Getting to Digital Preservation Tools that "Just Work"

Andrea Goethals
Harvard Library
90 Mt. Auburn St.
Cambridge, MA 02138
+1 (617) 495-3724
andrea_goethals@harvard.edu

Paul Wheatley
Paul Wheatley Consulting Ltd
Leeds, LS2 9BS
paulrobertwheatley@gmail.com

Stephen Abrams
UC Curation Center, CDL
415 20th Street, 4th Floor
Oakland, CA 94612
+1 (510) 987-0425
Stephen.Abrams@ucop.edu

Janet Delve
University of Portsmouth
Eldon Building
Winston Churchill Avenue
PO1 2DJ
+44 (0)23 9284 5524
janet.delve@port.ac.uk

Ed Fay
Open Planets Foundation
c/o The British Library, Boston Spa
Wetherby, West Yorkshire, LS23 7BQ
+44 1937 54 6013
ed@openplanetsfoundation.org

Cal Lee
UNC Chapel Hill
Manning Hall, Room 212
Chapel Hill, NC 27599
+1 (919) 962-8071
callee@ils.unc.edu

Dirk von Suchodoletz
Faculty of Engineering
Albert-Ludwigs University Freiburg
Freiburg i. B., Germany
dirk.von.suchodoletz@rz.uni-freiburg.de

ABSTRACT
In this paper, we describe a panel that discussed experiences, strategies, challenges and lessons learned maintaining and funding digital preservation tools that are available for use by the digital preservation community. The panel, together with the audience, explored the challenges and discussed potential solutions to developing more robust and sustainable support structures for these tools.

General Terms
Infrastructure, Communities, Digital preservation marketplace

Keywords
digital preservation tools, open-source, enhancements, software development, funding

1. INTRODUCTION
Many of the tools our institutions rely on for digital preservation planning and activities are maintained and funded by single institutions, or reliant on short-term funding. In some cases these tools had project or grant funding that has run out. This presents challenges for the maintaining institutions, funding agencies where applicable, and the digital preservation community as a whole which is reliant on these tools.

- In these cases the maintaining organizations have made their tools available to the community for use but often are not able to keep up with the growing and changing needs of the larger community. Even if the tools are open sourced, the maintaining institutions typically do not have the resources to test, incorporate and document the contributed changes in a timely way.
- Funding agencies and governments want the products they fund to remain relevant and usable and have broad impact well beyond the project funding period. They want to know that any preservation tools they fund can be sustained and improved over time.
- Organizations using the tools can get frustrated when the tools do not improve at the rate they would like and enhancements that they would like are not added. These organizations are not able to adequately improve their preservation practices and infrastructure in a timely way without having reliable tools that meet their requirements.

2. THE TOOLS
The panelists represented a variety of tools:

- BitCurator
- Emulation as a Service (EaaS) Project Tools
- File Information Tool Set (FITS)
- JHOVE
- JHOVE2
- KEEP Project Emulation Tools
- PLANETS Project Tools
- SCAPE Project Tools
- Unified Digital Format Registry (UDFR)

The panelists briefly described the purpose and status of the tools and the key challenges they have faced enhancing, funding, "mainstreaming", governing and providing roadmaps for these tools. In some cases they told success stories where they had

managed to forge sustainable models. The audience was invited to pose questions for the panelists and to contribute ideas for how these tools could be more easily improved and sustained.

3. DISCUSSION

The main points made by the panelists and audience members are summarized here.

3.1 Observations & Comments

- It was asserted that a very large amount of money has been spent on digital preservation tools that have resulted only in demonstrations and prototypes, and not usable code or well-used tools. Several people objected to this statement saying that there are many examples of tools developed within the digital preservation community that may not be perfect but that are widely used.
- The question of whether or not development within memory institutions is a good idea was raised. One responder said that in-house development is done because it is convenient, expedient or existing tools do not necessarily meet local requirements. While this can produce quick results it can pose quality and sustainability problems because of poor quality code, the need to maintain the code base long-term, etc. Others responded that it can be misleading to assume that in-house development within memory institutions is always done by librarians when it may be done by software developers and computer scientists as would be the case when developed by a commercial company.
- It was posited that developers and project managers could be trained in a month to produce good code adhering to best practices but this was disputed by some in the audience.
- The digital preservation community needs to become more proficient at hosting open source tools. While the open source approach for software development is favored by many institutions, some parts of it are not fully embraced because of a lack of resources to perform tasks such as code cleanup, adherence to good coding style, fixing reported bugs, and testing of patches.
- Management within organizations needs to be convinced to spend not only initial development resources on tool development, but also the ongoing maintenance costs which can be orders of magnitude more costly.

3.2 Success Stories

- There is widespread usage of some of these tools (e.g. JHOVE), both in stand-alone mode as well as integrated into larger repository systems.
- Some of these tools (e.g. BitCurator and the SCAPE tools) have found hosting environments (Educopia, OPF) and communities of use after their project funding ended.

3.3 Lessons Learned

- If you make the code available on an open source hosting platform like github be prepared for forks unless you have a clear documented process for developers to contribute code that can be easily integrated into the main branch.
- The key lessons learned from the UDFR registry are the importance of continued synchronization between registries such as PRONOM and UDFR to prevent the immediate divergence of format information, the role of the community and governance to sustain the registry, and dedicated evangelism to maintain interest and use of the registry.
- It can be hard to transition research projects into production tools. The focus of research is on new ideas, and the end products are seen to be publications and advanced degrees, and not necessarily the tools themselves. Often the tools remain prototypes which lack the documentation and commitment to ongoing maintenance that is needed for these tools to be used in production. There is little incentive for researchers to continue to work on tools that no longer are groundbreaking. This is a hard problem because it is appropriate for digital preservation research such as emulation to be conducted by academic or research institutions but for the reasons described it is difficult to transition this research into usable tools.
- Use cases should be well understood before technical development begins.
- Hackfests by design do not contribute to sustainable tools. They have been better at innovating new prototypes than enhancing existing production tools.

3.4 Ideas to Explore

- Up to now libraries and archives have carried the research and development cost for digital preservation. Can we find ways for other institutions (banking, etc.) with more money to contribute?
- Consider contracting development out to commercial companies. An example was given where commercial companies are asked to respond to an RFP developed by a group of memory institutions.
- Solicit and communicate stories where digital preservation tools have been found to be useful to institutions so that the stories can be shared with funders of the tools to encourage continued funding.
- Follow up after Hackfests with attention to clean up (documentation, testing, etc.).
- The digital preservation community could make a statement about adopting software development and testing best practices and funders could mandate that they be followed.