The National Archives

How to research and develop signatures for file format identification

November 2012

This publication is available for download at nationalarchives.gov.uk.

Contents

1 Introduction

Accurate file format identification is the starting point for any digital preservation process. It is difficult to manage the long-term preservation of digital objects without understanding their identity and characteristics. Automatic file format identification is supported by a number of tools, including PRONOM and DROID, developed by The National Archives[1].

PRONOM is an online technical registry, aiming to provide objective and definitive information about file formats, file format signatures, software products and other IT components. The data in PRONOM supports long-term access to, and preservation of, digital objects of cultural, historical or business value.

DROID (Digital Record Object Identification) is an open-source software application developed by The National Archives. DROID uses information from the PRONOM registry to identify file formats[2].

PRONOM and DROID must evolve if they are to remain accurate when new file formats emerge or old ones are revised. The National Archives digital preservation team regularly update the PRONOM registry with new and improved file format signatures. However there are a vast number of formats in existence, and this limits the coverage we can achieve. Recognising this, we actively encourage external submissions of file format signatures to the PRONOM database. We hope that this will:

- increase the usefulness of PRONOM and DROID by improving their coverage and accuracy
- respond to particular areas of interest within the digital preservation community
- capture information about specialist file formats not usually encountered within our own collections

2 What is the purpose of this document?

This document describes the procedures and practices The National Archives follows when conducting file format research. It provides good practice advice and guidance to researchers

[1] nationalarchives.gov.uk/information-management/our-services/dc-file-profiling-tool.htm
[2] nationalarchives.gov.uk/documents/information-management/droid-how-to-use-it-and-interpret-results.pdf

active in this area. We hope it will encourage users of the DROID and PRONOM services to conduct their own signature research and share this with the wider digital preservation community. This will improve the ability of the digital preservation community to identify digital files, through extending the coverage or increasing the efficiency of PRONOM and DROID.

This document gives a high-level overview of our approach. For a detailed description of internal file format signatures, the specific syntax employed by DROID and PRONOM, and an explanation of the rules governing the structure of the DROID signature xml files, see the Digital Preservation Technical Paper on Automatic Format Identification Using PRONOM and DROID[3].

3 Who is this document for?

This document is aimed at:

- researchers in the field of digital preservation
- anyone with an interest in accurate, automatic identification of digital file formats

The guidance offered here assumes some understanding of the internal structure of digital files. The reader should be familiar with the concept that all digital files, irrespective of their format, are composed of byte sequences, and that specific sequences of bytes may be characteristic of particular file types.

This guidance will help you collaborate with us to improve the accuracy of PRONOM and DROID, through contributing new signatures or improving the accuracy of the existing signature base.

4 What is a file format signature?

Before describing the methods we employ to conduct file format signature research, we must establish what we mean by a signature. File format signatures may be external or internal to the file. External signatures are filename extensions (such as `.doc` or `.pdf`) that can provide clues to the file format. Internal signatures are essentially byte sequence(s) common to particular file

[3] Digital preservation technical paper 1: Automatic format identification using PRONOM and DROID

formats which can be used to recognise digital objects encoded in that format. The aim for tools such as PRONOM and DROID is to provide unique signatures which identify specific formats.

DROID currently identifies digital objects using three distinct methods:

- external filename extensions
- internal byte sequences
- container signatures (which use a combination of internal and external elements)

5 External signatures

External signatures are derived from elements of the file which are found outside the bitstream of the file itself. The most useful example of this is the filename extension (or file extension) of the object.

File extensions are the portions of file names which appear after the full-stop (for example, word.doc). Common extensions include `.txt`, `.xls` and `.ppt`.

While file extensions can help identify the format of the object, it is not their primary function. Rather, their main purpose is to indicate to an operating system which software package should be used to open the file. As a result, there are a number of drawbacks to using file extensions as a means of file format identification:

- file extensions are not standardised or unique: unrelated file formats can have the same file extension
- different versions of the same file format often have the same file extension, but may be significantly different from one another
- file extensions are easily lost or changed, either by human interaction (including human error) or by automated processes

PRONOM stores information on known file extensions, which DROID uses to help identify file formats. Because of the issues outlined above, we regard these as ‘weak’ identifications, and place a greater emphasis on identifications achieved using internal signatures.

6 Internal signatures

Internal signatures are created from elements of the internal structure of a digital object. This internal structure is often governed by rules specific to a particular file format. If those rules can be identified, and if they can be shown to be unique to a particular format or version of a format, it should be possible to express them in terms of an internal file format signature. Some formats have been designed to include a pattern precisely for the purpose of identification which makes the process of signature development much easier. However, most formats were not designed in this way, and therefore the identity of the format must be inferred from common patterns which appear in the internal structures of the files.

All digital objects, irrespective of their internal data structures, can be represented as a series of bytes. Most internal signatures are composed of one or more byte sequences. These sequences may be simple, continuous strings of bytes, but more often consist of complex patterns of variations, gaps and alternative values.

A signature byte sequence is modelled by describing its starting position within a byte stream, the values of the bytes themselves and any positioning relative to the file or other patterns. Some signatures have more than one internal pattern, for example, a file format may contain a key sequence at the very start of the file and another key sequence at the very end of the file. An example of a simple internal file format signature is shown in Figure 1.

Figure 1: Example simple internal format signature for GIF 1987a

File Format: GIF 1987a

Offset: Absolute Beginning of File (Absolute BOF)

Sequence: 474946383761

Offset: Absolute End of File (Absolute EOF)

Sequence: 3B

Before this pattern can be interpreted and used by DROID, it must be converted into xml. The PRONOM registry automatically generates xml signature files extrapolated from the data contained in its database.

Figure 2 shows the same GIF 1987a file format identifier after its transformation into an xml DROID signature file.

Figure 2 Example DROID xml file signature sequence for GIF 1987a

```
<InternalSignature ID="18" Specificity="Specific">

     <ByteSequence Reference="BOFoffset">

          <SubSequence MinFragLength="0" Position="1"
                  SubSeqMaxOffset="0" SubSeqMinOffset="0">

               <Sequence>474946383761</Sequence>

               <DefaultShift>7</DefaultShift>

               <Shift Byte="37">2</Shift>

               <Shift Byte="38">3</Shift>

               <Shift Byte="46">4</Shift>

               <Shift Byte="47">6</Shift>

               <Shift Byte="49">5</Shift>

               <Shift Byte="61">1</Shift>

          </SubSequence>

     </ByteSequence>

     <ByteSequence Reference="EOFoffset">

          <SubSequence MinFragLength="0" Position="1"
                  SubSeqMaxOffset="0" SubSeqMinOffset="0">

               <Sequence>3B</Sequence>

               <DefaultShift>-2</DefaultShift>

               <Shift Byte="3B">-1</Shift>

          </SubSequence>

     </ByteSequence>

</InternalSignature>
```

Different versions of essentially the same signature may be created if a single signature is not flexible enough to capture all the variation which occurs in real files of that format. All of these variations need to be incorporated into a signature or signatures in order to achieve accurate file format identification.

7 Container file signatures

Container file signatures are used to identify formats such as ZIP and OLE2 compound files which act as containers for other digital objects. Container signatures may need to combine various elements of the internal structure of a file (such as byte sequences and internal folders) to construct a valid signature.

Having identified the presence of a container, the remainder of the signature might look for files expected within the container that identify its type, and may even look for further byte sequences within those files. These methods allow the identification of complex file formats and structures.

Figure 3 DROID XML for Microsoft Project 4.0/95 OLE2 container file identification

```
<ContainerSignature Id="4000" ContainerType="OLE2">
  <Description>
    Microsoft Project 4.0/95 OLE2
  </Description>
  <Files>
    <File>
      <Path>CompObj</Path>
      <BinarySignatures>
        <InternalSignatureCollection>
          <InternalSignature ID="310">
            <ByteSequence Reference="BOFoffset">
              <SubSequence Position="1" SubSeqMinOffset="40"
                   SubSeqMaxOffset="1024">
                <Sequence>
                  14 00 00 00 'MSProject.Docfile.4' 00
                </Sequence>
              </SubSequence>
            </ByteSequence>
          </InternalSignature>
        </InternalSignatureCollection>
      </BinarySignatures>
    </File>
  </Files>
</ContainerSignature>
```

Figure 3 provides an example container file signature in xml form for the Microsoft Project 4.0/95 format. A comprehensive breakdown of the signature is beyond the scope of this document; however the example indicates how container signatures work.

The format is first identified using a generic OLE2 signature. DROID then refers to the smaller subset of OLE2 container type file signatures to provide a more complete and accurate identification.

In the example shown, DROID looks for a folder named CompObj within the OLE2 container. Once it finds this folder, it searches the contents for the byte sequence

```
14 00 00 00 4D 53 50 72 6F 6A 65 63 74 2E 44 6F 63 66 69 6C 65 2E 34 00
```

(which would be represented in ASCII text as '`14 00 00 00 MSProject.Docfile.4 00`').

8 Tools and resources for signature research

A number of widely available tools and services can aid the file signature researcher.

8.1 Hex editors

A hex editor is a computer software program which can display the bytes that make up any digital object. Bytes can be viewed with their hexadecimal values, together with a rendering of those bytes as ASCII characters. Loading files into a hex editor enables the individual bytes of each object to be examined, and should enable patterns or byte sequences indicative of an internal signature to be compared and contrasted with other digital objects of the same format.

There are a number of freely available hex editors. The majority offer additional features of value to file format researchers, for example, the ability to compare byte sequences between files and the ability to search for specific hexadecimal and ASCII character strings. HxD Hex Editor is one such tool[4] - section 9.3 gives an example of its use.

[4] HxD Hex Editor mh-nexus.de/en/hxd/

8.2 Web analytic tools

Web analytic tools can help identify areas of interest to the user community. The National Archives uses Webtrends[5] software to analyse requests to the online PRONOM service. This helps us understand which file formats are commonly searched for or viewed, and provides a useful input into the prioritisation process (see section 9.1). Free or open-source tools are also available, for example, Google Analytics[6].

8.3 Internet search engines

Google's filetype operator can help locate sample files with a particular filename extension. For example, searching for `filetype:pdf` returns results with the filename extension `.pdf`. Right-click on the link(s), and select 'Save … As' to download the files.

Files of a particular format are often stored in server directories accessible over the internet. The directory pages typically include the phrase Parent Directory and the file extension. This provides a useful means of finding collections of files with a known extension.

For example, a search for the term `"Parent Directory" tif` will return a high proportion of directories containing tif format files, usually distinguished by a title of the form Index of directory name. Note that the double quotation marks are required around 'Parent Directory' to instruct Google Search to treat this as a phrase[7].

Google Search results may also be limited by a date range, using the Advanced Search options. This is a convenient way of finding examples of older versions of a format, which might be sparse in results returned from an unqualified search.

Files that are downloaded from internet sources should be virus checked as a matter of course, and you should consult your IT department before downloading unknown files from the internet. You should also ensure that your storage or use of these files does not breach copyright laws.

---

[5] webtrends.com
[6] www.google.com/analytics/
[7] support.google.com/websearch/bin/answer.py?hl=en&answer=136861

Storing the downloaded files systematically and in a specific location will make it easier for you to work with them and to dispose of them when appropriate.

9 Signature research

Sections 5-7 provide an overview of the three different methods of signature identifications used by DROID and PRONOM. We will now look at the methods we use to conduct file format signature research. These processes can be applied by other researchers undertaking similar work. While our methods are not definitive, they provide a useful benchmark for work in this field of digital preservation.

The process we apply can be briefly summarised as follows:

- prioritisation
- researching file format specifications
- collecting file samples
- investigating and developing internal byte sequences
- testing the signature

9.1 Prioritisation

The PRONOM database contains information about more than 750 different file formats. Although this represents many years of research, it covers only a small fraction of the thousands of different file formats in existence. One online file extension identifier service currently contains 26,024 records, with 51,537 registered file type records and 16,344 Program/MIME type records[8].

As it is not feasible for us to research this number of file formats, we have developed a method to help us determine which formats should be investigated as a matter of priority, and which are of lesser importance.

[8] FILExt.com, online file extension search engine - filext.com/
Figures stated are accurate as of 1 May 2012

We use three sources of information to determine the priority with which file formats should be investigated. These are:

- analysis of The National Archives digital collections
- user requests and community feedback
- analysis of PRONOM web hits

9.1.1 Analysis of The National Archives' digital collections

The National Archives is fully committed to the role of PRONOM and DROID in supporting the wider digital preservation community. However, first and foremost we must identify and manage the file formats we encounter in our operational work of accessioning, storing and preserving the digital records of the UK government and related public record bodies.

To this end, we regularly analyse the file formats in our collections (both the digital archive collection and our current business records) and also investigate the file formats in use across government. This enables us to prioritise research into formats that are likely to enter our archival collections. This is the primary driver for our research into file format identification.

9.1.2 User requests and feedback

The National Archives currently deals with a limited set of digital formats, although this is expanding over time. Other archives, repositories and research libraries deal with formats that we have yet to encounter.

Requests and feedback from the PRONOM and DROID user community provide a secondary source of information which helps us set priorities. Responding to user feedback and requests enables us to increase the coverage, usability and relevance of PRONOM and DROID. It also improves our ability to deal with these file formats in the future, in the event that they enter our own collections.

9.1.3 Analysis of website hits

The direction of file format research is also shaped by an analysis of searches and page views on the PRONOM registry website. This indicates the range of formats of interest to

our users, and enables us to target formats which are receiving a high volume of interest. It also helps us identify trends, such as an interest in video, audio or CAD file types.

9.2 Researching file format specifications

Once file formats have been prioritised for research, the next step is to look for a consistent sequence or pattern of bytes which can be interpreted as an internal signature. The most obvious approach is to refer to a published file format specification, to which a digital object must conform if it is to be recognised as an example of that particular format. Some formats, such as Portable Network Graphics files, include a signature specifically designed to allow identification. In other formats an incidental signature can be derived from elements of the internal structure. In either case, the byte sequence and order will, to some extent, be determined and constrained by the format specification.

Open and published format specifications may be found on the internet through simple search engines queries. Other specifications may be ISO standards and these will be published on the ISO website[9].

However, any type of file format specification is useful for the purpose of signature research. This holds true even if the specification is generalised or non-specific, or does not detail a specific signature sequence, or even if it is not an official specification. Many software enthusiasts have conducted their own format research and made their insights and findings available via the web. While their intended goal may not have been the identification of an internal signature, their work can be of significant help to us.

9.3 Collecting format samples

Even when an official specification for a format exists, it is advisable to have sample files to research and investigate - in many cases a format specification does not adequately describe every potential variation in the internal structure of the files. File format researchers generally use their own collection of digital objects, since these tend to be the file formats that are most interested in identifying, understanding and preserving.

[9] International Standards Organisation. www.iso.org

Ideally, the digital objects used to research and develop a signature should be drawn from as wide a range of sources as possible. This increases the chances that the observed patterns in the internal structure of the objects indicate an actual signature - rather than arising from the environment in which they were created, stored or used. If only a small sample of files, or samples from a small number of sources, are available for research, it can be difficult to establish that common patterns in the file structures reflect a signature which is persistent across all examples of that format. Even where a large volume of digital objects is available for research, it is generally desirable to collect and examine examples which were created at different times and under different conditions. Section 8.3 gives examples of using the internet to obtain sample files for research.

9.4 Investigating and developing internal byte sequences

Once a sufficiently large sample of files in a particular format has been assembled, work can begin on investigating the internal structure of that format in an attempt to identify common patterns.

A binary file signature is modelled by describing its starting position within a bit stream, and its values. The starting position may be absolute or variable.

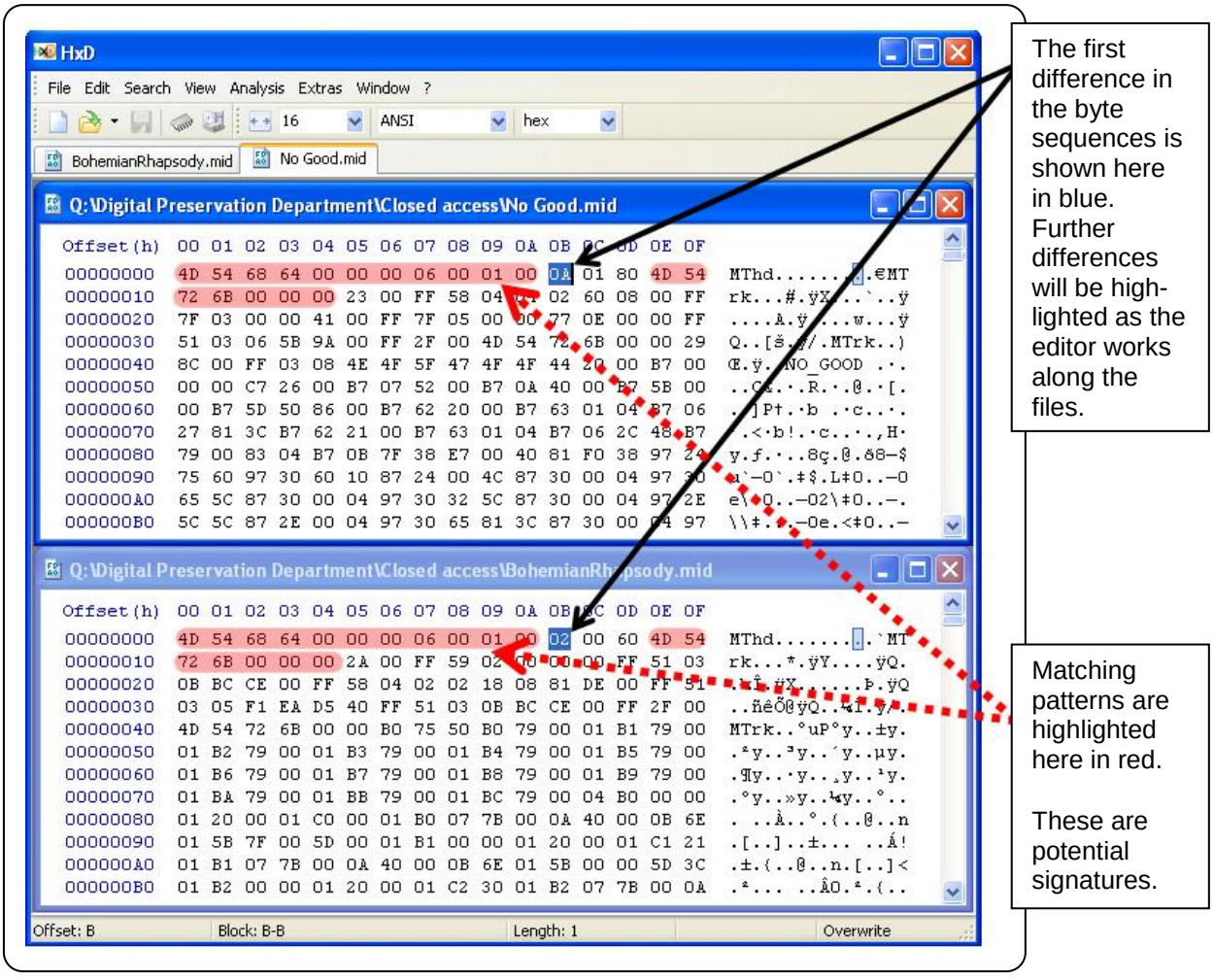
By absolute we mean that the sequence starts at a fixed position within the digital object. The fixed position can be at the start - ‘absolute Beginning of File’ (BOF) or end - ‘absolute End of File’ (EOF). Alternatively it may be at a fixed offset relative to the start or end of the file. A variable sequence, as the name suggests, can occur anywhere within the internal bitstream of the file.

For example, a signature sequence which always begins at the 36th byte of any file of that format would be described as an absolute (or fixed) offset, whereas a sequence which could start anywhere within the first 36 bytes would be a variable offset. While it is possible that an identifying sequence is completely variable in its location, a limit or range can generally be placed on variability.

The start and end sequences of a file often contain the key patterns from which internal signatures can be created, and these are the areas we look at first.

In the example in Figure 4, a hex editor has been used to compare the internal structures of two MIDI audio files. As can be seen, it is only at byte number 12 that the byte sequence begins to differ between the two files, this indicates that the first eleven bytes may form part of a signature. We can also see that bytes 15-21 are identical in the two files, so again this is a promising sequence on which to focus research.

Figure 4 A hex editor being used to compare the internal structures of two MIDI files



Further MIDI files can now be investigated, to see if the pattern identified above holds true for other examples of the format. Reference to specifications can identify further areas of the byte sequence to target, which helps make the overall task of signature development more manageable. In the above example of the MIDI file format, the eventual signature was discovered to be:

Offset: Absolute Beginning of File (Absolute BOF)

Sequence: 4D546864000000600[00:02]{4}4D54726B

This sequence was identified using 10 MIDI files downloaded from various internet sources, and with reference to the MIDI file format specification 1.1, released by the International MIDI Association. The availability of a specification gives a degree of confidence in the proposed signature, which in turn meant that a relatively small number of sample files were needed to construct and test the signature.

9.5 Testing

A signature is only as good as the testing that it undergoes prior to release into the PRONOM registry. The more rigorous the testing, the more likely that the signature will accurately identify a particular file format.

In the past, The National Archives has focused on internal testing of our signatures against file formats which either already existed on our network, or which had been specifically obtained to support testing.

Recently, this approach has been extended to add an element of external testing. We will now provide test signatures to interested organisations, which they can test against their own files, feeding back the results to The National Archives. The benefits of this new approach are two-fold:

- it increases the size of the sample being tested
- it improves the diversity of the test files, by bringing in files that have been created and held in different technical environments

9.5.1 Internal signature testing

Once a potential signature is identified, it is added to the test PRONOM database. This is a copy of PRONOM, into which test data can be added and from which test signatures can be generated without affecting the live system. There are two elements to the testing of internal signatures, each with a specific purpose.

First, the signature is tested against sample files in its target format. This validates the accuracy of the signature. If some files remain unidentified at this stage, this process allows us to identify any adjustments required to improve the success rate. It is not unusual for a test signature to initially identify only 90-95% of the files in the test sample.

During this process of refinement, several test signatures may be developed before all the digital objects in the test sample are successfully identified. In some cases, a single signature will not successfully identify the full range of digital objects of a particular format, due to variability in the content and placement of the identifying byte sequences. In these cases, variations of the general signature can be created and added to PRONOM to accommodate known variations in the file format.

Second, any new signature is tested against an internal test suite of a wide range of known file formats. This tests whether the new signature conflicts with any existing identifications. By testing against known files, false positive identifications will be immediately recognised as a potential error, and can be investigated and resolved promptly.

9.5.2 External signature testing

As part of the testing process, The National Archives makes DROID test signatures available for external testing. Testers can use the test signatures on their own digital collections, providing an extra layer of challenge and validation before the signature is formally released.

Our eventual aim is to create a File Format Validation Test Suite, populated with examples of the various file formats recorded in PRONOM. This too could be made available to interested parties to support their own signature research and testing, and would add a further layer of verification and confidence to the signatures developed for PRONOM and DROID. Interested researchers in the digital preservation community could contribute additional files to the test suite so that, over time, it would be populated with a wide and varied corpus of file formats.

An obvious hurdle to overcome before this validation suite could become a reality is the issue of ownership and intellectual property rights on the digital objects themselves. This would need to be satisfactorily resolved before the validation suite could become a reality.

10 Optimisation

The guiding principle of our file format research and the PRONOM service is to provide a means of identifying any given file format uniquely. To this end, PRONOM will attempt to generate an internal signature for any file format, irrespective of the length, complexity or content of that signature. A sequence within a signature may be only a few bytes in length, or may need to be 50 or more bytes in order to provide accurate identification. PRONOM and DROID are designed to handle these differences in scale and complexity.

However, large or complex signatures have a greater impact on the performance of DROID (because DROID will have to read and compare a greater number of bytes). With this in mind, we also work to optimise the file signatures in PRONOM.

Any signature in PRONOM may be enhanced or replaced over time, as we improve our understanding of the file format concerned, or in response to feedback received from users of the service. A signature might be rewritten to improve either its accuracy or its efficiency.

While accuracy remains the focus of our file format research, if it is possible to improve the performance of DROID by optimising internal signatures (for example by replacing variable offsets with fixed offsets, or removing wildcard elements) we will make efforts to do this.

DROID and PRONOM are designed primarily for file format identification. Neither formally validates files against a format specification or makes assertions about conformance with any specification. For this reason, when creating file format signatures, we do not attempt to capture every technical detail of the format. This approach would go far beyond our intended aims and would reduce the performance and efficiency of the tools. Rather, for each format, the signature encodes only the minimum amount of information about the common byte sequences in our sample files that will allow us to confidently assert whether a file either is or isn’t a further example of the format. A strong signature must be concise enough to be handled efficiently by DROID but detailed enough to avoid clashes or false-positive identifications.

11 Next steps

The following resources from The National Archives will help you get started with your file format research:

- further information about PRONOM nationalarchives.gov.uk/pronom/default.aspx
- further information about DROID nationalarchives.gov.uk/information-management/our-services/dc-file-profiling-tool.htm
- Digital preservation technical paper1: Automatic format identification using PRONOM and DROID nationalarchives.gov.uk/aboutapps/fileformat/pdf/automatic_format_identification.pdf

You can contact us at pronom@nationalarchives.gsi.gov.uk

Please get in touch if you have any queries about Pronom or DROID, if you would like to get involved in testing new signatures or if you wish to develop new file format signatures and contribute these to the PRONOM service.