Document Distribution Note
Release to Members: Immediate
Release to Public: 08/08/2013

Collaborative Approaches to Managing File Formats – A day of Action

28th January 2013, Wellcome Trust Conference Centre by James Hall, Borthwick Institute (Intern)

The Digital Preservation Coalition (DPC) met on the 28th of January 2013 to discuss the problems, along with possible solutions, surrounding the management of file formats. The conference was also an opportunity for anyone attending to "bring out your dead" files of any format and to have others look at them to see what can be done to recover them.

The importance of the topic of this event is that there is a vast multitude of file formats in existence, and anyone involved in the preservation of digital data needs to know the best ways of managing all the different formats as well as how to open older formats that are no longer used, and what the most common problems to affect different file formats are. All this helps recovery of data that might otherwise have been lost.

These notes are intended to provide an informal briefing for members of the DPC not able to attend in person. They only represent the sessions that the author was able to attend. For an authoritative and comprehensive report, readers are encouraged to contact the organisers or speakers directly.

The Nature of the Problem – A Presentation by Chris Rusbridge

The problem turns out to be, in its most simple form, what to do when a file format won't open. There are also other problems that can be linked to this such as problems relating to scale that can mean that managing large digital archives is a difficult task, however this was only mentioned and was not to be discussed.

The main point of the presentation was that if any member of the public were to attempt to open a file and find that it wouldn't; what should they do, where should they go. This problem can affect anyone who uses a computer and is especially prominent for those involved in digital preservation.

So there is a problem, and there are possible solutions; five of them. These solutions are: lists of file format specifications, tools for identifying file formats, tools for processing file formats, tools to migrate file formats into readable formats, and emulation tools. This gives rise to another problem; which are the best to choose and which are the best to trust. This problem comes about as there are many of each of these possible solutions. For example there are many lists of file formats: PRONOM, Wotsit, Library of Congress, and Jhove, to name a few, though there are many more. None of these are comprehensive, and with the example of identification tools no two will give the same result for all file formats.

So if all of these different solutions could somehow share their data, along with adding new data that experts, or others, may have, then they could become comprehensive. So this means that crowd-sourcing could be the answer, unfortunately this only works when whoever runs these methods of solving file format problems has the opportunity and wants to be greatly involved in formatting and inputting the data received.

Another option could be to go to the source of the file formats, to companies such as Microsoft, to see how willing they are to provide the original specifications. This relies on the companies having kept the original specifications, or at least having some members of staff who worked on them, who may remember at least part of how to create the file format specifications and is willing and able to provide their time. This would, at least, provide those trying to work with them a starting point.

So why not just compile all these sources together to just solve the problem! Well someone is attempting to, at least with creating a list of file formats. Jason Scott leads an archive team that tries to save file formats from discarded files. He has created a wiki called “Just solve” to just solve the problems, to access the mass of people needed to contribute in order to create a comprehensive list of file formats and is, apparently, dedicated to continue administrating the wiki in order to make it into this comprehensive list.

There are still problems even with this; firstly it is currently only a work in progress and so can be a useful tool to anyone with file format problems, but no more so than any other list of file format specifications. Secondly it requires a huge amount of people to take the time to contribute their knowledge to improve this list when many may not wish to, think that what they know can be useful, or even know that it exists at all. Finally it needs Jason Scott, or others, to dedicate their time to updating and administrating it; this may be happening at the moment but after time could falter. Chris Rusbridge seemed confident that it was one of the best places to go both to contribute to and get information from and would become at least a near comprehensive list of file format specifications.

Developments in PRONOM and DROID – David Clipsham, the National Archives

It is probably best to begin this section by clarifying what DROID and PRONOM are: DROID is a tool used for file identification, and PRONOM is a list of file formats.

The first, and probably most important, development for DROID and PRONOM was the hiring of a full time developer for the two services, the developer being David Clipsham who was the presenter for this topic. This means there is now one person whose job it is to manage and develop DROID and PRONOM, whereas before they were managed and developed by several members of staff when they had the time between their other duties.

Developments to DROID itself have been few over the past year, though there is work on making it compatible with Java 7, for which there have been problems, along with a wiki for DROID 7 for anyone to keep contributing to and a Google group support forum for it. DROID will be from now on available directly from The National Archive website rather than GitHub, where it had previously been available and the source code shall still be available from.

PRONOM on the other hand has had more development. There has been the addition of 946 file signatures, though the priority is on listing file formats that other information about the file types. There is also a new position of Digital Analyst Developer that is soon to be filled and whoever fills the position will be linking PRONOM to other data sources to identify further file formats. This should greatly increase the amount of information available from PRONOM in future and make it a more reliable source of file format specifications with a more comprehensive database in future.

David Clipsham also aimed to show attendees of the conference how to build new file format specifications in the two "Developing and sharing file signatures for PRONOM/DROID" sessions later in the conference. These sessions were done parallel to other sessions and I did not attend either of the sessions on creating file signatures for PRONOM/DROID as I was in the other sessions, though from discussions with the other guests it seems that these sessions were a success and very interesting.

CRISP – Maureen Pennock, the British Library

CRISP stands for Crowdsourced Representation Information to Support Preservation. The best way of explaining what CRISP is would probably be to define the various parts of it.

Crowdsourcing is a way of outsourcing tasks to a large group of people and combine their knowledge. There are two major types of crowdsourcing, these are: public crowdsourcing which can get an awful lot of attention, this has been used for projects such as creating a sound map of Britain where members of the public recorded sounds all over Britain; and expert crowdsourcing that will have a relatively low number of participants but a large amount of combined knowledge, this has been used for projects such as ChemSpider (a database of chemical structures).

Representation Information is what is used to map a data module into a more meaningful form for humans. These are file format specifications or emulation tools that allow humans to access the information from within the data module.

CRISP is therefore is a way of getting representation from a large collective in order to support the preservation of files that may no longer be able to be accessed otherwise. It is a way to get around the problems of file format lists as many have a lack of content, a lack of use, no way of contributing to or no backup, and none are comprehensive. This means that the sustainability of lists of file format specifications such as WotSit and PRONOM have come under question. All these lists also contain duplicate work as several would contain many of the same file formats although some would be the only ones with a specific file format.

CRISP is trying to rectify these problems by getting a large number of people involved and making it simple. It works by people contributing by tweeting a URL link that contains file format specifications along with @dpref, or by pasting a URL into a simple form (this form is available from the CRISP website and can be found by searching the internet for crowd sourced representation information) that only has one mandatory field and all others are optional extra information for if the contributor to add to if they feel like it. This information is then all fed into worksheets that are then checked through for mistakes, such as when someone has tweeted the wrong URL, and is then published on the web archive to be accessed by anyone having file format problems.

The initiative is designed so that anyone can get involved in it and as it is not a funded project it needs many people to get involved with it. This could be a problem as if it gets too few people contributing then it won't be sustainable as a project, though this is a recent project and has been publicising itself a lot through such organisations as the DPC and events such as this conference and so should be able to maintain a large input.

Crowd-Solving the File Format Problem – Paul Wheatley, Leeds University Library

Using crowd sourcing as a method of gaining information seems at least a popular idea, as it has come up in each of the presentations in this conference, along with a good one as the combined knowledge of a group is far greater than that of a single individual or of a few individuals. In the previous presentations it certainly seemed like a popular method of getting information about file format specifications along with other information relating to file formats in order to solve the problems surrounding them. Paul Wheatley suggests that there should be a step back first, that crowd sourcing should be used to discover, in more detail what the actual problems are. It seems a logical progression to find out what exactly the problems are, to think about them, and prioritise them before attempting to solve them.

It is also suggested that anyone with file format problems should share their images of their problem on the atlas of digital damages on Flickr where there are over 80 pictures of different error messages, physical damage to CDs and other portable data devices, obvious damage to files, along with many other ways of showing how digital data can be damage, lost or unsecure. This database of digital damages along with other databases gained by organisations such as the Open Planet Foundation (OPF) who have a "cabinet of [digital] horrors" help those who develop tools to read different file formats or solve problems with different file formats. This works by giving them an idea of all the problems and especially the most common problems so that they can then create ways of solving them. This seems extremely simple, but had barely been touched upon in any of the other presentations and perhaps this is something that should be thought about more before attempting to solve problems with digital data.

The OPF have a wiki (wiki.opf-labs.org) that allows anyone to contribute problems with digital data, along with solutions. These solutions can be both ones that have worked and failed attempts at trying to solve problems. All this information is extremely useful as anyone can see what problems other people are having and help them if they know how, get solutions for their own problems, and find out what not to do as it could further damage the data and there is no point in repeating the mistakes of others if their mistakes are easily accessible. This wiki is a very useful resource as it contains hundreds of issues and solutions. Another website with a similar idea is stackexchange.com where uses can share the same sort of information as in the OPF wiki and can vote on how useful answers were so other users can see which solution is likely to be the best for their problem. These websites are easy to use and contribute to allowing many people to help many other people in an easy and simple way.

Sharing all of this data is essential to helping resolving problems with digital data, because if nobody who has the knowledge and resources to resolve a problem knows about the problem then it will never get fixed. This just seems like an extraordinarily simple idea, but it could so easily be brushed over and forgotten about.

These services also have problems; there seems to be no way to contribute to all at once, if someone has an answer to a problem and adds it to one of these databases, they may not add it to any others and then that information would only exist on one and may not be found by those who need it. This

is a very similar problem to all the tools and lists that have been created to help solve file format problems.

Group Discussion

The group discussion brought up more problems relating to file formats and more ways to resolve the problems.

These solutions including tweeting about any problems as there are hundreds of people on Twitter who would be willing to help find a solution. Others were; going to see people face to face or to workshops that would allow whoever is helping to see exactly what the problem is and be able to help more easily. Though this could cause problems if the file that is experiencing problems may contain confidential information then it isn't possible to have somebody help rescue it if they are not allowed to see the contents.

It seemed that the main conclusion arrived at by this discussion is that currently the best way to resolve file format problems is to make others aware of the problem and ask for help if the solution cannot easily be found elsewhere.

First Parallel Session

The next part of the conference was split into two sessions that ran at the same time as each other, these were titled: "Contributing to Collaborative Initiatives", this session was run by Maureen Pennock and Paul Wheatley; the other was "Developing and sharing file signatures for PRONOM/DROID" this was run by David Clipsham.

I attended the session on "Contributing to Collaborative Initiatives" and so am unable to report on the other session.

For this session we had been encouraged to bring our "dead" files in order to attempt to share information and attempt to recover them. We were given the link for the Sustainable PReservation Using Community Engagement (SPRUCE) project and then given time to get on with it. I cannot give much more information about this session as we were not fed information but given a chance to test SPRUCE for ourselves.

Second Parallel Session

The two sessions here were a continuation of "Developing and sharing file signatures for PRONOM/DROID" (again with David Clipsham), and "Deploying tools for characterisation" run by Carl Wilson. I attended the latter session.

This session explained a further problem faced by digital preservation by showing how much data is created by the Large Synoptic Survey Telescope and giving an idea of how much data is transferred in one second on the internet. This gave some amount of context to how much digital data exists and gave me an idea of how many file formats exist.

Once again the problems surrounding file format identification tools were outlined, this time with examples; showing that a PDF was identified as being version 1.4 by ExifTool and

version 1.3 by Jhove. This led onto the application of File Information Toll Set (FITS); FITS is a tool created by Harvard University library that wraps Droid, Metadata Extra, Jhove, ExifTool, FFident, and File Utility together and runs all through the selected files.

This means that FITS gives the user a consolidated output of all the tools, giving them the combined identifying power of all tools together. This way all file formats that could be recognised by any of these tools would be recognised without having to run each tool individually.

It is good as it has all the benefits of each tool, but it also has all the downfalls and bugs of each tool. Every time two of the tools disagree on any piece of information about a file, they will still disagree, so any problems that any of the tools have will still be present in this wrapped package of tools. Another problem that FITS has is that it currently contains outdated versions of each of the tools as it is not being updated and can only currently be updated by Harvard University who hold the rights to it. Another problem it has is that it takes a lot of time to run as it contains six tools that each need to run over the selected files, this might only be a minor annoyance to any user with only a few files they wish to identify, but to anyone with a lot of files it could be very time consuming.

The web app; Clever, Crafty Content Profiling of Objects (C3PO) created by Petar Petrov is a Mongo Database that can be used to display the output from FITS, at least as I understood it, and allow the user to narrow down their data and view only the parts they want. This seems like an extremely useful tool as when working with FITS it would give the user the opportunity to see how many files of any format they have, what sort of file formats they have, where each of the tools has disagreed on any information about certain file formats, and I'm sure many other things too. This to me seems extremely useful, but I only have a limited understanding of the topic and a lot of the information explained by Carl Wilson was beyond my understanding and so I cannot comment any further on this.

Final Discussion

In this final discussion all were welcomed to ask any questions or voice any concerns they had about the topic of the conference. It turned out to be somewhat depressing. This was due to nobody being quite sure how to proceed from this point to make identifying file formats and knowing where to go to get the best tools to accomplish what was needed easier.

There were many suggestions and some of them seemed to have the majority of guests agreeing that they were approaches that should be acted upon. These included furthering attempts to get information on file formats from companies such as Microsoft, trying to collaborate in order to consolidate all the tools and lists in order to both prevent the duplication of work and to make it easier for those who use the tools to know which one to use. There was also a suggestion of the DPC becoming a lobbying committee to further raise awareness of these problems and attempt to get others involved in resolving them.

Final Comments from Presenters

The presenters were all given time at the end of the day to give some final thoughts and comments. The main consensus of these comments was that everyone who can contribute should contribute to

the knowledge bases and help by giving any information that they are able to, to any of the projects discussed throughout the day.

DigitalPreservationCoalition

Attendance:

28 people booked for the session and 12 completed an evaluation form. There was 1 no show on the day (Caroline Waterloo) and 1 who attended without registering (Angela Dappert).

Type of organisation:

Library	Archive	Educational Institution	Other (specify)	No response
1	8	3	1: Software Vendor	2

Role:

Librarian	Archivist	Conservator	Other	No response
0	5	0	4: Intern, PM, Digitalisation, IT guy (file format research)	3

What are your reasons for attending this event?

- Currently looking at tools for file format ID - thought this would help
- Presenter, Ambition to make the world better
- To understand more about how to add a signature to DROID/PRONOM. Talk with other organisations about their preservation issues.
- Learn more about file characterisation tools
- Learn more about identifying file formats
- To see what is current in file format identification
- Presenting
- Get ideas for dealing with 'problem files'
- Learn about developments, get raw material for DPTP. Broaden my understanding.
- I work in a digital preservation team so this is very relevant for me. Although ultimately as I work for TNA my focus is on DROID/PRONOM
- To learn, discuss and debate on file formats

On a scale from 1-5 how would you rate today's event (actual numbers):

	Not satisfied				Very satisfied	No answer
	1	2	3	4	5	
Relevance to you	0	0	0	6	6	0
Presenter(s)	0	0	0	5	7	0
Level of information	0	0	3	6	3	0
Venue and facilities	0	0	0	5	7	0
Value for money	0	0	0	2	9	1
Overall satisfaction	0	0	1	3	8	0

On a scale from 1-5 how would you rate today's event (%):

	Not satisfied				Very satisfied	No answer
	1	2	3	4	5	
Relevance to you	0	0	0	50	50	0
Presenter(s)	0	0	0	42	58	0
Level of information	0	0	25	50	25	0
Venue and facilities	0	0	0	42	58	0
Value for money	0	0	0	17	75	8
Overall satisfaction	0	0	8	25	67	0

DigitalPreservationCoalition

Which sessions did you find most useful?

The Nature of the Problem – Chris Rusbridge	5
Recent Developments with PRONOM and DROID – David Clipsham, National Archives	5
CRISP – Maureen Pennock, British Library	7
Crowd-solving the File Format Problem – Paul Wheatley, Leeds University Library	8
Discussion	2
First Parallel session:	1
- Contributing to collaborative initiatives	4
- Developing and sharing file signatures for PRONOM/DROID	2
Second Parallel session:	1
- Developing and sharing file signatures for PRONOM/DROID (continued)	1
- Implementing characterisation tools	7
Panel session and discussion	2

What would you be willing to contribute to the collaborative file format initiatives you've hear about today?

- Blogging about my work - must do this more. Liaising with TNA more regarding DROID/PRONOM -highlighting strange files.
- Anything I can, though I do not know how much my limited knowledge can provide
- Support to some leadership
- Development of additional internal signature files for data formats
- Could feedback errors/bugs in use of tools
- Yes but it is not something I would prioritise
- Signatures - once I've learnt how to create them
- Populate PRONOM with any contributions generated as a result of the day.
- If I can help but it's not clear what the payoff is
- Unfortunately I have very little time within my work schedule to meaningfully contribute to initiatives. However if there is anything I can do while I am doing signature research for DROID I will do it. I will also continue to respond to queries sent to TNA about DROID/PRONOM & general file format issues

Whom do you think should take leadership to ensure that common file format issues identified during the day are resolved?

- I have no opinion
- DPC? OPF? BL?
- TNA as PRONOM owner
- DPC
- DPC?
- DPC
- No opinion
- I have to say I think this is unrealistic

What might your institution be prepared to do to ensure that common file format issues are addressed?

- Blog about it?
- Publicise, develop, liaise with interested groups and forums, regarding our Common Pharma File Formats Initiative.
- Provide sample data and test tools

- See if they apply to our collection and feedback info to whoever asks
- Contribute information
- Provide assistance + PRONOM sigs where appropriate
- TNA will always be open to submissions to DROID/PRONOM & we also dispense digital preservation advice about format issues when we are asked. I think perhaps this hasn't been the case in the past but it definitely is now.

Is there anything you will do (differently) as a result of attending this event?

- Try harder - very difficult to collaborate due to time constraints but will try
- Attempt to use new tools.
- Think more about a "Share the data we've got" hackday
- I now know how to submit files to develop new formats. Can see value in proposing to integrate linked data PRONOM with CRISP outputs
- Reconsider security steps at ingest. Try tools. Speak to some about how they are doing things
- Make use of hex editor to identify unknown files
- Start trying to use FITS
- There are some new tools that I'll try. Work with existing systems rather than create new ones
- Will look again at file signatures (though I don't understand them). Will look at C3PO tool when available
- Not really but it has been interesting to refresh my knowledge of what is going on in this area. Even though a substantial part of my job is making sure our collections are identified I have very little time to dedicate to research/catching up on what's going on out there with file format ID/issues

Was there anything else that you would have liked us to have included?

- More of a chance to ID some of the 'dead' files I'd brought with me
- See comments
- More opportunity to practise - I had to miss one of the panel cessions (on C3PO) to go to the DROID ones and would have liked to have done both. Also more speakers who are actual practitioners would have been good.
- It was difficult being torn between the parallel sessions
- More discussion on what the problem actually is. We seem to have moved on from "obsolescence" to "characterisation"
- I think more should be made of the difference between identification and characterisation and what it is that users want from these tools. What level of information about a format?

What did we do well?

- Well organised as always, excellent chairing by William as always
- The presentations were very informative and good at grasping my attention
- Interesting line up of presenters with a good range of views
- Good level of knowledge and content
- Organising such a useful day
- I liked the format. Food was good
- Keep to time. Chaired discussion well
- Good to have it in the first place. Good speakers and time for discussion. Wouldn't expect anything less from DPC
- The event was very well organised and delivered. Good topics and discussion

Document Distribution Note
Release to Members: Immediate
Release to Public: 08/08/2013

Where did you hear about this event?

Archives-NRA	0
Digital preservation list	4
DPC discussion	6
Twitter	0
Preservation Advisory Centre website	0

Other: 2: APARSEN news e-mail and via supervisor

Any further comments

- Possibly a further session about security issues at ingest I.E. Different people donate data - how can we ensure it won't damage our systems/other data
- Thank you all. Really enjoyable event, David
- Rather depressing final session. Inconclusive. Not much 'action' for a 'day of action'

Individuals who wish to remain in contact with the DPC:

- James Hall: jimmy.hall@york.ac.uk
- Ash Hunter: already have details
- Lee Hibberd: already have details
- Jo Gilham: jo.gilham@york.ac.uk
- Ed Pinsent: e.pinsent@ulcc.ac.uk
- Ian Ireland: ian.ireland@ nationalarchives.gsi.gov.uk