Capturing practitioner needs

Mashups and hackathons

Community developments of preservation and curation services have failed to target the real needs of users and practitioners working at the coal face of digital stewardship. Many community tools lie abandoned and unused. We ran mashup events that invited practitioners to bring along their data and their preservation challenges, then worked to solve them by re-purposing and developing existing open source tools. OPF hackathons and the EU funded SCAPE Project also followed a practitioner driven approach.

Practitioners

The practitioners came from a variety of backgrounds and organisations that included Libraries, Archives, Museums and commercial entities. Some were experienced information professionals, others were taking their first steps as digital stewards. All had some responsibility for the management of digital data.


| 90 | Datasets |
| 150 | Issues |
| 90 | Solutions |

Activities from mashups, hackathons and the SCAPE Project are captured on the Open Planets Foundation wiki. The result is possibly the largest and most detailed record of practitioner requirements and current preservation practice on the internet.

Analysis of this data collates the individual practitioner issues into 5 key themes which are described on the right. The overriding focus of these themes is the need to characterise digital data and therefore better understand what it is and what condition it is in.

Quality assurance and repair of damaged or potentially damaged data or metadata

Appraisal and assessment in order to inform selection, curation and next steps

Locate data to preserve typically where mixed with other data across shared server space

Identify preservation risks in order to inform preservation planning

Long tail a cross section of miscellaneous preservation issues

Enhancing Characterisation for digital preservation

Many of the practitioner challenges were tackled as part of the events in which they were raised, but a dedicated characterisation hackathon and two small awards of £5000 allowed significant progress to be made in enhancing characterisation capability in 3 key areas. The results demonstrate the impact of a community focused approach which utilises shared problem solving and collaborative software development.

Solving PDF preservation

The Apache Preflight tool (which validates a PDF against the PDF/A standard) was wrapped as a PDF risk identifier, providing simple and configurable reports on where PDF’s could pose long term preservation risks. It was named PDF/En? Large scale testing and possible incorporation in key repository software was being explored at the time of writing.

Consolidating file format magic

The “big 3” file format identification tools, DROID, Tika and File, all have their own file format signatures or “magic”, stored in different formats. An unsatisfactory situation for users, and a state of perpetual effort duplication for the tool maintainers. A mapping was created from Tika to DROID magic. Although not a complete solution, it provided a large volume 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. Future work will aim to meet this goal.

Automating characterisation

FITS provides comprehensive data characterisation but has in recent years become out of date, as Harvard has struggled to support the codebase on its own. C3PO provides visualisation of FITS output. In tandem, they have the potential to provide a high impact characterisation solution. This toolset was brought up to date and opened up for future community support by wrapping new versions of the tools, adding Apache Tika, providing new documentation and simplifying the codebases.