CRISP: Crowdsourcing Representation Information to Support Preservation

Maureen Pennock
The British Library
Wetherby
West Yorkshire
0044 1937 546302
maureen.pennock@bl.uk

Andrew N. Jackson
The British Library
Wetherby
West Yorkshire
0044 1937 546602
andrew.jackson@bl.uk

Paul Wheatley
University of Leeds
Leeds
West Yorkshire
0044 113 243 1751
p.r.wheatley@leeds.ac.uk

ABSTRACT
In this paper, we describe a new collaborative approach to the collection of representation information to ensure long term access to digital content. Representation information is essential for successful rendering of digital content in the future. Manual collection and maintenance of RI has so far proven to be highly resource intensive and is compounded by the massive scale of the challenge, especially for repositories with no format limitations. This solution combats these challenges by drawing upon the wisdom and knowledge of the crowd to identify online sources of representation information, which are then collected, classified, and managed using existing tools. We suggest that nominations can be harvested and preserved by participating established web archives, which themselves could obviously benefit from such extensive collections. This is a low cost, low resource approach to collecting essential representation information of widespread relevance.

Categories and Subject Descriptors
H.3.m [INFORMATION STORAGE AND RETRIEVAL]: Miscellaneous

General Terms
Management, Documentation, Design, Experimentation, Human Factors, Verification.

Keywords
Representation information, crowdsourcing, digital preservation, web archiving, community engagement, social networking.

1. INTRODUCTION
Representation information (RI) is widely acknowledged as essential for digital resources to remain accessible into the future. The internet is one of the best sources of representation information, which is scattered around web in a variety of personal and organizational websites. Yet finding and navigating this information is not straightforward. We know from experience that the identification and collection of RI is highly resource intensive. Organizations collating and maintaining resources themselves have struggled to resource this work. The PADI site remained a key source of information on digital preservation for a number of years but was eventually closed and web archived when the overhead of maintaining the information became too great Furthermore, we know all too well that websites themselves are far from permanent. Vital online information about preservation tools and file formats can be transitory: here one day, 404'd the next.

Existing online community-created resources that link to online representation information sources go some way to addressing these challenges, though they are typically spread around quite thinly, with much duplication. A number of formal RI registries have been built but are sparsely populated, despite widespread community acceptance of the importance of RI, and there appears no overall consensus on the extent of RI required to support long term preservation and access.

The scale of this challenge requires a coordinated and collaborative effort across the wider preservation and curation communities, to establish an inclusive and (semi-)automated solution for RI collection and preservation. Encouraging more coordination will reduce duplication of resources and maximize effort in creating and maintaining the resources we need to make preservation effective.

2. DEFINING SHARED REPRESENTATION INFORMATION REQUIREMENTS
Representation information facilitates the proper rendering and understanding of content. In OAIS terms, RI is a distinct type of information object that may itself require representation information [1]. It can exist recursively until the knowledge base of the designated community dictates no further RI needs be recorded. As a result, the extent, size and boundaries of an RI collection are potentially immense. The vague boundaries and immense potential scope of an RI collection may be one of the reasons why RI collections have been so difficult to establish. We contend that the precise scoping of a core RI collection is the key to maximizing community input and establishing a successful well-populated collection. 'Core shared RI' is that which is most broadly relevant to the widest possible user base.

Brown, in his 2008 white paper on Representation Information Registries, defines two classes of structural RI: Descriptive and Instantiated [2]. These are defined respectively as information that describes how to interpret a data object (e.g. a format

specification) and information about a component of a technical environment that supports interpretation of the object (e.g. a tool or platform).

Descriptive structural RI such as format specifications, which are universally relevant for all objects of a given format regardless of the environment in which content has been used, are core shared RI. These are therefore our starting point for a core shared RI collection. We consider tools that support interpretation to be secondary shared RI, as whilst they are essential, their relevance is more likely to differ for different collecting institutions.

Format specifications are not just necessary for future access, but also contemporary preservation planning. The current SCAPE (Scalable Preservation Environments) project[1], funded by the EU, needs to collect format information to assist preservation planning and other processes. It is clear that the number of stakeholders with a vested interest in contributing to a shared format specification registry is extensive.

3. CURRENT INITIATIVES

The case for representation information has been well made elsewhere and will not be repeated here [3]. Numerous online RI resources have been established by the preservation community, each with slightly different foci, granularity and coverage. Here we introduce some of the key current resources.

3.1 Format registries

Several different format registry initiatives have been established in the preservation community over the past decade. These are now roughly consolidated into two initiatives: the UDFR and the proposed OPF format registry.

UDFR combines content previously collected in PRONOM and GDFR in a single, shared semantic registry [4]. Functional development is led by use cases. The system is highly structured with a well-defined ontology. It is publicly available and awareness of the resource is high, though the contributor base appears relatively low.

The proposed OPF format registry ecosystem will link existing sources of representation information and enable users to create linked data collections based on the information currently distributed across disparate resources [5]. Proposed components include the PLANETS core registry and PRONOM, in conjunction with a proposed 'registry of registries'. The success of the project is dependent upon successful population of supporting registries.

Whilst both are labeled 'registries', a corresponding repository element is typically able to store RI directly.

3.2 Tool registries

A number of tool registries have been established and shared across the digital preservation community. The following list is not exhaustive but exemplifies the range and scope of currently available online tool resources.

The Digital Curation Centre (DCC) Tools & Services site identifies and links out to a large number of curatorial tools for deposit/ingest, archiving/preserving, and managing/administering repositories.[2] Many of the tools were developed by and are well established in the preservation community. The site is managed by the DCC, though community nominations are encouraged by email.

A community wiki of precision digital preservation tools is provided by the OPF through the OPF Tool Registry.[3] This includes tools developed in the AQuA and SPRUCE mashups, as well as the SCAPE project.[4] Tools are categorized by function and simple user experiences described. Source code for some of the tools is hosted directly on the wiki. The site is manually populated by a small geographically distributed group of digital preservation professionals. Membership of the group is open to all, and all members have editing rights.

The Digital Curation Exchange Tool list is a flat though extensive list of links for tools and services relevant to digital preservation.[5] It includes many 'supporting' services and developer tools absent from other lists, such as storage solutions, core utilities, and office plug-ins. Description is minimal. The list is maintained by the membership, which is open to all.

Finally, an inventory of Partner Tools & Services is available from the NDIIPP website, which briefly describes and shares information about tools and services used in NDIIPP.[6] Entries are not categorized though the context of use is clearly identified. Some content is hosted directly on the site though many entries point to external links.

3.3 Other initiatives

The Library of Congress' (LoC) Digital Formats Sustainability site contains extensive format descriptions relevant to the LoC collection.[7] Format versions have their own entries. Descriptions link to format specifications published online and identify sustainability issues. Format specifications published on these pages are harvested by the LoC web archiving program. The site is maintained by LoC staff though community input is welcomed.

Twitter provides an unofficial forum for sharing information about digital preservation resources online, as do many personal collections of bookmarks hosted in social bookmarking tools.

Other file format resources are maintained outside of the digital community, the most comprehensive being Wikipedia. Wotsit.org maintains a similarly impressive array of format information. These appear to have been under-utilized in most digital preservation registry initiatives to date.

4. DRAWBACKS OF CURRENT APPROACHES

4.1 Lack of content

Almost without exception, the tool and format registries provided by the digital preservation community suffer from inadequate amounts of content. This observation seems at odds with the effort that has been devoted to existing registry initiatives where the focus has typically been placed on designing detailed data models

[1] SCAPE project website: http://www.scape-project.eu/

[2] DCC Tools & Services resource: http://www.dcc.ac.uk/resources/external/tools-services

[3] OPF Tool registry: http://wiki.opf-labs.org/display/SPR/Digital+Preservation+Tools

[4] AQUA http://wiki.opf-labs.org/display/AQuA/Home; SPRUCE http://wiki.opf-labs.org/display/SPR/Home.

[5] Digital Curation Exchange: http://digitalcurationexchange.org/

[6] NDIIPP Partner Tools & Services list: http://www.digitalpreservation.gov/tools/

[7] Digital Formats Sustainability: http://www.digitalpreservation.gov/formats/

and building systems to manage and publish the resulting RI. The result is theoretically capable replicas and systems, which are largely empty of their most important feature: the data. We suggest that the biggest challenges facing these initiatives are not related to managing or publishing RI, but in capturing and recording it

4.2 Duplication and reinvention

A considerable number of DP community-created web pages list digital preservation tools. Most have some unique entries, though many contain entries duplicated across other entries (albeit with slightly different descriptions). The result is that users are unable to easily find the tools they need and precious DP community resources are spent needlessly reinventing the wheel or aspects of the wheel. For example, more than one institution has developed its own checksum tool for digital preservation purposes.

4.3 Lack of use

It is undeniable that despite the massive investments made to establish representation information registries, the current initiatives are under-utilized. Much effort has been devoted over the past decade to developing new digital preservation tools and approaches, but insufficient attention has been paid to the needs of the users. The result is a mismatch between preservation tools, and user requirements.[8]

This may be down to insufficient understanding about use cases and requirements. RI repository use cases are undeniably unclear, though it may also be a case of chicken and egg: which comes first, the RI, or an understanding of how RI should be used? Perhaps the community still has insufficient detailed understanding of how RI fits into a preservation strategy and the relationship between RI requirements and different preservation strategies. Or is it perhaps a case that we have not yet reached the stage, from a temporal perspective, where we need much more than file format specifications. Whatever the reason, it will only be solved by greater collaboration and engagement with the user community.

5. ADVANTAGES AND DISADVANTAGES OF A COMMUNITY & COLLABORATIVE APPROACH

A community-based approach to collecting and managing representation information has potential to resolve many of the drawbacks in current approaches. For example:

- It is user focused, so the final data is more likely to meet the needs of end users and is therefore more likely to be used.
- It puts the initial focus on capturing content, thereby increasing the flow of incoming data and increasing the chances of reaching that critical mass.
- A single, concerted and collaborative effort will minimize efforts wasted through duplication and reinvention
- The end result is likely to be of a higher quality with less effort from any one participant (and therefore more distributed costs), as it has been refined by the crowd, with a higher number of contributions and expertise from a wider cross section of the community.

The risks of a communal and collaborative approach however, cannot be overlooked:

- There may be difficulty reaching consensus about the level and granularity of RI resources required.
- Without sufficient refinement by a number of contributors, content may be of poor quality.
- Success depends on reaching a critical mass of contributions. If this is not reached, the solution may hold few advantages over other approaches.

Individual organizations that have hosted community discussion forums have typically struggled to reach a critical mass of contribution to make the forums a success. This has been the experience of even those with sizeable and engaged communities such as the Digital Curation Centre, the Digital Preservation Coalition or the Open Planets Foundation. The recent proposal for a digital preservation themed Stack Exchange site seeks input and engagement from across the international digital preservation community. While still requiring further support to reach a functional beta stage at the time of writing, it has been successful in soliciting widespread international support and shows promise for a broad community driven approach. However, it has yet to be seen whether this widespread 'show of hands' will translate into active and participatory membership.

Collaborative collection approaches must target content at a level of granularity most likely to be relevant to the majority, in order to engage as broad a swathe of the community as possible. We propose that success at this level is most probable if it is a) simple, b) does not require extensive input from contributors, and c) makes use of existing tools and networks. Our answer to this is CRISP.

6. CRISP: A COMMUNITY APPROACH TO COLLECTING REPRESENTATION INFORMATION

CRISP utilizes the power and wisdom of the crowd to identify and share online resources of representation information, beginning with file format specifications. We have selected format specifications as they are the lowest common denominator of representation information: as previously argued, files of a given format and version share a core RI requirement for the same format specification, regardless of the more extensive environment in which they were produced (the RI for which is more likely to differ across different environments and uses). Access to format specifications is necessary for all preserving institutions. This initiative is therefore broadly relevant and with a clearly defined scope.

CRISP is in the early stages of development. The main objective of the initiative is to address the gaps in collection content currently evident in global format registries managed by the digital preservation community. We will, in essence, get the data. Once we have it, we will store it in a preservation-capable store. We expect to expand our scope to preservation tools in the future, but the initial focus is limited to an achievable and easily defined set of data, namely the format specifications. Our solution has yet to be fully implemented but we are confident that it is sufficiently robust and reliable to serve our needs.

Content will be crowd-sourced via two mechanisms that will make it easy for interested parties to participate. The primary method of submitting information is via an online form, hosted on

[8] Mashup events have provided a useful forum in which to engage with considerable numbers of users, capture and publish their requirements and explore solutions by utilizing existing open source software).

the Open Planets Foundation website.[9] Minimum data requirements have been set purposefully low. The only compulsory fields are a) URL and b) tag(s), though additional fields are available and contributors are encouraged in particular to tag their entries by format to support classification and curation at later stages. Registration is not required prior to nomination. This, alongside a small minimal requirement for input and a simple, straightforward interface, ensures the barriers to participation are as low as possible.

The form links directly to a Google spreadsheet, which is publicly available so participants have access to all nominations and are able to re-use the data if desired. A small number of super-users will be identified to promote the initiative and curate the spreadsheet. De-duplication algorithms will eliminate multiple entries for the same resource whilst maintaining the tags applied by different proposers to ensure broad classification relevance.

The second, more experimental approach is via mentions of the @dpref Twitter account. Tweets to this account will collated and added to the spreadsheet. We were hoping to use a social bookmarking system like Delicious or Diigo, but we found them to either be unreliable or have too high a barrier to submission. Both also failed to have suitable methods for exporting the curated dataset. A Google spreadsheet offers the functionality and access that is needed.

We propose that the repository element of the equation is served by the existing power of well-established web archiving systems, which will harvest sites listed in the spreadsheet and store them as part of an RI 'collection'. This will, in the first instance, be undertaken by the UK Web Archive. As the spreadsheet will be publicly available and the contents broadly relevant, we hope that the initiative will be more widely adopted by the global preservation community in the near future and that other web archiving institutions will also avail themselves of the resource. By remaining neutral in terms of ownership, it is anticipated that buy in across the community will be increased.

We are not the first group to propose use of web archives for collecting representation information. The subject has been raised more than once in the IIPC Digital Preservation Working Group. More recently, the web archiving team at the Library of Congress has begun archiving web pages identified in the Digital Formats Sustainability site. However, web archiving alone will not solve the challenge of resourcing and broad relevance to the community. Crowdsourcing has been used by cultural heritage institutions to meet other objectives in recent years, for example correcting OCR text, and has successfully increased the amount of manpower available to an initiative whilst simultaneously raising awareness of the content and increasing use. There is no reason to believe this approach will be any different.

Our proposal is simple, and we are confident that its simplicity will be the key to its success.

7. ISSUES

The main advantages of our approach stem from its low cost, clearly defined scope, and broad relevance. However, we appreciate that it is not without issues:

- There is the risk that the community will not get on board with the initiative. Without a critical mass of participants, the initiative will not reach the critical mass of content required.
- Champions and curators are required for sustained community engagement and curation of the data prior to harvest: there are costs associated with this
- Legislative issues may prevent interested web archives from sharing their RI collections publicly, lowering the incentive for input from non-crawling institutions
- An automated solution is required to clearly identify openly licensed content that can be freely republished
- There is a risk associated with using free online tools and services, which may be withdrawn or the data lost with no compensation or backups.

These issues will be managed as the initiative develops.

8. CONCLUSION

CRISP offers a low cost and simple solution to the problem of identifying and collecting essential representation information commonly required by the collecting institutions. The main risk lies in garnering sufficient community engagement to ensure RI sources are nominated. If the community does not buy-in to the proposal, then population of the established representation information repositories will continue at the very slow pace we have seen to date. Similarly, without better community engagement, it will be difficult to clearly identify use cases and encourage use of the repositories. Without this, they will fail to be truly integrated into the preservation solutions currently being developed. CRISP is the first step in solving that problem.

9. ACKNOWLEDGMENTS

Our thanks to colleagues in the web archiving and digital preservation teams at the British Library for their input to this idea.

The SPRUCE project is funded by JISC.

10. REFERENCES

[1] OAIS standard: http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683

[2] Brown, A. 2008 'White Paper: Representation Information Registries' Planets project publication http://www.planets-project.eu/docs/reports/Planets_PC3-D7_RepInformationRegistries.pdf

[3] See for example the OAIS model and Brown (2008), op cit.

[4] UDFR http://www.udfr.org/; GDFR http://gdfr.info/; PRONOM http://www.nationalarchives.gov.uk/PRONOM/

[5] Roberts, B. 2011 'A New Registry for Digital Preservation: Conceptual Overview'. http://www.openplanetsfoundation.org/new-registry-digital-preservation-conceptual-overview

[9] The form is available at http://www.openplanetsfoundation.org/testbed/digital-preservation-reference-stack-collection-form