Supporting practical preservation work and making it sustainable with SPRUCE

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

Maureen Pennock
British Library
Boston Spa
Wetherby
+441937546302
m.pennock@bl.uk

ABSTRACT
The SPRUCE Project has applied community oriented approaches to support and sustain digital preservation activity. An emphasis on practitioner requirements and focused agile development has enabled the updating and refinement of key digital preservation toolsets that meet user needs. The capture and sharing of these requirements has provided a detailed snapshot of current curation practice, providing insight into practical practitioner needs for those able to fund and support tool and service development. A variety of collaborative initiatives have developed online resources and forums for supporting digital preservation activity. SPRUCE has begun constructing a toolset to support managers and practitioners in making the case to fund and sustain digital preservation activity.

As SPRUCE enters its final half year, this paper provides an outline of key achievements as well as thoughts on the effectiveness (or otherwise) of some of the more innovative or unconventional approaches taken by the Project.

Keywords
Digital preservation, requirements, agile development, Hackathon, Mashup, business case, collaboration.

1. INTRODUCTION
The SPRUCE Project [1] is a two year collaboration between the University of Leeds, the British Library, the Digital Preservation Coalition, the London School of Economics and the Open Planets Foundation. SPRUCE is funded by Jisc with the aim of supporting digital preservation activity and making that activity sustainable. The project is primarily focused on supporting grass roots preservation activity, by connecting individuals responsible for managing digital data with domain experts, technical experts and a supportive community of peers. Both face to face events and collaboration and support via online communication tools, social networking and web based resources are being employed by the project. Development of a set of resources to support teams and organizations to articulate their case for resourcing digital preservation will help to make this supported preservation activity more sustainable.

This paper focuses in some detail on software tools of relevance to the long term preservation of digital content. There is insufficient space to describe each of these tools in detail, and so it is recommended that readers without experience of these tools utilize a reference resource such as the OPF Tool Registry [2] to provide context to the observations in this paper.

2. SOLVING PRACTITIONER CHALLENGES WITH AGILE DEVELOPMENT AND COLLABORATIVE EVENTS
A core part of meeting SPRUCE's aims has been delivered through the use of agile events, including Mashups and Hackathons. At the time of writing, SPRUCE has delivered 2 Mashups with a third planned for July 2013. The events bring together practitioners (who contribute digital data and preservation challenges) and developers (who apply tools to solve the practitioners' challenges). They support expert attendees in expanding understanding and tackling complex challenges, and help staff from organizations taking their first steps in digital preservation activity. The format of these events was covered in detail in a paper from the AQuA Project at iPRES 2011 [3], so this paper will concentrate on the outputs of these events to date. As well as resulting in many useful outcomes for each individual practitioner or developer, a vital output from the events has been the capture and sharing of practitioner requirements with the wider community.

Practitioner requirements were captured from each SPRUCE Mashup event as well as from AQuA Project Mashups (where the format was first developed), Open Planets Foundation (OPF) Hackathons, and practitioner needs generated by the EU funded SCAPE Project. This totaled over 140 different preservation challenges or "issues", sourced from over 100 practitioners, who represented over 70 different organizations.

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 two or three day event. 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. SPRUCE collated this data on a single wiki page using Confluence tagging functionality. The result is a detailed record of practitioner requirements and current preservation practice [4] that provides an essential companion to this paper. The solutions to the practitioner derived issues are one of the most obvious and valuable outputs from the SPRUCE Project. Solutions range from fully functioning technical solutions that have since been adopted and embedded in practitioner's organizations, promising prototypes or demonstrators, and also experiments that presented a dead end. For example, a particular tool or approach was explored, but it was decided (often following testing with actual data from the practitioner) that it did not lead to an effective outcome. Capturing the evidence of where a particular tool did not

work well to solve challenges with particular data was seen to be as useful as capturing success stories. Both cases can useful inform (and provide evidence based lessons learned) for other practitioners.

2.1 Understanding and Addressing Practitioner Needs

As the data captured on practitioner needs grew, it was felt that further benefit could be gained from a more detailed understanding of what digital preservation practitioners most needed help with. This became a key focus to explore and report on for the project. What are the priorities for supporting digital preservation practitioners, and what could be done to meet these priorities?

Analysis was performed by SPRUCE on the preservation issues data (i.e. the , with a view to informing the direction of digital preservation tool development. 5 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.

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 that SPRUCE should host a developer only event which would enable a more concerted effort to update and enhance key digital preservation characterisation tools. Further development work was 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 and are described below.

2.2 Solving the PDF Preservation Problem

PDF issues were a recurring theme in previous Mashup and Hackathon events. The majority of solutions explored the use of 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. JHOVE is able to validate a PDF file against the PDF standard. Practitioners wanted to assess a PDF file against an agreed list of genuine preservation risks. Although these two use cases are similar (and indeed overlap) they are not identical; a common misconception which has led to cases of practitioners migrating perfectly renderable PDFs that JHOVE had assessed as invalid (eg. Friese [6]). Therefore the largest of the four groups at the characterization Hackathon wrapped Apache Preflight as a PDF risk analysis tool. An evaluation with large volumes 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 as part of the final SPRUCE Mashup, and the OR2013 developer challenge (both in July 2013).

2.3 Consolidating File Format Identification

The “big 3” file format identification tools, DROID, Tika and File, all have their own file format magic which is used to distinguish between each different file format. This leads to the different format identification tools sometimes 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 unhelpful and confusing results for the tool users. Addressing this problem would be a big win for practitioners. The group made considerable progress in mapping Tika magic to DROID magic. Although not a complete solution (due to the complexity of the challenge), 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.

2.4 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.

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. These were ongoing at the time of writing, but considerable progress has already been made (including bringing the wrapped tools within FITS up to date).

The end result should provide a comprehensive assessment and characterization capability with across the board applicability for a large number of practitioners.

2.5 Evaluation

SPRUCE feels it has demonstrated the value of developing software based on comprehensive requirements from practitioners. The real effectiveness of the resulting tool enhancements will become clearer over the final term of the Project, but SPRUCE has clearly demonstrated that significant progress can be made with limited resources if a collaborative and well targeted approach is taken.

The growing popularity and success of activities with some similarity in approach, for example the North American CURATEcamp events [8], reinforces this position. The recent

audio visual focused CURATEcamp day [9] attracted over 150 viewers and a smaller but considerable number engaged via IRC and Google Hangouts.

Home grown preservation tools (meaning those created by this community) are often created with an initial burst of development work, sometimes funded by a specific organization, sometimes with external funding. Whichever the funding source, sustaining the effort, and consequently the tool, can be a challenge. Maintenance and enhancement over time, can however be possible with community contributions and occasional small injections of funding, as SPRUCE has demonstrated.

More effective support in managing tool development, perhaps provided by a coordinating organization, has the potential to make it far more realistic for effective tool maintenance to be performed with these small contributions of effort from across the community (and in particular from occasional Hackathon events). Automated builds, regression testing (essential when making changes and improvements to a complex tool such as FITS) and provision of a consistent test corpora could all play a useful part. SPRUCE partner, the Open Planets Foundation, is seeking to take on this role and has plans to establish supporting activities over the coming months. For example see [10].

3. ONLINE AND REMOTE COLLABORATION

The SPRUCE Project has explored taking some of the positive community experiences from its face to face events and applying them in alternative channels. SPRUCE contributed in a variety of ways to a number of online initiatives. Some were created and launched by SPRUCE, some came about in partnerships with other like minded individuals and organizations, and some were simply promoted by SPRUCE. A single page on the SPRUCE wiki brought together links and publicity to all of the initiatives described below [11].

3.1 Initiatives

A recurring theme at Mashup events, Hackathons and during lively digital preservation discussions on twitter [12] was the need for sharing example files to enable preservation challenges to be collaboratively explored and also to support the development and testing of digital preservation tools. Whilst much larger test corpora, such as the somewhat ubiquitous Govdocs [13], provide material for high volume tool testing, the exchange of small numbers of files exhibiting characteristics of interest seemed to be largely supported via private channels. The OPF established an area on Github as a simple tool to crowd source and manage files of this nature [14]. The only practical constraint is that contributed files must be made available under a CC0 license.

A variety of initiatives relating to Representation Information (RI)[15] were launched during the last year. SPRUCE developed cRIsp in partnership with the UK Web Archive, in order to crowd source RI with as lower barrier to participation as possible [16]. The OPF hosted preservation risk focused pages on its wiki [17]. Jason Scott and the Archive Team launched Just Solve (the file format problem) [18]. And finally, a semantic wiki version of a more formal RI registry was completed by the UDFR project [19]. SPRUCE was not directly engaged with these last three, but it did help to publicise them.

Stack Exchange was quite widely advertised (with support from SPRUCE) as a potentially useful question and answer site for digital preservation topics and via the Libraries and Information Science Stack [20] has accumulated a valuable reference resource for the DP community.

COPTR [21] An ongoing initiative proposed and led by SPRUCE with support from Aligning National Approaches to Digital Preservation is aiming to collate the contents of existing tool registries and reduce some of the unhelpful duplication present in the myriad of existing registries. Four organizations (Open Planets Foundation, Digital Curation Centre (UK), Digital Curation Exchange and Library of Congress /NDSA) who host some of the best existing tool registries have committed to participate in COPTR following production of a wiki based demonstrator [22]. Tool data from these organisation's registries is at the time of writing being collated in advance of production of the COPTR registry.

A single blog post from Barbara Sierman entitled "Where is our Atlas of Digital Damages" [23] prompted two related initiatives. The first captured stories of digital damage, the second focused on images. The latter of these utilized a Flicker group to crowd source images of digital preservation challenges, broken files or "glitch art" [24]. SPRUCE contributed to the latter, publicizing it, collating images from individual contacts and establishing a twitter bot to tweet about new images in the Atlas (which at the time of writing has 132 followers).

3.2 Evaluation and lessons learned

Many of the initiatives were a quiet success, with contributions and interactions from a cross section of individuals from the community. The Format Corpus has gradually received contributions from many quarters (233 commits at the time of writing), and now provides a host of assorted broken files, obsolete files and sets of files exhibiting preservation relevant characteristics (for example the "PDF Cabinet of Horrors" [25]). Contributions of files and usage of files in the corpus was observed during many of the other collaborative events and initiatives described in this paper. Just Solve did not appear to be well supported by the digital preservation community (meaning memory organizations) but delivered the most convincing results of the RI initiatives. cRIsp, launched to an enthusiastic response from the iPRES2012 audience but received a disappointing response from the "crowd". The Atlas of Digital Damages holds 90 images and has 63 members at the time of writing and has received praise in particular as a resource for assisting in communicating the basics of digital preservation visually and in an engaging manner. Although the DP content on Libraries and Information Science Stack was considerable (49 questions) both it, and the proposal for a dedicated digital preservation Stack, were closed after only a short time in beta. Only a quarter of those who signed up to the DP Stack to say they were committing to use the site, actually joined the short lived beta. A poor result, but one that was unfortunately not helped by inflexible moderation and management from Stack itself, that closed the beta without supporting healthy meta discussions with much needed moderation support.

A striking observation for SPRUCE was the substantial lack of formal institutional support for the majority of these initiatives. With a small number of notable exceptions, any success was typically made possible by a cross section of enthusiastic individuals. SPRUCE efforts to enlist support from preservation organizations often fell on deaf ears. When organizational contacts were pushed, it was clear that the unconventional or innovative nature of some of these initiatives was not always viewed favorably. Ownership was also highlighted as an issue.

While organizations were happy to talk the language of collaboration, they were typically reluctant to contribute resources or support to online locations beyond their own organizational URL. This unfortunately explains one of the key reasons behind the current state of online preservation resources where a large number of organizations host very similar information on a variety of topics such as: Getting started in DP, information about DP tools, recommended formats, and so on. As illustrated in the tool registry case (see section 3.1), organizations have not only failed to collaborate in this sphere but they are actively competing with each other. This leaves practitioners struggling to find the support they need. Changing this mindset will be a gradual process requiring direct advocacy and exemplars to illustrate the value of breaking the constraints of walled gardens and competition, and stimulating real collaboration.

Using existing technology and neutral locations to host content related activities was a key theme in the most successful of the initiatives. For example the Atlas utilised Flickr, Just Solve used only a wiki, Format Corpus took advantage of Github functionality. As well as making the setup and management of these initiatives cheap and simple, it provided the community with interfaces and tools with which they were already familiar and were straightforward to use.

4. SUSTAINING THE PRESERVATON ACTIVITY

SPRUCE is building a toolkit of resources that will help managers and practitioners make a convincing case to fund and sustain digital preservation activity. At the time of writing, this toolkit is at an early stage of development, but two ongoing activities are building the evidence base and foundation for this work.

Whilst the main focus of SPRUCE Mashup events has been to understand and solve practical digital preservation challenges, a secondary aim has been to support practitioners in building embryonic business cases. Mashup sessions have included four stages including a benefits brainstorm and alignment exercise, a stakeholder analysis, a skills gap analysis and an elevator pitch. This final stage challenges practitioners to summarize their case in a 60 second pitch to a senior manager. As with the other Mashup activities, results are captured on the SPRUCE wiki [26].

Two SPRUCE funding awards have targeted business case activities, and have taken the form of case studies examining new or expanded digital preservation activity. As well as resulting in sharable exemplar business cases, the process and lessons learnt in their development have been captured. At the time of writing these results are being finalized and will be made available shortly.

5. RECOMMENDATIONS ON CONNECTING THE COMMUNITY

A number of SPRUCE blog posts [27] and presentations have highlighted the challenges of communication and coordination, and what goes wrong when there is inadequate support for these mechanisms that are essential to a healthy community [28]. Duplication and the waste of precious resources are particularly concerning outcomes.

Through its focus on community and collaborative solutions, SPRUCE has made some valuable contributions to the communication required to break away from these negative outcomes. At the lowest level this may simply involve connecting community members with relevant contacts based on an awareness of activity right across the community. For example connecting a user experiencing a particular preservation challenge to an appropriate tool they weren't aware of; making a software developer aware of sources of feedback published elsewhere on some of their code; joining up developers or projects with common aims; heading off new developments, where existing solutions already exist. Connections of these kinds can be important but low key, although they can establish the foundations for far greater partnerships. For example, a weekend twitter conversation between SPRUCE and parties interested in improved format identification led to organic organization of a remote Hackathon, run with members of CURATEcamp [29] that developed new format signatures, facilitated Format Corpus contributions of ebook and video format files and prompted the first step towards opening up the FITS tool to wider community development. The latter of these leading to significant FITS improvements (see section 2.3)

SPRUCE argues that there is a case for a dedicated "digital preservation community manager". SPRUCE has experimented with playing this role and has shown how valuable it is in coordinating activities across the community and in different projects/initiatives. But, as is typical in digital preservation, the role has been funded by a project with a finite lifespan. Ideally this role therefore needs to be adopted by a more sustainable, long term organization such as the OPF, the DPC, or perhaps the ANADP initiative.

6. CONCLUSIONS

SPRUCE activities and related community focused initiatives have met with mixed results so far. Those organizations and individuals that have engaged with SPRUCE activities appear to have got significant value from them. Event feedback in particular was consistently high (for example see feedback responses to SPRUCE Mashups [30]). However a recent SPRUCE Mashup had to be cancelled due to low levels of user registration, suggesting that communication and breaking out to a wider audience remains a significant challenge. Involvement and engagement has not been widespread across the community known to be working in this field.

SPRUCE suggests that barriers to collaboration are gradually being removed and that sufficient value has been obtained from the approaches described in this paper to warrant continued persistence in community collaboration.

7. ACKNOWLEDGMENTS

Thanks to the many people who participated in SPRUCE events and led or contributed to the other collaborative initiatives listed in this paper, without which this work would not have been possible.

8. REFERENCES

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

[2] OPF Tools Registry, http://wiki.opf-labs.org/display/TR/Digital+Preservation+Tool+Registry

[3] Wheatley, P, Middleton, B, Double, J, Jackson, A and McGuinness, R, People Mashing: Agile digital preservation and the AQuA Project. In: *IPRES 2011: 8th International Conference on Preservation of Digital Objects*, 1-4 November 2011, Singapore. http://eprints.whiterose.ac.uk/43837/

[4] Digital Preservation and Data Curation Requirements and Solutions, SPRUCE wiki, http://wiki.opf-

labs.org/display/REQ/Digital+Preservation+and+Data+Curation+Requirements+and+Solutions

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

[6] Friese, Y, Hunger for Automation – The first migration actions in our Rosetta Digital Archive, IDCC 9, 2013. http://www.dcc.ac.uk/sites/default/files/documents/idcc13posters/Poster213.pdf

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

[8] CURATEcamp, http://curatecamp.org/

[9] AV CURATEcamp day, http://wiki.curatecamp.org/index.php/CURATEcamp_AVpres_2013

[10] Webinar: Software Development with OPF, http://www.openplanetsfoundation.org/events/webinar-software-development-opf

[11] Collaborate with the digital preservation community, http://wiki.opf-labs.org/display/SPR/Collaborate+with+the+digital+preservation+community

[12] Blog summary of twitter discussion regarding obsolete Powerpoint 4 (Mac) files, Rusbridge, Chris, http://unsustainableideas.wordpress.com/2012/10/02/powerpoint-4-0-story-so-far/

[13] Govdocs, http://digitalcorpora.org/corpora/files

[14] OPF Format Corpus, Github, https://github.com/openplanets/format-corpus

[15] OAIS standard, CCSDS, http://public.ccsds.org/publications/archive/650x0m2.pdf

[16] Wheatley, PR, Pennock, M and Jackson, AN (2012) cRIsp: Crowdsourcing Representation Information to Support Preservation. iPres 2012, 1-5 October 2012, University of Toronto, Canada, http://eprints.whiterose.ac.uk/74679/

[17] OPF File Format Risk Registry, http://wiki.opf-labs.org/display/TR/OPF+File+Format+Risk+Registry

[18] Just Solve, http://fileformats.archiveteam.org/

[19] UDFR, http://www.udfr.org/

[20] LIS Stack, http://libraries.stackexchange.com/

[21] COPTR blog post, http://bit.ly/14yVzRz

[22] COPTR demonstrator, http://wiki.opf-labs.org/display/coptr/Home

[23] Digital Damages, Barbara Sierman, http://digitalpreservation.nl/seeds/where-is-our-atlas-of-digital-damages/

[24] Atlas, Flickr, http://www.flickr.com/groups/2121762@N23/

[25] PDF Cabinet of Horrors, contributed to the OPF Format Corpus by Johan van der Knijff, https://github.com/openplanets/format-corpus/tree/master/pdfCabinetOfHorrors

[26] SPRUCE Business Case for DP, http://bit.ly/Z9X8xL

[27] SPRUCE blogs, http://openplanetsfoundation.org/blogs/paul

[28] Bacon, J, The Art of Community, O'Reilly, http://www.artofcommunityonline.org

[29] CURATEcamp file id Hackathon, http://bit.ly/Ye6XQk

[30] SPRUCE Mashup Feedback, http://wiki.opf-labs.org/pages/viewpage.action?pageId=13041673