Enhancing characterisation for digital preservation

Paul Wheatley
University of Leeds
Brotherton Library
Woodhouse Lane
+441133435562
p.r.wheatley@leeds.ac.uk

Gary McGath
Independent Software Developer
42 Heather Court
developer@mcgath.com

Petar Petrov
Creative Pragmatics
Kolonitzgasse 9/11-14
Vienna
petar@creativepragmatics.com

ABSTRACT

Advances in digital preservation software tools have sometimes been slow and or poorly directed. The result has been a lack of tools that meet practitioner needs, and a surplus of tools that have very few users and little practical application. The Jisc funded SPRUCE Project has championed the recording and sharing of practitioner requirements, and the development of solutions to meet those requirements using agile hackathon or mashup style events. This poster will provide a visual summary of requirements identified by practitioners, and will describe four resulting tool developments that significantly advance our digital preservation capability.

Keywords

Digital Preservation, Hackathon, Mashup, User Requirements, Digital Preservation Tools

1. BACKGROUND AND REQUIREMENTS

Practitioners responsible for managing digital data rely on automated software tools [1] to perform many of the key functions that are typical in archival and preservation workflows. Home-grown digital preservation tools have not always developed at a pace or with coverage sufficient to meet practitioners' needs. Steve Knight observed in last year's iPRES opening keynote: "We are still pretty much talking about the same things. Tools like DROID and PRONOM etc. didn't work properly then, and they still don't work properly now" [2]. The problem has not just been that tool development has been lacking, but that the focus and direction of development energies have been poor. Opportunities to build on existing toolsets and incorporate digital preservation requirements have been missed. Even where potentially useful tools have been developed, they have often struggled to find a user base. Examples of duplication and lack of coordination are common[1].

Over the last couple of years the SPRUCE Project [3] has been championing collaborative events that place a strong emphasis on meeting practitioner needs by re-using and enhancing existing software tools. By facilitating the cooperation of both practitioners and software developers, the outcome of tool development has had increased impact and value.

This poster will provide the background to the practitioner requirements and subsequent development (described in sections 2 and 3 below) by outlining the requirements capture process and then highlighting statistics on the number of events at which requirements were gathered (14), the number of practitioners who contributed requirements (100), and the number of organizations which the practitioners represented (70).

2. WHAT ARE THE PRIORITIES FOR DIGITAL PRESERVATION PRACTITIONERS?

Practitioners were asked to bring their digital preservation challenges to the Open Planets Foundation (OPF) hackathons, AQuA Project mashups and SPRUCE Project mashups that were held over the last couple of years. Further challenges were contributed by the EU funded SCAPE Project. Some constraints were placed on the scope and focus of these challenges, mainly related to the scale of challenges that could realistically be addressed in a 2 or 3 day hackathon. Practitioners were otherwise left to contribute whatever digital preservation challenges they wanted to have addressed.

All of these challenges (and related descriptions of the data on which they are focused, and the solutions developed to solve them) were captured in different locations on the OPF wiki and were then collated on a single wiki page using Confluence tagging functionality [4]. The result is a detailed record of practitioner requirements and current preservation practice.

Five key themes were drawn from the 140+ preservation issues identified by practitioners:

- Quality assurance and repair of damaged or potentially damaged data or metadata
- Appraisal and assessment in order to inform selection, curation and next steps
- Locating preservation worthy data, typically where mixed with other data across shared server space
- Identifying preservation risks in order to inform preservation planning
- A long tail of miscellaneous issues including contextual issues, data capture, embedded objects, and broader issues around value and cost

The overriding focus of these themes is the need to characterize digital data and therefore better understand what it is and what condition it is in. This understanding is typically required before subsequent steps in preservation and curation are undertaken.

This poster will summarize these prioritized practitioner needs, and highlight their relevance for steering future tool development activity.

3. CHARACTERISATION TOOL DEVELOPMENT BASED ON PRACTITIONER NEEDS

Many of the practitioner challenges were tackled as part of the events in which they were raised, with a range of outcomes. Some

[1] For example see "Digital Preservation Cost Modelling: Where did it all go wrong?", which references ~17 different costing models/tools developed to meet very similar aims: http://www.openplanetsfoundation.org/blogs/2012-06-29-digital-preservation-cost-modelling-where-did-it-all-go-wrong

resulted in completed tools that were subsequently put into production use at the practitioners' organizations. Some provided proof of concepts or prototypes pointing the direction for future development. Some resulted in unsuccessful approaches, and some remained unsolved.

Analysis of the practitioner needs provided a review point at which to consider next steps for further exploitation of the best work taken on during the hackathon and mashup events, and to consider how the high priority needs could be addressed more effectively. Given the clear need for better characterization, it was decided to host a developer only event which would enable a more concerted effort to update and enhance key digital preservation characterization tools. Further development work could be supported through SPRUCE Awards of up to £5000, which were made available under a funding call for event participants.

A dedicated characterization hackathon was hosted by SPRUCE and the University of Leeds in March 2013 [5]. It was attended by a group of experts including representatives from many of the high profile, home grown digital preservation characterization tools including: JHOVE, JHOVE2, DROID, FIDO, C3PO and FITS. The theme of the event was to coordinate and combine efforts and technology to improve characterization capability.

Four key areas were tackled at the event which are briefly summarized below.

3.1 Solving the PDF Preservation Problem

PDF issues were a recurring theme in previous mashup and hackathon event theme that resulted in a variety of experiments. The majority of these utilized Apache Preflight (or related PDFBox libraries) suggesting this technology had considerable potential. The practitioner challenges also highlighted the inadequacy of existing community solutions. JHOVE for example provides very detailed output for PDFs, but without a clear focus on preservation risks (the main practitioner need) and with data on some risks lacking. Therefore the largest of the four groups at the characterization hackathon wrapped Apache Preflight as a PDF risk analysis tool. Evaluation with large amounts of real data and possible incorporation into key repository technologies to achieve maximum impact for UK Higher and Further Education practitioners (eg. EPrints and DSpace) is being explored at the time of writing.

3.2 Consolidating File Format Identification

The "big 3" file format identification tools, DROID, Tika and File, all have their own file format signatures or "magic" [6], stored in different formats. This data is used to distinguish between different file formats. This leads to the different format identification tools reporting different results for the same file. Each tool has strengths and weaknesses present in its file format magic. Combining the magic would enable a significant improvement in identification coverage and a reduction in inadequate and confusing results for the tool users. Both would be big wins for practitioners. The group made considerable progress in mapping Tika magic to DROID magic. Although not a complete solution, it provided a lot of valuable data for the DROID team to collate and enhance the DROID magic, taking us much closer to a single source for file format magic.

3.3 Wrapping Tika for use in FITS and C3PO

The final two groups looked at addressing the complex picture [7] surrounding the key preservation tools: Apache Tika, FITS and C3PO. All of these tools have considerable potential to deliver effective digital collection assessment via automated characterization, but their current status presents a variety of challenges for end users. FITS for example wraps a number of out of date tools, while C3PO does not offer many extension points.

Two groups of developers at the characterization hackathon focused on incorporating the Apache Tika characterization tool into FITS and C3PO with the aim of making use of the better performance Tika provides and reducing metadata sparsity. Follow up SPRUCE funding awards were granted to address a variety of issues with FITS and C3PO, with the aim of refreshing this toolset. As well as enhancing the functionality and capability of the tools work behind the scenes on the source code and on new documentation has simplified the process for other developers to add support for new tools. This should make future development and support from the community (rather than just the original authors) a more realistic prospect. The OPF will continue to provide coordination, code management, testing and quality assurance to support this process. Further hackathons (such as iPRESHack [8]) will provide stimulus for new community sourced developments.

The poster will summarize the tool developments in these four areas, demonstrating how a strongly practitioner led approach can result in well focused tool development and a high impact for the end user.

4. REFERENCES

[1] Digital Preservation Tools http://wiki.opf-labs.org/display/SPR/Digital+Preservation+Tools

[2] Steve Knight quote in: Angevarre, Inge, NCDD Blog, http://www.ncdd.nl/blog/?p=3338

[3] SPRUCE Project, http://wiki.opf-labs.org/display/SPR/Home

[4] Digital Preservation and Data Curation Requirements and Solutions, OPF wiki page, http://wiki.opf-labs.org/display/REQ/Digital+Preservation+and+Data+Curation+Requirements+and+Solutions

[5] SPRUCE Hackathon Leeds: Unified Characterisation, http://wiki.opf-labs.org/display/SPR/SPRUCE+Hackathon+Leeds%2C+Unified+Characterisation

[6] File format magic, Wikipedia http://en.wikipedia.org/wiki/File_format#Magic_number

[7] To fits or not to fits, Petar Petrov, http://www.openplanetsfoundation.org/blogs/2012-07-27-fits-or-not-fits

[8] iPREShack, hackathon at iPRES2013 http://wiki.opf-labs.org/display/SPR/iPREShack+-+SPRUCE%2C+OPF+and+CURATEcamp+hackathon+at+iPRES2013