

Novice to Know-How Module Text

Course 3: Using DROID

Module 1: What is DROID?

1. Introduction,

In the previous course we looked through the range of issues around preserving the bitstreams of digital content, including what files and formats are, what risks they face, and basic workflows for preserving content. Work around these issues requires us to understand the digital content we want to preserve, and to capture information about it. But doing this our self by hand, file by file, would be incredibly laborious, in fact prohibitively so. Therefore, finding software that can help us generate the information we need is essential. Thankfully, there are a number of software tools available, and in this course, we're going to look at one of the easiest to access and use, DROID.

2. What is DROID?

DROID is a software tool developed by the Digital Preservation department of The UK National Archives. It performs automated batch identification of file formats, a process also known in digital preservation as characterization. DROID is free to download from The UK National Archives' website and is supported by a thorough user guide. There is an active community of DROID users and the tool has also been integrated into a number of larger repository systems.

3. A Little on Characterization,

On the last slide we mentioned that the analysis undertaken by DROID is often called characterization by digital preservation practitioners. Fundamentally, Characterization is about understanding the digital content we wish to preserve. This will then allow us to assess risks, plan our preservation processes, and take action as needed.

The information generated by the characterization process can answer questions such as:

- How many files are there?
- How big are the files?
- What file formats are they?
- When were they created?
- When were they last edited?
- Is there any dynamic or interactive content?
- Does the digital content contain personal information?
- And, are any of the files encrypted?

This information represents a lot of the basic metadata that we would want to save for our digital content.

Characterization tools are also very useful when working at scale and we want automation of processes where possible.

4. Other Characterization Tools,

Before we delve a bit more into DROID, it is worth noting that there are many other characterization tools available, each with slightly different functionality. Particularly in relation to the richness of the metadata they generate. DROID is a great place to start, but as you progress further down the digital preservation road, you may wish to investigate other options. Other tools include:

- Apache Tika
- EXIF
- FITS
- JHOVE

For a comprehensive list of characterization tools, you can visit the COPTR tools registry. A link is provided in the resources for this course.

5. How does DROID work?

DROID uses three different methods for identifying file formats, these are an analysis of the file's:

- Extension
- Signature, and
- Container

The metadata provided by DROID for each file will say which method was used to identify the individual files.

If DROID identifies the file by its extension this means that the format was identified purely on the basis of its file extension. Such an identification may not be reliable, as file extensions can easily be changed, and many file formats and versions of file formats have the same extension.

A signature identification means that a format was identified by finding a specific pattern in the byte sequence, usually in the header of the file. The sequence is unique to a particular file format and version. This method is much more reliable than identification by extension only.

A container identification means that a format was identified by finding embedded files, often with signatures of their own, inside the main file. For example, OpenDocument word processing files are actually zip files containing xml files, images or other resources used in

the document. A container identification would identify the main file as an OpenDocument file, not a zip file. This method is very reliable, as not only does the broad type of container have to be identified (for example zip), but the zip file must then be opened, and files inside scanned for further identifications to be made.

6. DROID and PRONOM,

To allow DROID to make its identifications it needs access to information about file formats and their characteristics to use for comparison. For this, DROID uses PRONOM, a technical registry also developed and maintained by the UK National Archives.

PRONOM is a large database of information on file formats and the software products that support them. A PRONOM record can include information such as the version of the format, what compression and encoding standards are used, if a specification can be accessed and where it can be found, and who owns or manages the specification. This is all information that can be very useful for digital preservation.

When generating metadata about files it has analysed, DROID not only lists the method of identification used, but also includes a unique identifier for the format in PRONOM. This allows a link to be made between the metadata and the corresponding format record in PRONOM.

PRONOM records for commonly used file formats generally include more detail than rare or niche formats. The UK National Archives does, however, welcome contributions from the community to help enrich the data held in PRONOM.

7. Why Use DROID?

DROID is a great tool to use when starting out in digital preservation for several reasons:

- First, it is free to download, easy to access, install and set-up, and there is an excellent user guide.
- Second, it has a straightforward, simple user interface. Many other characterization tools do not have a user interface at all, working only from command line instructions.
- It also can identify one of the biggest ranges of file formats amongst similar tools, thanks to the richness of the data in PRONOM.
- Next, we can have confidence in its continued support by The UK National Archives, as they use it in their own digital preservation processes.
- Finally, it can produce, with a high-level of reliability, the basic metadata we need for understanding the digital content we have. It will then allow us to export this metadata in different formats including comma separated value files that can be easily used in Excel or uploaded to a database. It can also produce a variety of summary reports.

8. What information can DROID provide?

A DROID analysis can produce up to eighteen pieces of metadata for each file. This includes:

- The file name, size, last modified date, and file path: this records what the file is called, how big it is, when it was last edited and where the file is stored on the system
- The identification method used and the status of the DROID analysis: letting us know how DROID identified the file and whether the analysis was successful or if there was an error or problem accessing the file
- The file type, extension, format, version and PRONOM identifier: this covers some key characterization information, so we know what type of file it is and what format (including version) it is. An identifier that links to the relevant format description in PRONOM is also included. Allowing access to more information.
- DROID can also generate checksums according to three different standards to facilitate integrity checking.

9. What else is covered in this course?

In the remainder of this course we'll look at how to use DROID and its key functionality in detail. The modules will cover:

- Downloading and installing DROID
- Opening and setting-up DROID
- Creating a DROID Profile
- Running DROID and exporting data
- And, creating DROID reports

Each module has a set-by step demonstration of the relevant functionality, and all modules except 'Downloading and Installing DROID' will also take you through a task to "try out" the functionality.

There is also a link to the DROID User Guide in the resources for this course.