The International Journal of Digital Curation
Volume 7, Issue 1 | 2012

Towards the Development of a Test Corpus of Digital Objects for the Evaluation of File Format Identification Tools and Signatures

Andrew Fetherston and Tim Gollins,
The National Archives

Abstract

The digital preservation community currently utilises a number of tools and automated processes to identify and validate digital objects. The identification of digital objects is a vital first step in their long-term preservation, but the results returned by tools used for this purpose are lacking in transparency, and are not easily tested or verified. This paper suggests that a test corpus of digital objects is one way of providing this verification and validation, ultimately improving trust in the tools, and providing further stimulus to their development. Issues to be considered are outlined, and attention is drawn to particular examples of existing digital corpora which could conceivably provide a useable framework or starting point for our own communities needs. This paper does not seek to answer all questions in this area, but merely attempts to set out areas for consideration in any next step that is taken.

Introduction

There is a growing demand from the digital preservation community for greater transparency regarding the accuracy, verification and testability of results obtained from file format identification and validation tools.[1] Tools such as PRONOM, DROID, JHOVE, Unix File Utility, FIDO are already widely used in digital preservation, digital records management and digital curation workflows, and provide a valuable resource for individuals and organisations charged with managing digital objects. However, the basis for the results returned by such tools remains largely hidden from end users. Even where tools have been independently tested and results documented, it has been impossible for other individuals to repeat and verify the results.[2] In a large part this is due to such testing being conducted on collections of files which are unavailable publicly, such as collections on a personal hard drive, or files which, although publicly accessible (e.g. they have been harvested from web-based content), may not be individually identifiable, and cannot necessarily be guaranteed and verified as the same files which were used for the original testing.

This lack of transparency and the absence of consistent, repeatable and testable evidence means that confidence in the results obtained by such tools relies on trust in the institution which supplies them, rather than empirical evidence of their accuracy. While there is no suggestion that such tools are not put through numerous internal testing and evaluation procedures by their developers, or that the independent external testing has not been rigorous and extensive, information about such procedures remains inaccessible to the wider community. This in turn acts as a barrier for further developments and improvements of the tools themselves, since they cannot be evaluated in any assured or accepted way.

One method to provide this additional testing and validation would be the creation and maintenance of a test corpus of digital objects, with known provenance and identity, against which new tools, identifications and characterisation information could be tested. Such results could then be used as a baseline, against which further improvements and enhancements of file format identification tools could be tested and evaluated in an objective, open, testable and repeatable manner (Garfinkel, Farrell, Roussev and Dinolt, 2009).

This document will outline the benefits of such a resource; discuss the practicalities of setting up and maintaining a test corpora; suggest ways in which the resource could be structured for maximum benefit to the digital community, and highlight the factors that need to be considered in order to create a useful and useable product.

[1] See recent postings on the Open Planets Foundation blogs: http://www.openplanetsfoundation.org/blogs/2011-02-17-call-test-set-files

[2] Examples of independent testing of file format identification tools can be seen in Ford (2011) and van der Knijff (2011).

 doi:10.2218/ijdc.v7i1.211

The Test Corpus

Why do we Need a Test Corpus?

Accurate file format identification must be the starting point for any digital preservation process. Without knowing the format types of the digital objects we are responsible for, it is difficult for any organisation or individual to develop and implement plans for preservation and access. The positive identification of digital objects allows those of us entrusted with their long-term preservation to develop targeted approaches to their management, ensuring finite resources are expended only on managing and maintaining the file formats we possess now, or are likely to possess in the future. Likewise, for organisations accepting digital objects as part of an accessioning process, the ability to identify potentially 'difficult' digital objects, from a preservation/presentation stance, can empower those repositories to make informed decisions regarding their acceptance.

Despite the obvious benefits in utilising automated file format identification tools and processes in order to acquire this identification information, until now there has been little attempt by those involved with the development of such tools to provide quality assurance and validation for the identifications assigned. While collaboration between developers and end users has allowed for the development of ever improving identification methods, and an ever expanding list of identifications, much of the testing and provenance information remains inaccessible to the wider community, leading to a disjunction between the results generated by such tools and the perceived accuracy and confidence in those results.

The development of a test corpus of authenticated digital objects with a known provenance, from which regular and predictable identifications could be generated, is a potential solution to this perceived failing. Such a resource will provide an objective means by which to judge the accuracy of any identification, and will:

- Allow users to have a greater confidence in the results they achieve with the tools by enabling them to conduct their own validation and testing;
- Enable developers working to improve file format identification techniques to rigorously test their work against known objects;
- Encourage collaboration and dialogue in the digital preservation community, with an emphasis on developing better file format identification methods.

Outline Methodology

The digital preservation community is not alone in observing the benefits of a test corpus for the evaluation of tools. Other groups which are actively engaged in this area include the digital forensics, information retrieval, and text-compression communities.[3] As a result, there is a potential to develop and learn from existing

[3] For a digital forensics perspective, see S. Garfinkel, P. Farrell, V. Rousseu, and G. Dinolt (2009); for information retrieval see the Text Retrieval Conference website at http://trec.nist.gov/; and for text compression communities see The Canterbury Corpus at http://corpus.canterbury.ac.nz/index.html

methodologies and practices that have already been developed in these areas. In order for this to happen however, the digital preservation community must be clear in what it is seeking to achieve with such a corpus, and consider the practicalities involved in its creation and ongoing development. In particular, the following areas need to be addressed:

- Source and location of corpus
- Selection criteria of contents
- Provenance and metadata
- Persistency
- Usability
- Security and access control
- Maintenance
- IPR and copyright concerns

Source and location of corpus

The corpus should be comprised of a known and well provenanced representative collection of digital objects, from which recognised and repeatable results could be achieved. As such, its main purpose would be to provide an effective method for the testing and validation of object identification and validation tools, and any file signatures such tools may utilise. This collection would be open for anyone to utilise, but be added to only by a trusted and validated process, following pre-determined procedures. A trusted location for the collection is required, one which is accessible to any interested stakeholder. A web hosted site is the natural choice.

A number of digital corpora already exist, varying in size, content and purpose.[4] Some are relatively small in size, while others contain thousands of files. While there are obvious uses that those of us concerned with file format identification can put these resources to, they do not adequately answer the specific needs identified above, nor should we expect them too, since they have been created with other purposes in mind. For example, metadata and provenance information is lacking for many files, and there is no attempt to capture a wide sample of formats, or examples of different versions of a format, which is crucial for digital preservation purposes. We believe that the digital preservation community therefore cannot rely on utilising existing corpora to conduct file format identification tools testing, but instead must create a new corpus of objects for that specific purpose.

[4] Examples include the Waterloo Repertoire, which provides a collection of image files for a quantitative comparison of image compression programs at http://links.uwaterloo.ca/Repository.html; the Image Spatial Data Analysis Group Conversion Software Registry, which provides information on file format conversion processes, with some sample formats at http://isda.ncsa.uiuc.edu/software.html; and the Digital Corpora Govdocs1, which contains nearly 1 million digital objects comprised of files gleaned from the .gov web domain at http://digitalcorpora.org/corpora/files.

 doi:10.2218/ijdc.v7i1.211

Provenance/selection criteria

Digital objects in the collection would need to have associated metadata for authenticity purposes. This could comprise the characteristics or instance properties of the digital object and details about the creating application and operating system the object was produced in. A list of metadata requirements should be agreed by the community managing the test corpora, which any subsequent additions to the collection would need to adhere to.

Selection criteria for the content of the corpora should also be agreed upon by the individuals or organisations responsible for the content. Issues such as ownership, content type, and file sizes need to be considered, and a methodology constructed to ensure a representative sample of files is collected.

However, there is also a danger that enforcing too strict an approach to file type submissions will reduce the potential value of the corpora. For example, by insisting on having a known provenance and associated metadata for every digital object, there is a potential risk that the result will be a homogenous collection of widely available file formats, reducing the diversity of the collection and potentially limiting the value of the corpus for the digital preservation community and tool developers.

A compromise position may be the most useful approach to adopt in this area. One possible solution would be to create a core collection of files with strong provenance and metadata information, while requiring less provenance requirements for the remaining collection.

Validation

A collection of well provenanced digital objects will provide a resource that can be utilised for validation testing as well as identification testing purposes. Ensuring details of the creation of the digital object are captured and recorded in the associated metadata for that object will provide a resource that can be interrogated and tested by validation tools. Any discrepancies or anomalies in the validation of particular objects can then be investigated with the knowledge of how the object was created. Characterisation of digital objects, and the investigation and identification of different characteristics within the same file format, will also be aided by such a resource.

Persistency

The test corpora and the results obtained from the objects need to be persistent and retrievable over time. Accurate version control is therefore a fundamental requirement for any corpora. The content of a test corpus will, by necessity, expand over time to incorporate new file formats, but the availability of all previous versions of the collection will enable continued verification of the results obtained by file format identification tools at any given point during the history of the corpora. This improves the integrity of the results, allowing them to be verifiable and repeatable. Utilising a resource similar to ApacheSubversion, for example, could provide the tools and processes necessary to implement this feature.[5]

[5] See http://subversion.apache.org/#site-overview for details of an open sourced version control system which could potentially be used in this space.

At the same time, the results obtained by varying tools and versions could also be documented and stored. This would clearly show any areas where a tools performance has changed over different versions, and would also indicate potential areas for future development and improvements.

Usability

Consideration of the internal organisation of the digital corpora is required, and this needs to be agreed by all contributors. For example, how should the test objects be stored and listed in the collection? By the order they were submitted to the test corpus? Or grouped by the submitter? Or ordered and grouped by some other criteria, such as file format, file type or MIME type?

While the strength of the test corpus is based on the authenticity and size of its collection, the ability for users to 'cherry pick' particular formats of interest to them is also a valid use of the service, and should be factored in to the structure of the corpora. Creation of sub-sets of data will allow tools to be tested against certain criteria, for example against audio formats only, or against digital objects larger than 1MB in size. This will improve the practical use of the collection for tool development and testing purposes.

While the test collection should be large enough to include versions of all file formats currently identifiable by automated tools, including example file types from different sources which may contain variations within the internal structure of the file, the size of the corpora should not become so large that it impedes the usefulness and accessibility of the content. There is work to be done in this area to assess and determine what constitutes a significant sample of files in order to produce significant and trustworthy results. An added complication is that such a figure will vary depending on the file format itself, the number of software applications it can be created by, and the variability of its content and internal structure.

The value of the test corpus could be greatly enhanced by the inclusion of known 'problem' files, such as corrupted files, encrypted or password protected files, and raw bit streams. Tools could therefore be tested not only against the positive identifications achieved, but also against other criteria, such as avoiding false positive identifications, or identifying corrupt files.

As mentioned previously, the ability to provide a standard test collection for tools is one of the prime motivations for this resource. As such, the corpora could be used to test file identification and validation tools against various categories and criteria, such as accuracy of identification, speed of processing, or percentage of false positives. An agreed baseline or standard for published test results would also need to be agreed (i.e. tested on a standalone PC, with x value RAM and y value processor speed).

Security

The integrity of the collection needs to be strictly controlled so that files can only be added by the previously described procedures. While any user would be permitted to download files from the corpus, only trusted users would be able to add or upload to it, following prescribed procedures, thus controlling the content and maintaining the

integrity of the collection. Fixity checking will also be required to confirm to the user that what they have downloaded is an accurate copy of the corpora at a particular time.

The potential increase in threat to heritage collections from cyber attack needs to be considered to ensure that such a corpus does not become a means by which computer viruses or malware could enter or be spread around such organisations. All files submitted to the corpus should be quarantined for a period of time (30 days appears to be an accepted length), to ensure any danger of virus/malware is minimal.

Standard backup and secure storage processes will need to be in place to ensure the business continuity of the collection in the event of disaster.

Maintenance

The object corpora could potentially grow to a large size, and may require management for an indefinite period of time. Commitment to maintaining and providing access to such a resource is not to be taken lightly, and the organisation or community responsible for the upkeep of the resource must ensure that processes and strategies are in place to allow continuous access over the long-term.

There must also be a commitment within the community to use the corpus as the default standard for testing and tool refinement purposes. The value of the corpus is intrinsically linked to the amount of use it is put to, and its value may be lessened should it fail to be used by a significant proportion of the community.

IPR and copyright

Copyright and Intellectual Property Rights (IPR) relating to the digital objects and their content is an important consideration for this type of test corpus, and must be dealt with before any files can be made publicly available.

Professional advice in this area will need to be sought as part of the initial planning process to ensure that owners and contributors to the digital object collection do not open themselves up to unnecessary risk in this area. It is currently unclear whether a Creative Commons Licence could be used for this purpose, and whether it is robust enough to deal with all potential eventualities in this space. Certainly, any contributors to the test corpus would be required to provide a declaration regarding their rights to distribute the particular digital objects they wish to contribute, deferring responsibility away from the service providers.

Data protection issues and the sensitivity of information contained in the digital objects is also a consideration. Clearly, no material of a sensitive or personal nature should be part of the corpus, which will help to ensure that data protection and sensitivity issues are avoided.

Access rights

The issue of access control has already been touched on, but there is another aspect of this which should be considered separately in relation to IPR and copyright. There is a potential for any licensing along the lines of creative commons to become void if individuals or organisations utilise the digital objects covered by such licensing for

commercial gain. This is certainly a potential outcome if developers are encouraged to utilise the test corpus in order to improve signature development and automatic object identification tools. Again, professional advice and guidance should be sought regarding this issue.

Metrics

Having outlined the various practicalities to consider in this space, it is worth briefly looking at the potential measurements and metrics which could be used on the test corpus, and which would have a beneficial impact on the assurance, validation and future development of file format identification tools.

One assessment would be in the area of precision and recall, something which is well known and utilised in the realm of information retrieval, and which it is proposed could be easily applied to this particular area of research.[6]

In basic terms, precision and recall analysis is a method to determine the accuracy of results obtained from a particular tool or resource, based on the ability of that tool to return or identify relevant results.

Recall can be defined as a measurement of the ability of a tool or resource to correctly return all relevant results within a given set of data, regardless of the number of incorrect results that may also be returned. So, as an example, a file format identification tool that correctly identified all JPEG files in a sample of digital records would have a recall value of 1.0, regardless of whether it also incorrectly identified a number of Tiff files as JPEG.

Precision, on the other hand, can be defined as a measurement of the ability of a tool or resource to return only correct results. So, again using the same example as before, a file format identification tool which correctly identified only JPEG files as JPEG files, and did not misidentify any other file formats as JPEG files, would have a precision score of 1.0, regardless of whether it had also missed some JPEG files altogether or incorrectly identified them as different file formats.

Precision and recall results can offer useful information when assessed individually, but in many cases their real value is seen when they are taken together to form a single measurement, known as the F-measure or F-score. This is a combined score based on both precision and recall results, and is used to give an overall assessment of a particular tool.

However, this method is not particularly well suited for the purposes of the assessment of file format identification tools, since it gives equal weight to both precision and recall results. In the area of file identification, it is our belief that the non-identification of files (a true negative) is preferable to a misidentification (a false positive). While a failure to return a positive identification may result in some form of manual intervention and further assessment of the digital object, a false positive identification could potentially result in the mismanagement of the object, with associated consequences in terms of its long-term preservation. Therefore, in order to assess file format identification tools and their results in a more accurate and realistic

[6] Much of the following discourse draws on Manning, Prabhakar & Schütze (2008).

 doi:10.2218/ijdc.v7i1.211

way, more emphasis needs to be given to the precision value in any assessment. The F (0.5) measure, a variant of the F-score, which weights precision twice as much as recall, may be a suitable alternative to the F-score in this type of evaluation. However, there is a danger in assuming this (or any other) value without due consideration, and more research could be undertaken to determine a suitable weighting score for the particular assessment required in this area.[7]

For classification purposes, the results generated by a tool can also be categorised as one of four possible outcomes:

- True positives: correct result
- True negatives: correct absence of a result
- False positives: object identified incorrectly
- False negatives: absence of a result where object should have been identified.

This type of analysis, sometimes known as sensitivity and specificity analysis, and related to precision and recall, can help add further depth and clarification to an assessment of the relative merits of different file format identification tools.

Sensitivity relates to the tool's ability to identify positive results, while specificity focuses on a tool's ability to accurately identify negative results. In the case of file format analysis, this could mean the ability of a tool to correctly distinguish between a valid file and a corrupted or incomplete file of the same format. So, again to use JPEG as an example, a tool could be assessed on its ability to correctly identify all JPEG files in a collection (its sensitivity value), but it could also be assessed on its ability to return an accurate non-identification for corrupted JPEG files within the same collection (its specificity value).

As with all of these assessments and analysis, they can only really be as accurate as the test data itself, which is why the authenticity and provenance of the test objects is of prime importance in this work. It is also one of the main reasons why current digital corpora may fall some way short of providing the raw data for such testing and evaluation.

While we may not be in a position to demand full provenance of all the digital objects added to the test corpus, for reasons previously outlined, we should certainly look to establish a core set of objects which can be used as the basis for the establishment of a baseline of measurements and evaluations. By being assured of where, when and how this core selection of objects was created, more in-depth analysis of file format identification tool performance can be achieved. Pooling the results of various tools tested against such a core sample can allow for the generation of 'standard scores' for each criteria, which could then be utilised by tool developers to indicate where particular tools were falling below the standard, and therefore help focus and direct targeted development work to these areas. Such scores and values could then be extrapolated across larger collections with an added degree of confidence in the results generated.

[7] Further information concerning the F-measure and its variant forms can be found at van Rijsbergen, (1979).

Conclusions

This document has touched on some of the areas for consideration regarding the development of a test corpus of digital objects for validation and research purposes.

The development of a digital corpus has the potential to offer significant benefits to the digital preservation community. The test corpus proposed will provide the quality assurance on file format identification methods and tools that is currently lacking in this area, and will be accessible for all groups and individuals concerned with the management of digital records. The additional transparency in the processes by which file format identification tools assign IDs will greatly enhance the confidence of users of such tools. The test site should also act as a stimulus for further file format research and signature development, and provide an environment for collaborative work within this field.

However, the work involved in setting up such a resource is not to be taken lightly, and will require a reasonable investment in time and resources, both in the initial planning and set-up stage, and in ongoing maintenance and support. It is also imperative that a consensus is achieved across the digital preservation community as to the use and upkeep of the resource, with the usefulness of the resource directly correlating with the number of individuals and organisations making use of, and adding content to, the corpus.

Acknowledgements

The comments and postings which appeared on the Open Planets Foundation blog under the title ‘Call for a Test Set of files’, acted as a stimulus for this paper. In particular, the role of Dirk von Suchodoletz, the initiator of the blog post, is acknowledged. In addition, the authors would like to acknowledge input and comments received from various colleagues within the digital preservation community which have helped, directly or indirectly, to influence this paper. In particular, the input of Jay Gattuso of the National Library New Zealand, who participated in several preliminary discussions regarding the use of digital corpora for digital preservation purposes, is acknowledged.

References

Ford, K.M. (2011). The application of file identification, validation, and characterization tools in digital curation. (Thesis). University of Illinois at Urbana-Champaign, USA. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/24301/Ford_Kevin.pdf?sequence=1

Garfinkel, S., Farrell, P., Roussev, V., & Dinolt, G., (2009). Bringing science to digital forensics with standardized forensic corpora. Paper presented at the Digital Forensics Research Workshop (DFRWS). Montreal, Canada. Retrieved from http://www.dfrws.org/2009/proceedings/p2-garfinkel.pdf

Manning, C.D., Prabhakar, R., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. Retrieved from http://nlp.stanford.edu/IR-book/

van der Knijff, J.M. (2011). Evaluation of characterisation tools. Part 1: Identification. SCAPE (Scaleable Preservation Environments) Project. Retrieved from http://www.openplanetsfoundation.org/system/files/SCAPE_PC_WP1_identification21092011_0.pdf

van Rijsbergen, C.J. (1979). Information retrieval. London, UK: Butterworths. Retrieved from http://www.dcs.gla.ac.uk/Keith/Preface.html