metadata.

Technology

The Metadata extraction Tool uses a combination of Java and XML. It can be used through the graphical user interface, at the command line, or incorporated into other programs as required.

The Tool is distributed as Free Software under the Apache Public License (version 2).

Digital preservation at the National Library of New Zealand

The National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act 2003 requires and mandates the Library to take responsibility for the collection and management of digital material 'in perpetuity'.

Preservation in perpetuity

Preserving digital material in perpetuity is much more difficult than simply collecting it. Digital material is by its nature ephemeral, fragile and poorly suited to preservation.

The Library has addressed this challenge by developing Tools and processes to support digital preservation, including:

- Preservation Metadata Schema – High-level overview of the minimum set of metadata necessary to preserve digital material
- Preservation Metadata Data Model – Conversion of the schema into a practical model for a database to hold preservation metadata
- Metadata Extraction Tool – Based on the data model, automates extraction of preservation metadata from digital material.

The Metadata Extraction Tool was designed with the needs of the wider digital preservation community in mind, and future development will be informed by the community.

Why automate?

Digital material is well suited to automated processing. Automating the extraction of metadata provides accuracy and reliability, and allows the use of extracted metadata during other automated processes such as ingesting material into a repository.

Automation is integral to implementing a digital preservation metadata strategy:

- It mitigates the risk of human error.
- It allows the processing of far more material than manual processes.
- It produces output that can be adapted to suit other business processes.
- It provides standardised, transparent and auditable processes.

The current growth of digital material makes it impossible for manual processing to generate preservation metadata. The Metadata Extract Tool provides a simple way of automating preservation processes.

Further information

The Metadata Extraction Tool software and documentation (User's Guide, Installation Guide, Developer's Guide, Change Log and Software Architecture) is available from http://meta-extractor.sf.net/.

More information on the National Library of New Zealand can be found on its website http://www.natlib.govt.nz/.

For more information, contact metadata-extract@natlib.govt.nz.

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND