NLA Software and File Formats Knowledge Base

Dr Mark Pearson
National Library of Australia
Parkes Place West
Canberra, ACT 2600
Australia
+61 2 6262 1080
mark.pearson@nla.gov.au

Gareth Kay
National Library of Australia
Parkes Place West
Canberra, ACT 2600
Australia
+61 2 6262 1031
gareth.kay@nla.gov.au

ABSTRACT

This demonstration will showcase ongoing work at the National Library of Australia to develop a software and file formats knowledge base for digital preservation purposes. This project involves empirical research into the capabilities of software applications in relation to file formats. We will talk about the types of information we capture in the knowledge base and describe the steps we are taking to transform it into a machine-actionable graph database, a prototype of which will also be demonstrated.

General Terms

infrastructure

Keywords

Software, file format, knowledge base, graph database.

1. INTRODUCTION

The National Library of Australia has an ongoing project to develop a knowledge base detailing relationships between software applications and file formats. This paper describes the drivers and strategic goals for developing the knowledge base and the rationale behind taking an empirical approach to its development.

2. SCOPE OF THE WORK

The project involves detailed empirical research into the capabilities of selected software applications with respect to selected file formats. The research is primarily format-driven since the primary long-term goal is to be able to successfully access content stored in digital files. Priority file formats have been chosen based on business needs and the composition of the NLA's digital collections.

For each major abstract content type (images, textual documents, videos, spreadsheets, maps etc.) we have chosen a small number of the most predominantly used applications in order to investigate their capabilities with respect to the associated file formats. The applications chosen may be proprietary in nature or open source.

Details such as release dates, versions, vendor support, licensing status and dependencies are recorded both for formats and applications. Due to business needs the data gathered from the research is initially being recorded in a multiple worksheet Excel file in semi-structured format. Development of a prototype graph database together with software modules capable of importing data from the Excel file is taking place in parallel with the empirical work.

While Excel is not a suitable platform for a production knowledge base, its use in the development phase does have some advantages: as our understanding of the problem domain improves through empirical contact with it, we can experiment with changes to our data model at very little cost. When we come across aspects of the software/file-format relationship which we judge might be significant to future preservation decision making but which the current iteration of the model provides no structured way to record, we can adapt the model accordingly.

Two very useful by-products of the empirical work are: a growing corpus of files in various formats and format versions containing known content which we have created ourselves and which we can usefully employ in testing software package capabilities; and a growing collection of VMWare virtual machine images for various current and historical operating system environments.

3. GOALS AND DRIVERS

The long-term strategic goal is to build machine-readable knowledge bases to aid us in: determining our *level of support* [1] for different file formats; analysing the NLA's digital collection materials for preservation risks; and planning and executing preservation actions on digital objects which comply with the documented *preservation intents* [2] for those objects.

A more immediate goal is to replace an existing Drupal-based software/formats knowledge base which is limited in its ability to express arbitrary relationships between entities and is not suitable for machine querying or complex queries.

There is much existing work in the area of technical registries [3][4][5][6][7] and the NLA is actively involved in other work in this area through collaboration with organisations such as National and State Libraries Australasia [8]. While the outcomes of this project will provide practical benefits for the NLA they will also hopefully provide food for thought for the wider community in its efforts to develop open, maintainable linked data technical registries.

4. THE KNOWLEDGE BASE

Functional relationships – A key function of the knowledge base is to map out the capabilities of software applications in relation to the file formats they are (or claim to be) able to handle. To gather this data we investigate certain *functional relationships* for each software/format combination. These relationships are used to describe capabilities exhibited by an application in relation to a format. Currently, we investigate four relationships: *import*; *render*; *edit*; and *save*. These relate to whether an application can parse a given format and build a 'meaningful' internal representation of its content; render that internal representation; allow a user to make changes to the content; and save it to the format, respectively.

The process of documenting these functional relationships involves as a first step harvesting information (where available) from vendor documentation and/or running the software and noting the formats listed in the 'Open' and 'Save as' menus. Such entries in the knowledge base are assigned a confidence value of 'untested' as we don't know how *well* the software opens or saves a given format. For file formats which are considered high priority by the NLA the functional relationships are empirically tested with the aid of the *test file corpus* (described below). Such entries are assigned a confidence value of 'tested' and are more detailed in nature.

Preservation notes – During the process of investigating the functional relationships issues which could affect the suitability of either an application or a file format for future preservation actions sometimes arise. Examples could include rendering issues; discrepancies between documented and actual software functionality; software/hardware dependencies; installation issues; and/or the inability to preserve certain properties of content which may have been deemed significant by the preservation intent statements associated with the content type. These issues are recorded in semi-structured format in the 'preservation notes' field. Crucially these notes can act as triggers for reassessing the knowledge base schema if it becomes clear that the current schema provides no structured means for recording such details.

Test file corpus – A useful by-product of this project is a growing benchmark corpus of test files in selected file formats, created in software packages which have been documented in the knowledge base. When new test files are created, their content is carefully crafted in accordance with current preservation intent statements for the content type. Put more simply, the content is chosen so that we can test how well an application maintains important features of that content when importing, rendering, editing or saving it.

What makes this corpus particularly valuable is that each test file is linked within the knowledge base to the software version used to create the test file, the operating system and environment in which the software was run, as well as the format and version the file is written in. Another feature is the inclusion of screenshots showing the content from each file rendered in the software it was created with. This additional resource allows for the detection of content loss or rendering issues when a file is opened in a different application.

5. GRAPH DATABASE

When the empirical part of this work began we did not have a suitable database in which to record our findings and for this reason, as mentioned above, the data is currently recorded in an Excel file in semi-structured format. However, in parallel with the empirical work we have also been developing a set of software modules for importing and transforming the Excel data into a *directed property graph*: a "key/value-based, directed, multi-relational graph" [9]. In such graphs both vertices and edges may have arbitrary sets of key/value attributes.

A number of database systems supporting the property graph model are currently available [10] but the system we have chosen initially – OrientDB [11] – supports vertex and edge types which can have inheritance relationships. Vertices, edges and vertex/edge types can all be dynamically added and removed. This makes it ideally suited to problem domains where the schema has not been strictly defined or may be a 'moving target'. It is also open source, released under the Apache License, Version 2.0.

OrientDB supports the Tinkerpop Blueprints property graph Java API [10] - described as "JDBC, but for graph databases" and both an SQL-based query language extended with features for graph traversal; and Gremlin [11] – a graph traversal language.

6. THE DEMONSTRATION

The demonstration at iPres 2014 will address in more detail the data we are capturing in the knowledge base and the nature of the functional relationships we test for each software/format combination. The prototype graph database will also be demonstrated with example queries.

7. REFERENCES

[1] Pearson, D. 2012. The Adventures of Digi: Ideas, Requirements and Reality. Presentation at *Future Perfect 2012*, Museum of New Zealand Te Papa Tongarewa, Wellington. https://www.nla.gov.au/content/the-adventures-of-digi-ideas-requirements-and-reality

[2] Webb, C., Pearson, D. and Koerbin, P. 2013. *"Oh, you wanted us to preserve that?!"* Statements of Preservation Intent for the National Library of Australia's Digital Collections'. D-Lib Magazine, Vol.19 1/2, Jan/Feb 2013.

[3] *PRONOM technical registry* - http://apps.nationalarchives.gov.uk/pronom/

[4] *Unified Digital Format Registry (UDFR)* - http://www.udfr.org/

[5] Anderson, David and Delve, Janet (2012) *The Trusted Online Technical Environment Metadata Database* – TOTEM. In: Kölner Beiträge zu einer geisteswissenschaftlichen Fachinformatik . Verlag Dr. Kovac, Hamburg. ISBN 9783830064183

[6] http://www.openplanetsfoundation.org/blogs/2010-12-08-breaking-down-format-registry

[7] McGath, G. 2013 *The Format Registry Problem.* Code4Lib Journal. Issue 19, 2013-01-15. http://journal.code4lib.org/articles/8029

[8] http://www.nsla.org.au/projects/digital-preservation

[9] *Defining a Property Graph* - https://github.com/tinkerpop/gremlin/wiki/Defining-a-Property-Graph

[10] *Tinkerpop Blueprints* home page - https://github.com/tinkerpop/blueprints/wiki

[11] *Orient Technologies* home page - http://www.orientechnologies.com

[12] *A Graph Traversal Language* - https://github.com/tinkerpop/gremlin/wiki