# People Mashing: Agile Digital Preservation and the AQuA Project

Paul Wheatley
British Library
Boston Spa
Wetherby
+44(0)1937546254
paul.wheatley@bl.uk

Bo Middleton
Brotherton Library
University of Leeds
Leeds
+44(0)1133436386
m.m.middleton@leeds.ac.uk

Jodie Double
Brotherton Library
University of Leeds
Leeds
+44(0)1133437783
j.l.double@leeds.ac.uk

Andrew N. Jackson
British Library
Boston Spa
Wetherby
+44(0)1937546254
andrew.jackson@bl.uk

Rebecca McGuinness
Open Planets Foundation
Boston Spa
Wetherby
+44(0)1937546254
rebecca@
openplanetsfoundation.org

## ABSTRACT

Manual quality assurance (QA) of digitised content is typically fallible and can result in collections that are marred by a variety of quality and access issues. Poor storage conditions, technology obsolescence and other unforeseen problems can also leave digital objects in an unusable state. Detecting, identifying and ultimately fixing these issues typically requires costly and time consuming manual processes. An inadequate understanding of potential tools and their application creates a barrier to the automation and embedding of preservation processes for many collection owners. The JISC funded [1] Automating Quality Assurance Project (AQuA) [2] applied a variety of existing tools in order to automatically detect quality and preservation issues in digital collections and work to bridge the divide between technical and collection management expertise. Two AQuA Mashup events brought together digital preservation practitioners, collection curators and technical experts to present problematic digital collections, articulate requirements for their assessment, and then apply tools to automate the detection and identification of the content issues. By breaking down the barriers between technical and non-technical practitioners, the events enabled grass-roots digital preservation collaboration between the two communities. This paper describes the AQuA Project's novel approach to agile preservation problem solving and discusses the incidental benefits and community building that this strategy facilitated.

## 1. THE CHALLENGE

Creating a digital object via digitisation is prone to mistakes and the introduction of quality issues. In recent years, increasingly ambitious digitisation programmes (such as the recent JISC eContent Programme [3]) have turned digital content creation into a mass production activity. Known quality issues include missing pages, duplicate pages, incorrect de-skew, out of focus images, incorrect or incomplete metadata, the infamous "thumb in picture" and a variety of other processing or corruption problems have been introduced with mass digitisation.. (see Figure 1). Collection curators and technical staff are now faced with detecting mistakes and quality issues on a large and ever expanding scale.

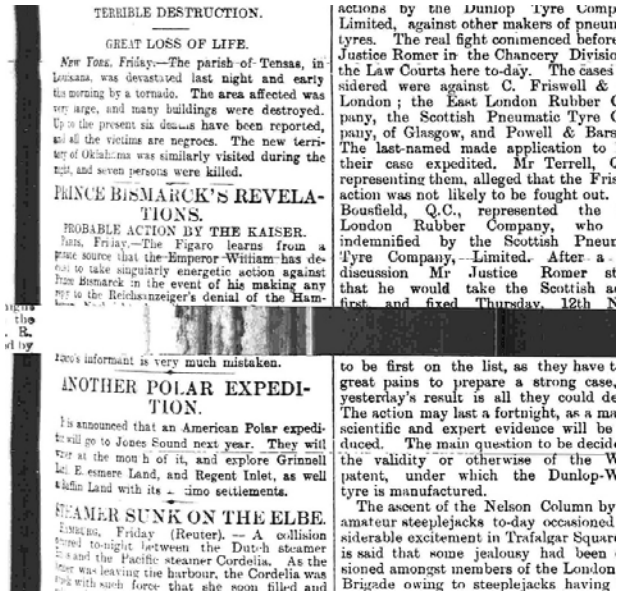
Undetected digitisation quality issues can become digital preservation issues later in the lifecycle and these are often problems that are hard to rectify once the source material has been re-shelved and the digitisation activity has been closed. With only manual content checking to mitigate these issues, there is a serious risk of erroneous or poor quality content making it through to the end user's screen. Timely and automated identification of problematic scans would enable re-digitisation at comparatively low cost as opposed to costly retrospective rescanning years later.

Preserving an existing digital object (whether digitised or born digital) typically requires a number of processing steps, before it can be safely placed into a digital repository. Each of these individual operations has the potential to malfunction, sometimes with disastrous results for the resulting preservation effort. Whenever digital content is acquired, created, moved, un-packaged, processed, migrated, curated, repackaged or otherwise changed, problems can occur and collection damage can result. Culprits include software bugs, network dropouts, full disks and human error.

Detecting these issues requires thorough content checking at key lifecycle stages. File hashing and file manifests can support efficient digital object integrity checking, but many operations in a preservation workflow will legitimately alter the digital objects, resulting in a necessary recalculation of file hashes. Manual checking of content is a typical method of catching systematic errors, but suffers from a number of drawbacks. Human effort can be costly and this makes it difficult to scale this approach up to support the QA of large collections. A visual check can sometimes be subjective and QA problems can be quite subtle and hidden. Sampling approaches can also be used, but this leaves


*iPRES2011*, Nov. 1–4, 2011, Singapore.

blind-spots where issues can remain undetected. A more thorough and automated QA check may prove to be unaffordable unless built into core business practice for the collection managers and their institutions.



TERRIBLE DESTRUCTION.

GREAT LOSS OF LIFE.

*New York, Friday.*—The parish of Tensas, in Louisana, was devastated last night and early this morning by a tornado. The area affected was very large, and many buildings were destroyed. Up to the present six deaths have been reported, and all the victims are negroes. The new territory of Oklahoma was similarly visited during the night, and seven persons were killed.

PRINCE BISMARCK'S REVELATIONS.

PROBABLE ACTION BY THE KAISER.

Paris, Friday.—The Figaro learns from a private source that the Emperor William has decided to take singularly energetic action against Prince Bismarck in the event of his making any reply to the Reichsanzeiger's denial of the Ham-

actions by the Dunlop Tyre Comp Limited, against other makers of pneu tyres. The real fight commenced before Justice Romer in the Chancery Divisio the Law Courts here to-day. The cases sidered were against C. Friswell & London ; the East London Rubber pany, the Scottish Pneumatic Tyre pany, of Glasgow, and Powell & Bars The last-named made application to their case expedited. Mr Terrell, representing them, alleged that the Fris action was not likely to be fought out. Bousfield, Q.C., represented the London Rubber Company, who indemnified by the Scottish Pneu Tyre Company, Limited. After a discussion Mr Justice Romer st that he would take the Scottish a first and fixed Thursday, 12th

burg's informant is very much mistaken.

ANOTHER POLAR EXPEDITION.

It is announced that an American Polar expedition will go to Jones Sound next year. They will winter at the mouth of it, and explore Grinnell Land, E. esmere Land, and Regent Inlet, as well as Baffin Land with its Esquimo settlements.

STEAMER SUNK ON THE ELBE.

Hamburg, Friday (Reuter). — A collision occurred to-night between the Dutch steamer Holland and the Pacific steamer Cordelia. As the latter was leaving the harbour, the Cordelia was struck with such force that she soon filled and

to be first on the list, as they have t great pains to prepare a strong case, yesterday's result is all they could de The action may last a fortnight, as a ma scientific and expert evidence will be duced. The main question to be decid the validity or otherwise of the W patent, under which the Dunlop-W tyre is manufactured.

The ascent of the Nelson Column by amateur steeplejacks to-day occasioned siderable excitement in Trafalgar Squar is said that some jealousy had been sioned amongst members of the London Brigade owing to steeplejacks having

**Figure 1: A portion of a digitised newspaper image exhibiting damage arising during post processing**

If manipulating content increases the chance of damaging it in an unforeseen way, leaving it untouched over time raises the potential for obsolescence issues to be encountered. The critical questions facing digital preservationists include: will this content render correctly on the user's computer? Is it likely to render correctly in 5, 10 or 20 years time? If not, why not and what can be done about it? A variety of more technical proxies are typically raised in an effort to begin to answer these challenging questions. What is the file format? Does this file validate to its file format specification? Are there any external dependencies? There is therefore a need to assess or characterize digital content in order to gain a better understanding of its properties, analyze potential risks and inform subsequent preservation planning and remedial preservation treatments.

Quantifying the incidence and impact of these problems is difficult, particularly with regard to quality rather than preservation issues. The authors had encountered quality or processing problems at their respective institutions. However, organizations are usually not pro-active about broadcasting what might unfairly be seen as bad news stories. Anecdotal evidence suggested these issues were not uncommon elsewhere and documented QA work such as that by Riley and Whitsel, 2005 [4] also implies the existence of a challenge to be met. But prior to the AQuA Project, the real significance of these issues for memory and higher education institutions remained somewhat unclear as this was a collection management issue that was rarely discussed openly and more importantly discussed between technical and non-technical staff.

## 2. POTENTIAL SOLUTIONS

The authors felt that potential existed to apply existing software tools to many of the problems outlined above in addition to engaging collection curators in preservation planning who are normally excluded from hackathons. Several pre-requisites in terms of knowledge, access to data and expertise would need to be met for significant progress be made during the events:

1. A good understanding of the specific QA and preservation challenges faced by institutions.
2. Access to samples of problematic digital collections where these challenges were present, to support solution testing
3. Knowledge of likely toolsets that might provide useful solutions
4. Effort to progress solutions

The authors identified a potential funding stream from the Joint Information Systems Committee (JISC) that matched well with the problem space. The University of Leeds led a successful bid, partnering with the University of York, the British Library and the Open Planets Foundation. Funding conditions restricted the project length to 6 months and a modest budget. These constraints would make it difficult to gather QA and preservation problems and associated content, discover likely toolsets, apply them to the problems and evaluate the results all within the limited project length. Recruitment of project staff would be challenging due to a very short project lead time, and finding sufficient staffing expertise to meet the pre-requisites listed above nigh on impossible. The collaborators (represented by the authors of this paper) therefore pursued a more agile approach which would engage with practitioners and experts from other institutions in 2 mashup events that would each be 3 days in length. This would have the added benefits of gaining buy in to project outputs by getting potential users involved in creating the solutions, while facilitating knowledge sharing and collaboration.

## 3. THE AQUA MASHUP APPROACH

The AQuA approach has its origins in the Hackathon [5], where software developers meet up to solve technical challenges over a short period of time. Hackathon events have become increasingly popular in recent years as a way of removing the overhead of traditional project based development and enabling rapid prototyping and development through a combination of collaboration and friendly competition. The digital library community has begun to embrace the Hackathon concept, with projects such as DEVCSI [6], working actively to develop a technical community via supporting activities such as Hackathons and programming challenges.

The advent of the open data and linked data approaches has encouraged the creation of a similar event model to the hackathon but with a focus on exploiting open interfaces, mashing up data from several sources and providing new and often innovative services. Data Mashup [7] events, like Hackathons, typically provide supportive environments for participants to collaborate in small teams and compete to win challenges.

The Unconference [8] approach, demonstrated in the repository community by the CURATEcamp [9] events, seeks to break away from the pre-planned and often rigid structure of typical face to face meetings and support a more agile and bottom up approach.

The AQuA Project Mashups drew on elements of these existing approaches, while adding some new concepts in order to meet the challenges described above. Rather than being purely technically focused AQuA invited software developers as well as digital preservation practitioners and curation staff and gave them specific roles to play during the events. Instead of setting challenges for the attendees, we asked them to bring along issues they needed solutions to be developed for and spent time capturing and recording these in order to support future work. Although not quite a Hackathon, Mashup or Unconference, the authors settled on describing the events as Mashups.

# 4. THE AQUA EVENTS

The AQuA Project organized two Mashup events. The first was held at Weetwood Hall in Leeds for 18 attendees in April 2011. The second event was held at the British Library in London for 30 attendees.

## 4.1 Mashup Event Planning

A substantial amount of pre-event planning focused ensuring the attendees understood the expectations from the team and that the event ran smoothly.. A strict "no observers" rule required that every attendee had to either bring collection content with them and champion it at the event, or have the skills to play a developer role.

## 4.2 Mashup Event Format

The first day of each AQuA Mashup focused on setting the scene and capturing the digital preservation challenges that would be tackled. After a brief introduction to outline the structure of the event the focus was quickly placed on the participants, who gave lightning talks to the group. Attendees playing the role of Collection Owners were asked to bring along samples of problematic digital collections and talk about the issues they had. Technical attendees were asked to talk about their skills, experience and interests. Over lunch the facilitators matched up the attendees into teams, ensuring that each Collection Owner was supported by a Developer. Working in small groups, and in some cases individual teams, details of the collections samples brought to the event were discussed. QA and preservation issues were identified and recorded, and potential avenues to explore in solving the challenges were noted. From this brainstorm, teams were able to select a challenge they were interested in tackling and begin work on it. The Developers began to seek out useful software tools to apply in order to tackle the identified issue, while the Collection Owners recorded the results of the brainstorming and progress made with solutions.

The second day had much less structure, allowing the Developers plenty of opportunity to progress their technical work, while liaising closely with the Collection Owners on their teams. Collection Owners had the opportunity to work further on capturing their preservation issues and broadening the perspective to explore contextual challenges. Institutional constraints would inevitably impact on the technical solutions being developed and how they could ultimately be embedded into existing workflows.

The third day initially provided some time to wrap up development work, focus on capturing, and where possible visualizing, the results. A small group brainstorm was facilitated to consider the next steps once the event had concluded. Lightning talks to report back results to the group were followed by opportunities to evaluate the solutions and discuss the AQuA approach and events. Prizes for the best work by a Developer and the best work by a Collection Owner were voted on by the attendees themselves.

A strong focus was placed on capturing all event outputs on either the project wiki or Git code repository as they were developed or understood. A key concern of the authors in focusing project development effort into short lived Mashup events was that useful work might easily be lost if not captured straight away. Post event wiki gardening was planned to ensure a clear and meaningful record of results was captured and publicly available [2].

# 5. PROJECT RESULTS

## 5.1 Collections, Issues and Solutions

The AQuA Project wiki [2] contains descriptions of the outputs of the project events. Each of the digital content samples brought along to an AQuA event is listed and described under the Collections section. This described the basic details of the collection and provided a high level description of its characteristics. Preservation or QA challenges were termed "Issues" and listed under a related wiki page. These issues were related to specific collections using hyperlinks. All Issues were recorded in a standard proforma, capturing a detailed description of the preservation or QA challenge as well as possible approaches for tackling it. Where AQuA was able to explore a solution to the issue, a further "Solution" wiki page was produced. This described the approach taken and provided a link to the Solution itself and contained review notes on how well the Solution had solved the related Issue. The resulting network of Collections, Issues and Solutions provides a permanent record of the AQuA results.

## 5.2 People Mashing

A key aim of the project was not only to develop some solutions to the QA and preservation challenges identified, but also to facilitate collaboration, knowledge sharing and hopefully lasting relationships between the attendees.

Mahey and Walk [10] identify a need to break developers out of constrained development and problem solving cycles and exploit their wider capability while also developing them as individuals. They go on to describe how face to face events, amongst other possibilities, can facilitate collaboration, knowledge sharing and develop a support community. AQuA took this further by breaking down the barriers between technical and none-technical staff, creating an environment where participants were happy to ask questions without fear of judgement, and encouraging agile problem solving. AQuA dubbed this approach "People Mashing".

Non-technical staff developed skills to articulate issues and technical staff were able to develop preservation tools that would have an impact beyond the event. Participants commented in both a discussion at the end of the second AQuA event and in anonymous feedback that they were keen to encourage and maintain the community that the events had begun to establish.

# 6. REVIEW AND LESSONS LEARNT

## 6.1 Feedback, Review and Refinement

Survey Monkey was used to gather feedback from attendees at both events and time was made at the end of the London Mashup to discuss as a group how the event went and what the organizers and attendees should do next. Several planning and review meetings were held between events where the schedule was

revised and each session updated to take advantage of the experience of running the first event and the feedback received. Scaling up aspects of the first event to work with double the number of participants for the London Mashup was a key challenge.

## 6.2 What worked well

The popularity of the events and the presence of an array of both preservation and quality issues in participants' collections vindicated the project focus. Indications that these issues were actually a significant issue for many institutions were confirmed.

The events yielded a significant number of functional preservation solutions, some prototypes that required further work and a number of promising problem/solution explorations that can all be found on the wiki. Several participants were keen to stress that they would be taking home workable solutions that they could put into practice straight away. Peer review by the collection owners of the solutions developed for them was largely positive, although many noted that more development and support would be needed and illustrates a long-term challenge from the events to continue testing and refinement of tools in production environments.

Capturing a record of each Mashup using Collection/ Issue/ Solution proformas worked well in providing structure and clear aims for the events while ensuring that the valuable work performed was not lost at the end of the Mashups. The resulting documentation should be useful in supporting adoption and re-use of AQuA results by the Open Planets Foundation and other interested parties.

Many attendees gave very positive feedback about the collaborative and inclusive nature of the events. Several comments focused on the benefits of the agile approach to working. One attendee commented "Putting 30 people into a room, some with problems and some who can write solutions is extremely eye opening. I've learnt that free from restrictions on infrastructure and process … prototyping can solve a varied number of non-trivial problems quickly."

A number of the solutions developed took a genuinely innovative approach, such as the RDF visualization of characterization results [11] produced at the London Mashup. Encouraging participants to work on new problems, often outside their comfort zone, and discuss their approaches with others helped to facilitate this.

## 6.3 What worked less well

Collection Owners weren't challenged enough on the second day when the focus was on progressing the technical solutions. More sessions focusing on preservation planning and next steps would have made better use of their time and given them a tangible piece of work to take back to their institutions.

Following the first Mashup, it was clear that development time at the event needed to be maximized and as a result lightning talks for reporting back were minimized. This was probably a mistake as it would have increased interaction between the teams and sharing of ideas between developers.

Formal checkpoints between Developers and Collection Owners may have helped to reduce the length of development cycles, although many teams worked closely enough for this not to have been a significant issue.

Conference venues were used to host both events which precluded late night coding sessions. Several of the Developers were disappointed not to be able to keep working into the evening on the second day. Focusing the first evening on a meal and social event to encourage networking and the second as all night hack time would have been a good compromise.

Three days is also a long time for participants to abandon their day job and join a Mashup or Hackathon event. A number of interested parties would like to have joined one of the AQuA events but were unable to convince their manager to release them for the duration. On the other hand, fitting a structured event into less than three days would have been challenging. Project funding to cover accommodation and catering helped participants to justify time on AQuA as there were few additional costs to them.

Good Wi-Fi is essential at an event of this kind. Signal strength problems were encountered at the London event and a backup plan had to be put into action at short notice. Having a reserve ready to go is recommended.

# 7. NEXT STEPS

At the time of writing the AQuA Project Team is planning a follow up event that will focus on evaluating adoption of project results. It will consider what barriers there are to further development or re-use of the tools with the aim of targeting effort from the Open Planets Foundation, JISC and others on appropriate support activities.

Given the success of the AQuA events in beginning to build a community of digital preservation practitioners, maintaining the momentum with further face to face events would be desirable. All but one of the attendees who completed the feedback survey for the London event stated that they would like to attend more mashup events of the same AQuA format. Since the completion of the AQuA Project itself, the OPF and the Digital Preservation Coalition have announced a new event that has adopted the AQuA mashup format and approach [12].

# 8. REFERENCES


[1] *Grant 15/10: JISC infrastructure for education and research programme*, http://www.jisc.ac.uk/fundingopportunities/funding_calls/2010/10/grant1510.aspx

[2] *Automating Quality Assurance Project*, http://wiki.opf-labs.org/display/AQuA

[3] *JISC eContent Capital Programme*, http://www.jisc.ac.uk/fundingopportunities/funding_calls/2011/06/econtentcapital.aspx

[4] Riley, J and Whitsel, K, 2005. Practical quality control procedures for digital imaging projects, OCLC Systems & Services Volume: 21 Issue: 1. http://www.dlib.indiana.edu/~jenlrile/publications/imageqc/qc.pdf

[5] *Hackathon*. Wikipedia. http://en.wikipedia.org/wiki/Hackathon

[6] *Developer Community Supporting Innovation Project*, http://devcsi.ukoln.ac.uk/blog/about/

[7] *Mashup (digital).* Wikipedia. http://en.wikipedia.org/wiki/Mashup_%28digital%29

[8] *Unconference*, Wikipedia. http://en.wikipedia.org/wiki/Unconference

[9] *Curate Camp*, http://curatecamp.org/about

[10] Mahey, M and Walk, P. 2010. *Why UK Further and Higher Education Needs Local Software Developers.* Ariadne, Issue 65. http://www.ariadne.ac.uk/issue65/mahey-walk/

[11] Cliff, P and Fay, E. *tiff2RDF - visualising image collection consistency*. http://wiki.opf-labs.org/display/AQuA/tiff2RDF+-+visualising+image+collection+consistency

[12] *OPF and DPC Hackathon: Practical Tools for Digital Preservation*, http://www.openplanetsfoundation.org/community/opf-events/hackathon-practical-tools-digital-preservation