The National Archives

# DROID: How to Use It and How to Interpret Your Results

This guidance is produced by the Digital Continuity Project and is available from www.nationalarchives.gov.uk/dc-guidance

Any enquiries regarding the content of this document should be sent to digitalcontinuity@nationalarchives.gsi.gov.uk

**CONTENTS**

# 1. Introduction

**Digital continuity is the ability to use your information in the way you need, for as long as you need.**

If you do not actively work to ensure digital continuity, your information can easily become unusable. Digital continuity can be put at risk by changes in your organisation, management processes or technology. You need to manage your information carefully over time and through changes to maintain the usability you need.

Managing digital continuity protects the information you need to do business. This enables you to operate accountably, legally, effectively and efficiently. It helps you to protect your reputation, make informed decisions, avoid and reduce costs, and deliver better public services. If you lose information because you haven't managed your digital continuity properly, the consequences can be as serious as those of any other information loss.

## 1.1 What is the purpose of this guidance?

This guidance forms part of a suite of guidance[1] that The National Archives has delivered as part of a digital continuity service for government, in consultation with central government departments.

This piece of guidance provides you with practical information and support to help you complete the four-stage process of managing digital continuity.[2]

You should already know about identifying your information assets, relating them to business needs and mapping your technical dependencies.[3] This guidance will help you to use the DROID file profiling tool to further examine your information, and will enable you to:

- understand the drivers for file profiling
- understand how to run DROID
- interpret your results, avoiding common pitfalls.

[1] For more information and guidance, visit nationalarchives.gov.uk/digitalcontinuity
[2] See *Managing Digital Continuity* nationalarchives.gov.uk/information-management/our-services/dc-guidance.htm
[3] See *Identifying Information Assets and Business Requirements* nationalarchives.gov.uk/documents/information-management/identify-information-assets.pdf and *Mapping the Technical Dependencies of Information Assets* nationalarchives.gov.uk/documents/information-management/mapping-technical-dependencies.pdf

## 1.2 What is DROID?

DROID (**D**igital **R**ecord **O**bject **ID**entification) is a file profiling tool developed by the National Archives. It is in widespread use across the world, in cultural institutions, local and central government departments and other public bodies, and has been embedded into some commercial information management products.

DROID scans files, collecting information about them into a profile which can later be explored, filtered, exported and reported on. Millions of files can be profiled, and many different profiles can be reported on at the same time. It will also look inside archival files (such as 'zip' files), and examine the files inside them too.

One of the most important functions DROID performs is to identify what format a file is written in, even if the file name extension is wrong or missing. Where possible, identifications are made beyond the broad type, down to the version level. For example, it can tell you that a document is written in a very old version (e.g. Word 6.0), not just that it is a Word document.

DROID can currently identify over 250 file formats, and this number is growing all the time. Updated format signatures are automatically downloadable from the National Archives' PRONOM[4] service.

This guidance is based on DROID 6, while some of it may also be applicable to earlier versions of DROID (or indeed later ones), certain specific features may not be available, or may behave differently.

DROID is available for free download from the source-forge website: http://droid.sourceforge.net. If you have problems downloading this, please contact us at digitalcontinuity@nationalarchives.gsi.gov.uk.

[4] PRONOM is a registry of technical information about file formats and software, available at: nationalarchives.gov.uk/pronom

## 1.3 Who is this guidance for?

This guidance is primarily aimed at information managers wishing to manage digital information stored in files. It may also be useful for archivists, IT or digital preservation professionals who want to gain information about their file usage. Note that some of this guidance goes into technical detail, so information managers may want to seek assistance from IT, for instance with help in installing DROID (see Appendix). IT professionals may also find it helpful to refer to this guide, but will also have “Help” files on hand within DROID.

# 2. File profiling

To ensure the digital continuity of your information, you should understand the formats your information is encoded in and how files are being used in your organisation.[5] Profiling your files collects all of this information into a single database, and allows you to report against all of your files and formats.

## 2.1 Reasons for profiling

There are many different reasons why you may wish to profile your files and folders. From an information management and digital continuity perspective, you may want to profile your files to help you:

- understand your information and identify risks and issues (section 2.1.1)
- assure your technical environment supports your information (section 2.1.2)
- audit information management policy compliance (section 2.1.3)
- reduce data volumes (section 2.1.4)
- detect duplicates (section 2.1.5).

### 2.1.1 Understanding your information

Understanding what sort of files you have, the volume of data held in them and areas of high and low usage can help you make informed decisions and set policies about managing the information your organisation creates and uses. Reporting on profiles enables you to:

- understand data volumes by year or month, and how they are changing
- understand what different formats information is stored in, and how much space they take up
- locate areas of high and low usage by examining last modified dates
- understand what sort of information your organisation is creating. Examining the mime-types (see section 2.2.9) of files is useful here, as they provide broad classifications of content.

You may be able to understand which departments in your organisation are creating most data, according to where information is stored. This can help you see who needs your resources most – or who is overusing them.

[5] See more information in our guidance *Evaluating Your File Formats* nationalarchives.gov.uk/documents/information-management/evaluating-file-formats.pdf

You can also use this knowledge of your information to assess the risks to your information – for instance, if information is stored in legacy formats, you may need to convert them to avoid the risk of technical obsolescence.[6]

### 2.1.2 Assuring your technical environment supports your information

You may find that information is stored in a variety of formats, some of which may not be fully supported by your current software, or which are not your preferred formats. Even if you have a format policy to control which formats users save information in, files can be submitted from external organisations or members of the public. By filtering your profiles, you can:

- **look for older versions of file formats you support**. These may be possible migration candidates[7]
- **look for the file formats for which you do not have official software support**. These may be legacy files, or files submitted externally. They may also be candidates for migration
- **pay particular attention to files which DROID cannot identify at all**. In some cases, this may simply be because DROID has no signature for them, or which are unusual in some way. However, bespoke file formats, or very niche file formats will probably not be identified, and may be at particular risk of obsolescence
- **reduce software licensing costs**. For example, if there are no recently modified files in a given format, it is possible that no-one in the organisation is using this software anymore. You may be able to replace this software with viewers (rather than full functionality software), or migrate these files to another format.

### 2.1.3 Auditing information management policy compliance

You may have policies governing what sorts of information can be recorded in your systems. For example, you may not allow users to store music collections on your servers, or you may not allow users to save emails into email archive files (e.g. PST files), but instead require that they are stored in a records management system. By examining the file formats of files in use across your systems, you can detect whether these policies are being adhered to.

[6] See *File Format Conversion* nationalarchives.gov.uk/documents/information-management/format-conversion.pdf

[7] See more on migration in *File Format Conversion* nationalarchives.gov.uk/documents/information-management/format-conversion.pdf

If you generate “content hashes” with DROID (see section 2.2.11), you can also use these to link to forensic hash databases. These can allow you to detect whether the files in your systems are common, well known files (such as Windows system files). Knowing which files are well known outside your organisation can support information policy and decision making. For example, you may discover that a lot of storage space is being taken up with multiple copies of files which are easily replaced by installation CDs. Files which are not well known probably contain unique content, and would be hard to replace if deleted. Content hashes can also allow you to detect whether the files in your systems contain known illegal content, and in some cases, malware such as viruses.

There is a variety of content hash databases available. Two widely used ones are:

- National Software Reference Library: www.nsrl.nist.gov/
- Hashkeeper: www.justice.gov/ndic/domex/hashkeeper.htm

Note that DROID does not link your files to these hash databases. It merely generates a compatible hash for each of your files, which you can then use to link with them. You will require additional technical assistance to perform these links.

### 2.1.4 Reducing data volumes

By examining your file usage, you may be able to free up expensive storage. You can use DROID to:

- find files and file formats which occupy most space
- find the oldest files
- detect duplicates (see section 2.1.5)

If you identify areas of files, or formats which are taking up too much space, there are several actions you can take, including:

- deleting files which are not required. You must assess the business use of files and whether they are candidates for permanent preservation. You must not delete files simply because they are old or large. However, knowing the age/size of files can help you determine which files should be assessed first
- compressing files which are taking up a lot of space. For example, plain text files (e.g. csv data sets) can be compressed very efficiently into zip files with no loss of data
- converting files to a format which takes up less space

- moving the files to safer storage, e.g. a file archive. See the Digital Continuity Framework Lot 7 for digital archiving solutions.[8]

### 2.1.5 Duplicate detection

It is very common to find that files are duplicated in different areas of your filing systems. This can happen because many users save the same files from email attachments, or they take a backup copy of files while they are working on them, but don't end up changing most of them.

There are several ways in which duplicate files can be detected. The simplest method is to search for folder names containing words like, "backup", "temp" and "old", as users frequently name folders or files with these words if they intend them to be temporary copies. Another, more time consuming method, is by examining the names of files and folders. If there are areas with very similar (or identical names), then you may have duplicate information within them. However, both of these methods can only give you an indication that there may be duplication and a high degree of manual review will still be required to assure yourself that the file contents are really duplicated.

Another method of duplicate detection is to use content hashes (see section 2.1.11). If two files have the same hash value, then they are overwhelmingly likely to have identical content. DROID can generate content hashes for your files. If you export your profiles to a CSV file and import them into software like Excel or Access (see section 4.4), you can query for files which have the same hash.

If you do find duplicate files, you must decide how to deal with them. Clearly you will need to keep at least one of them, but you will have to decide which, if any, can be safely removed. There are risks to digital continuity in deleting files, so you should take into account several considerations before deleting duplicates:

- not all users may have access to all copies of a duplicate file
- the duplicate files may have different metadata. For example, the location of a file in a filing structure can provide essential context. Or there may be important audit metadata showing that different people opened the different duplicates
- a duplicate file may provide context to the surrounding data it was stored next to. The loss of this context could render the meaning of the surrounding files unclear or unintelligible

[8] See digital archiving solutions on our Digital Continuity Framework www.buyingsolutions.co.uk

- large areas of duplication may be for test or development environments, and therefore will be necessary for systems development.

You can mitigate some of the risks related to loss of context by leaving shortcuts (or symbolic links in a UNIX file system) to the “master file” when you delete a duplicate.

## 2.2 What information does DROID gather?

DROID collects a variety of information about your files and folders, including:

- type
- file name
- file name extension
- file name extension mismatch warning
- location
- file size
- last modified date
- number of format identifications
- file formats
- identification method
- content hash
- job status.

The following sections describe the meaning of each of the above items of information and where any issues can arise in understanding and interpreting it.

### 2.2.1 Type

DROID categorises the files and folders it profiles as being one of three types:

- file
- folder
- archival file.

Files have format identifications, but do not have other files or folders inside them. Folders do not have any format identifications or sizes, but can contain other folders, files and archival files inside them. Archival files are like folders, in that they can contain other folders, files and archival files inside them, but they are also files, so they have format identifications and a file size. Presently, DROID can look inside zip, tar and gzip archival files.

### 2.2.2 File name

The name of a file, folder or archival file is its name, independent of its location on a disk or inside an archival file. It includes any file name extension as part of its name. DROID treats all filenames as case-sensitive. For example, “MYDOCUMENT.DOC” and “mydocument.doc” are regarded as different file names.

### 2.2.3 File name extension

File extensions are a convention to indicate the broad type of a file (or archival file) by appending a short string to a file name, separated by a full stop. DROID converts all file extensions it extracts to lower case, to facilitate sorting, reporting and filtering.

Note that DROID will only extract an extension for files, not folders, and that there must be at least one character before the full stop – there must be a name before the extension.

### 2.2.4 File name extension mismatch warning

Sometimes file extensions are incorrect for the type of the file, or are missing where there should be one. If DROID detects that the file extension for a file name does not match the formats it has identified, it will issue a file extension mismatch warning. For example, if a file called “myfile.doc” is identified as a spreadsheet, then a file extension mismatch warning will be issued.

### 2.2.5 Location

DROID records the location of every file and folder it profiles. It records location in two ways, using a Uniform Resource Indicator (URI), and a file path where one exists. Like file names and extensions, DROID treats file paths and URIs as case sensitive.

There are two ways of recording location because not all files and folders have a file path, although this is the usual method of identifying location in a file system. Any file, folder or archival file which is inside another archival file does not have a defined file path, as it is inside the archival file, not directly in the file system.

For example, if we have:

1. a folder called “Folder” on the “C:\” drive of a Windows computer
2. a file called “Document.doc” inside “Folder”,
3. an archival file called “Archive.zip” inside “Folder”

4. a spreadsheet called “Spreadsheet.xls” inside “Archive.zip”
5. a folder called “Another folder” inside “Archive.zip”
6. a picture called “Large picture.jpg” inside “Another folder”

Then we have the following file paths and URIs:

| | File path | Uniform Resource Indicator (URI) |
|---|---|---|
| **1** | C:\Folder | file:/C:/Folder/ |
| **2** | C:\Folder\Document.doc | file:/C:/Folder/Document.doc |
| **3** | C:\Folder\Archive.zip | file:/C:/Folder/Archive.zip |
| **4** | | zip:/file:/C:/Folder/Archive.zip!Spreadsheet.xls |
| **5** | | zip:/file:/C:/Folder/Archive.zip!Another%20folder/ |
| **6** | | zip:/file:/C:/Folder/Archive.zip!Another%20folder/Large%20picture.jpg |

Only files, archival files or folders which are directly accessible in the file system have a file path. Those files and folders which are inside the zip file do not have a file path, but do have a URI, which tells you that they are inside the zip file, where they can be found in it, and where the zip file they are inside is to be found.

The prefixes of a URI tell you what sort of resource is being described by the URI, and the exclamation marks indicate where one type of resource is contained by another. For example, for “Spreadsheet.xls”, we can see that there is a file, C:/Folder/Archive.zip, with the prefix file:/. The exclamation mark (!) tells us that the spreadsheet is contained by the Archive.zip file, and the first prefix zip:/ tells us the type of the containment is a zip file. Note that spaces in URIs are encoded by “%20”, and folder separators are always forward slashes.

### 2.2.6 File size

The size of a file or archival file is recorded as the number of bytes used by the file. Files can have a size of zero (no content, just a record in the file system). Folders do not have a size.

The size of an archival file is the size of the archival file itself, not the sum of the sizes of its contents. For example, zip files compress their contents, so the sum of the sizes of the files inside a zip file will be bigger than the size of the archival file itself.

### 2.2.7 Last modified date-time

Most files, folders and archival files record the date and time on which they were last modified. This is not the same as the date a file was originally created, or the date on which a file was last read. Unfortunately, due to limitations in Java 6, DROID can only acquire the last modified date, even though the other dates may be present on the file system.
It is possible that not every file, folder or archival file will have a last modified date. For example, in some cases, resources inside archival files may not record this date.

It is important to note that last-modified dates can be changed when files are copied from one server to another, so this date may not reflect the last date a user actively modified the content of a file. Also, the content of a file (the data within it) may actually be older than the file itself – if a file was copied, or simply typed up manually from an older piece of content.

Some files may have noticeably inaccurate dates, e.g. 1 Jan 1970. In this case, the files will be newer than indicated. This error will likely be caused by the battery failing on the internal clock of the computer from which the document was uploaded.

### 2.2.8 Number of format identifications

DROID attempts to identify the format of files, including archival files, but not folders. The number of identifications DROID records for a file can vary. It can have

- zero, if DROID can't identify a format at all
- one, if it is unambiguous
- more than one, if DROID can't unambiguously decide what format it is in.

The latter situation can happen for three reasons.

1. A format is identified purely on the basis of its file extension, so multiple versions of a file format may match the same extension.
2. A format has several versions which are very similar and hard for DROID to distinguish between, so DROID will simply report all the possible versions.
3. A file may contain patterns, purely by chance, which appear in more than one file format.

### 2.2.9 File formats

When DROID identifies a file format, it records four pieces of information:

- format name
- format version
- PRONOM Unique Identifier (PUID)
- mime-type.

The format name is simply a human-readable name given to a file format or family of file formats, for example, “Microsoft Word”. The format version is the version of the format, for example “97-2003”. The PUID is a globally unique, persistent identifier for a file format and version, assigned by the National Archives through its PRONOM file format registry.[9] For example, the PUID for the “Microsoft Word 97-2003” file format is “fmt/40”.

PUIDs are guaranteed never to change, although new PUIDs may be defined. Clicking on a PUID in DROID will take you to the relevant page for that file format on the National Archives PRONOM website. The website will also help you with some file format names that you may be unfamiliar with. For example, you may see files identified as ‘OLE2 Compound Document Format’ (PUID fmt/111). This is a file format created by Microsoft used to contain other files and folders. It is often used in older Microsoft Office formats, but it is also used by other applications, some not written by Microsoft.

Finally, the mime-type is another scheme for identifying broad types of files in use on the internet. They are assigned by a body called the Internet Assigned Numbers Authority. Mime-types are quite broad classifications, so many different file formats will have the same mime-type. For example, the mime-type for “fmt/40” is “application/msword” – which is shared by all other Microsoft word formats.

### 2.2.10 Identification method

DROID has three different methods of identifying file formats:

- extension
- signature
- container.

An “extension” identification means that a format was identified purely on the basis of its file extension. Such an identification may not be reliable, as files can be named in any way, and extensions do not identify formats down to the version level, so such identifications can be quite broad, and may result in multiple identifications.

[9] PRONOM nationalarchives.gov.uk/pronom/

A “signature” identification means that a format was identified by finding signature patterns inside the file which are known to occur in particular file formats and versions. This method is quite reliable, as it is fairly unlikely that by chance a file will happen to have a pattern belonging to a different file format than its own.
A “container” identification means that a format was identified by finding embedded files (possibly 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 (e.g. zip), but the zip file must then be opened, and files inside scanned for further identifications to be made.

### 2.2.11 Content hash

DROID can optionally generate a content hash of the contents of each file and archival file, using the industry standard “MD5” algorithm. A content hash is a long number hat can be used to identify the content of the file. It is extremely unlikely that two different files will have the same content hash. The odds of two different files having the same hash by accident are a bit less than one in a billion billion.

Content hashes can be used to detect files with duplicate content (see section 2.1.5 on duplicate detection), or can be linked to forensic hash databases to find or exclude files which are widely used (and therefore not unique to your organisation) or which contain illegal content (see section 2.1.3 on auditing information management policy compliance).

Content hashing is turned off by default, as producing a hash requires reading the entire file, which will slow down DROID significantly.

### 2.2.12 Status

As DROID profiles your files and folders, it records whether the profiling was successful or not. There are four different statuses which a file or folder can have (see below):

| | |
|---|---|
| **Done** | Success, DROID could read the file or folder, and no errors occurred while trying to do so. It does not mean that DROID could identify the file format. |
| **Not found** | The resource was moved or deleted before it could be profiled. |
| **Access denied** | The operating system refused read access to DROID. You may want to change the permissions on those resources to enable profiling. |
| **Error** | An error occurred while trying to read the file. Anything that prevents DROID from profiling a file or folder (other than being not found or access denied) will result in an error status. You may be able to determine the cause of the error by examining DROID's error logs, although these are quite technical in nature. |

# 3. How to profile your files

This section will explain the main steps you should take to profile files using DROID, and the various options you need to consider. It will not explain how to use the software itself in great detail. For this level of detail (including screenshots), please refer to the help included in the DROID software itself, or for information on installing DROID, go to our Appendix.

## 3.1 Running DROID

To run DROID, go to the folder in which it is installed. If you are using Microsoft Windows, then double-click on the file called “droid.bat”. If you are running Apple Mac, or Linux, double-click on the file called “droid.sh”.

These files can also be run directly from the command-line, instead of double-clicking on them through your Graphical User Interface.

## 3.2 Creating a profile

DROID automatically creates a new profile for you (a blank tab) when it is first started. You can create more than one profile at a time, which will open in separate tabs.

You have several options you need to consider before you create a profile to use. You should decide whether to:

- look inside archival files (zip, gzip and tar files)
- generate a content hash for each file it processes.

If you look inside archival files, you will get a much better picture of what files and formats you have in your organisation, although the trade-off is that it will naturally take longer to profile, as you will be examining more files.

If you generate a content hash, then you can use this to detect duplicates, or link to forensic hash databases. However, generating a hash slows down DROID considerably, as the entire file must be read to do this. If you do not need the content hash, it is recommended that you turn off this option.

Note that if you need to change the profiling options, you must do this before you create a profile to use, as newly created profiles take the default options and cannot be changed after they are created. Profile options can be changed in the Preferences window, under the Tools menu.

## 3.3 Adding files and folders

Once you have created a new profile with the correct options set, you should tell DROID which files and folders you want it to profile. You can add individual files and folders to DROID. If you tick the “Include subfolders” box, and you select folders, DROID will not only process the files inside that folder, but also all the sub-folders inside it (and their sub-folders). Note that you can’t see the files and subfolders inside a folder you add until you actually begin profiling those folders.

If you accidentally add a file or folder you don’t want to profile, you can simply remove it. Note, however, that you can’t add or remove files or folders after a profile has been started – only while you are specifying what you want to profile.

## 3.4 Starting a profile

Once you have added files or folders to your profile, you can start profiling, by clicking the start button. Once you have started profiling, you can no longer add or remove files or folders from the profile.

You can see results appear dynamically in your profile, as DROID runs across your files and folders. As results appear, the icons for the files and folders move from being grey to colour, and the information it has found can be viewed while profiling is running. You cannot save, filter, report or export profiles which are running. If you had defined a filter on a profile you have re-started, it will be disabled when you start the profile.

At the bottom of your screen is a progress bar, showing the files and folders DROID is profiling. Initially, this progress bar may show very little, as DROID attempts to estimate how many files and folders it has to process. When it has produced this estimate, you will see the progress bar moving across (slowly, if you are processing millions of files), and the file name being processed will appear in the bar.

Please note that if you are profiling inside archival files (e.g. zip, gzip or tar), then the progress bar will estimate too low a number of files. This is because DROID does not know how many files are inside other archival files until it actually profiles them. Therefore, in these cases the progress bar will appear to stop, or may reach completion before the profile has actually finished. However, even in these cases (unless you have a very high proportion of files inside archival files), the progress bar should give you a rough indication of how long you can expect the profiling to take.

## 3.5 Pausing and resuming

If you need to, you can pause a running profile. Note that it may take a little time for the pause to take effect, as DROID queues up work which runs in parallel, and it must wait for all these tasks to complete before it can pause.

When a profile is paused, you can save, filter, export and report the results you have so far. This can be a good strategy to ensure you don't lose work, when profiling a very large number of files. When DROID is paused you can save the results so far, then resume profiling. If you have saved your profile so far, you can even shut down DROID and the machine it is running on, open the paused profile at a later date, and resume profiling.

In some extreme cases, you may be unable to restart DROID profiling once it's paused. This can happen if someone deletes or radically renames the areas it was profiling while it is paused. DROID will try to step back down your folders, to restart profiling where it can, but clearly if the structure it was in the middle of profiling no longer exists, then it won't be able to restart. In practice, this situation has never been seen, but it has been tested.

If a profile is paused, you can resume it simply by pressing the start button again.

## 3.6 Throttling

Occasionally, you may wish to keep DROID running but make it run slower, to take the load off the computer on which it is running, or the network or file servers that deliver the files you are profiling. In these cases, you can "throttle" DROID back, by moving the throttle slider control at the bottom of the main window. This control tells DROID to add a delay between processing each file (specified in thousandths of a second).

## 3.7 Recovery

If the entire area DROID is trying to profile becomes unavailable, DROID will attempt to pause the profile, allowing you to save the work done so far, and to restart it if the area becomes available again.

If you come back to a profile you left running and discover it is now paused, this is probably because the area you were profiling had become unavailable. In general, if you are running DROID for extended periods of time (days or weeks), you should check it occasionally, to see that it is still running. If not, save your work so far, and resume.

## 3.8 Opening and saving your profiles

You can save profiles to a file on your disk, and open them again for later analysis, or to resume a paused profile. These files have a ".droid" file extension. If you close down DROID and have not saved your profiles, DROID will prompt you to save them.

You can open as many profiles as you like in DROID. Each profile will open in a separate tab.

Be aware that if you are profiling sensitive information using DROID, then these files will also contain the names, locations, sizes and all the other information DROID gathers about them. You should ensure that you save these files in locations which are subject to the same level of security control as the information being profiled, and which are only accessible to an appropriate set of authorised users.

# 4. Exploring your results

Once you have profiled your files, DROID offers several different and complementary ways to explore these results, including:

- on-screen exploration
- filtering
- reporting
- exporting.

## 4.1 On-screen exploration

The Graphical User Interface of DROID offers a way to navigate and explore your results on screen, in the same way that file managers on Windows, Mac or Linux offer.

### 4.1.1 Viewing files, folders and archival files

Your profiled files are displayed as files and folders, nested within one another. To open a folder or archival file, click the plus sign beside it. To close it again, click the minus sign.

All the information collected by DROID is visible in this view. Note that file extension mismatch warnings (see section 2.2.4) are displayed in the file extension column as a yellow warning triangle icon. You can sort the information on the screen by clicking on the column headers.

### 4.1.2 Format identification information

If a file has no identifications, then the file will have a grey dot in the identification count column. If a file has a single identification, then the file will have a green dot in the identification count column, and format information will appear in the format columns.

If a file has multiple identifications, then the format columns will be blank, and a number in brackets will appear in the identification count column, indicating the number of different identifications made. Clicking on this number will bring up a small window containing all the identifications made.

### 4.1.3 Going to the files on your computer

If you want to look at the files themselves on your computer, DROID provides a way to open the nearest folder containing the selected file or files. Select the “Edit/Open Containing Folder...” menu or right-click and select “Open Containing Folder”. Your normal file manager will open at the folder which contains the item. If you select a file inside an archival file, then you will be taken to the folder containing the archival file.

## 4.2 Filtering your results

If you have profiled a lot of files, it can be time consuming manually looking for files which are of interest to you. If you apply a filter to a profile, then all the files and folders which don’t meet your criteria are filtered out, leaving only the ones of interest to you.

### 4.2.1 Defining a filter

Filters consist of one or more filter criteria (rules for filtering). Each filter criterion consists of three things:

| | |
|---|---|
| **Field** | The type of information to filter on |
| **Operator** | What sort of comparison to make |
| **Values** | The value(s) to compare against |

For example, you could define a filter like this:

| | | |
|---|---|---|
| File size | greater than | 1000000 |

If you also only wanted to look at files with the extension “doc”, you could add another filter criterion, giving this filter:

| | | |
|---|---|---|
| File size | greater than | 1000000 |
| File extension | equals | doc |

You can add as many filter criteria as you like to narrow down on the files or folders you are interested in. In the above example, you are looking for files which meet *all* of the criteria: files of a size bigger than 100000bytes *and* which have a file extension equal to “doc”. You

can also look for files which meet *any* of the criteria. For example, you may define a filter like this:

| | | |
|---|---|---|
| File extension | equals | doc |
| File extension | equals | DOC |
| File extension | equals | xls |
| File extension | equals | XLS |

This filter would find files which had any of the file extensions. Whether a filter looks for all or any of the criteria is set by simply switching the any/all option on the filter editing window.

### 4.2.2 Managing filters

Each profile has its own filter, independent of other profiles, allowing you to define different filters for different profiles. You can save and load filters you use frequently, so you don't have to keep manually creating them, and you can also copy a filter to all open profiles, if you want them to have the same filter on each. Filters can also be turned on or off quickly using a button in the graphical user interface.

### 4.2.3 Filters apply to everything

If you define and enable a filter, it applies to what you see on-screen, any reports you generate (see section 4.3), and any exports you perform (see section 4.4). This allows you to produce reports and exports over only the items of interest. If you want reports and exports across the entire profile, simply temporarily disable any filter you have defined for that profile.

However, there is one important difference when filtering items on-screen. Any folders or archival files which are needed to let you navigate to the unfiltered items are also displayed (or you could never get to the unfiltered items). These items are displayed with greyed-out icons, and without any information other than their name, to indicate that they do not meet your filter conditions, but are needed for navigational purposes. They will not appear in any reports or exports you perform.

## 4.3 Reporting on your results

Exploring and filtering your profiles allows you to see the kinds of files you have. However, if you want broad statistics over all of those files, then DROID offers a variety of reports to let

you quickly see the bigger picture. If you have enabled a filter, then the report will be on the filtered items only, letting you get different statistics for different sets of files and folders.

### 4.3.1 Report statistics

Reports offer the following statistics (where possible):

- count of items
- total size of items
- minimum size of items
- maximum size of items
- average size of items.

Note that the count of items may include folders as well as files (depending on the report you are running). Also note that the count is all the files profiled, which may include files inside zip files or other archival files. Hence, the count may be higher than a count of files provided by your operating system. If you have excluded processing archival files when processing, then the count should be the same as that reported by your operating system.

You can report over one or more profiles at the same time. If you are reporting over more than one profile, then you will be given statistics for each profile, and the totals across all profiles.

### 4.3.2 Breakdowns

Some reports offer these statistics broken down by another value. For example, DROID can produce statistics broken down by the year files or folders were last modified.

One important feature to note is, if your report is broken down by file formats (file format PUID or mime-type), then you will probably find that the final total count and sizes of the files are bigger than you will see in other reports. This is because files can be identified as more than one potential file format. Hence, when breaking down by file format, the same file can appear against more than one file format in the statistics. When adding up the totals for each breakdown, a file can be double-counted in the final totals. This is not incorrect, as the report is producing totals per breakdown, then totalling across all the breakdowns. However, you should be aware of this when interpreting the results.

### 4.3.3 Included reports

The reports included with DROID are:

- total numbers and sizes of files and folders
- total unreadable files
- total unreadable folders
- count and size of files by the year last modified
- count and size of files by the month last modified
- count and size of files by the year and month last modified
- count and size of files by the file format PUID
- count and size of files by the file extension
- count and size of files by the mime-type
- comprehensive breakdown (all of the above in a single report).

### 4.3.4 Exporting your reports

When you build a report, it will appear on screen in a new window. However, you will probably want to show the report to other people, or import the results into another program for further analysis (e.g. Excel). You can export your report to a file in several formats:

| | |
|---|---|
| **Portable Document Format** | PDF |
| **A web page** | HTML |
| **Plain text** | TXT |
| **DROID report XML** | XML |

For human-readable reports suitable for showing to other people, then the PDF or HTML options are the best choice. If you want to process the report statistics further using commonly available application software, then the TXT format is the best one to choose. This file can be directly imported into spreadsheets, where you can produce charts or different statistics.

The DROID report XML format is suitable for machine-processing using XML processing software or XML transformation scripts, allowing you to transform the report data into any other formats you may wish to produce. In fact, all of the above export options (except PDF) are produced by running what is known as “eXtensible Stylesheet Language Transformation” scripts (XSLT). These transformation scripts are found in the “`report_definitions`” folder and sub-folders in the DROID working area. You can define your own scripts, and add them to these folders to produce your own report exports in different formats you may require.

Finally, one particular report has an additional transformation script provided out-of-the-box. DROID ships with a “PLANETS XML” export option, available when running the “Comprehensive Breakdown” report. PLANETS[10] is a European-wide digital preservation project, which co-funded the recent DROID development.

## 4.4 Exporting your results

Exploring, filtering and reporting over your results can tell you a lot about the files and folders you have. However, sometimes you will want to work with all the information collected using other software. For example, you may wish to load the information into Excel

### 4.4.1 Multiple profiles

You can export one or more profiles to the same file. When you export, you can select which of the profiles you have open will be exported.

### 4.4.2 Comma Separated Values file

DROID can export file and folder information into a Comma Separated Values (CSV) file. This is a text file, with the columns of information written out separated by commas. There are two options for exporting to a CSV file:

- one row per file – each line in the CSV file represents an individual file or folder. If a file has more than one format identification, these identifications will be written out as extra columns at the end of each line. The number of lines in the CSV will be the number of files and folders in the (filtered) profile(s).
- one row per format identification – each line in the CSV file represents a unique format identification. If a file has more than one format identification, then that file will have more than one line in the CSV file. The number of lines in the file will be the number of different format identifications made in the (filtered) profile(s).

## 4.5 Analysing DROID information in external applications

### 4.5.1 Using exported CSV files

CSV files are a standard format, not controlled by DROID. You may find that if you double-click on a CSV file in your normal operating system, it automatically opens in an application

[10] PLANETS www.**planets-project**.eu/

(e.g. Excel, or Access). Which application opens a CSV file is configured in your own operating system, not by DROID.

Some applications may have limits on how large a CSV file they can import. For example, Microsoft Excel 2007 can load up to a million rows at one time, whereas Microsoft Excel 2003 can only import 65,536 rows. If your export contains more than the number of rows your application supports, you could consider splitting the file into separate files using a text editor, or you could filter your profiles to narrow down the files which appear in it before exporting.

One useful feature in Microsoft Excel and Access is the ability to split information in a column into more columns. For example, you can split up the file path into more than one column, with each column holding an individual folder in the path. This enables you to filter on particular locations more easily.

To do this, first make sure you have inserted a lot of blank columns to the right of the column you are splitting, to hold the split columns. Then highlight the file path column, and select "Text to Columns". When prompted, specify that the column is "delimited", and select Other, and type in a backslash "\" in the box provided. You should see that your file path is split into separate columns, with each column containing an individual folder name. With the data represented like this, you can easily filter on particular folders to find areas of interest.

### 4.5.2 Connecting to a DROID database

If you need to perform specialised queries on the droid database which DROID itself does not provide, you can connect to the DROID database using external software. DROID profile files (with the .droid extension) are in fact mini-databases, stored in a 64-bit zip file. Using appropriate software (e.g. WinZip), you can unzip a .droid file to another location on your disk. There should be a folder called "db" where you unzipped the .droid file, which is the DROID profile database.

DROID databases are Apache Derby 10.7 databases,[11] which many other database tools can connect to.

[11] See: http://db.apache.org/derby/

To connect to a DROID database, point your database tool at the "db" folder you unzipped, and connect to it using an Apache Derby 10.7 database connector. The username and password you will need to connect to a DROID database is:

Username: `droid_user`
Password: `droid_user`

A tool we have used successfully to do this is DbVisualiser.[12] However, note that Apache Derby 10.7 is a fairly recent release of the Derby database at the time of writing this guidance. You may have to update your database tool with recent Derby drivers to access it.

In DbVisualiser, you must download the 10.7 Derby binary distribution from Apache Derby, and then unzip the folder containing the file "`derby.jar`" into the correct sub-folder of DbVisualiser. If DbVisualiser was installed in `C:\Program Files\DbVisualiser`, then the correct subfolder is "`C:\Program Files\DbVisualiser\jdbc\derby\`". Once DbVisualiser has the latest database driver files, it will be able to read Derby 10.7 databases.

[12] See: http://www.dbvis.com/

# 5. Further reading

All Digital Continuity Project guidance is available at nationalarchives.gov.uk/information-management/our-services/dc-guidance.htm . Guidance that may be particularly useful to you includes:

*Evaluating Your File Formats* nationalarchives.gov.uk/documents/information-management/evaluating-file-formats.pdf

This guidance will help you to evaluate your file formats from a digital continuity perspective and to employ various strategies to maintain the continuity of your digital information.

*File Format Conversion* nationalarchives.gov.uk/documents/information-management/format-conversion.pdf

This guidance explains the issues in migrating information between different file formats. It will enable you to understand why, when and how you should convert file formats, and what you should convert them to.

*Risk Assessment Handbook* nationalarchives.gov.uk/documents/information-management/risk-assessment-handbook.pdf

Practical information and support to help you assess and manage risks to digital continuity – information on creating a framework for managing risk, carrying out a risk assessment, and mitigating risk.

# Appendix: Installing DROID

DROID is an open source tool, available for free download from the source-forge website: http://droid.sourceforge.net. If you have problems downloading this, please contact us at digitalcontinuity@nationalarchives.gsi.gov.uk.

To install DROID, download the latest zip file from the above website. Unzip the files into a suitable folder somewhere on your computer. DROID itself only needs read permissions to this folder when running (it does not write to its installation folder).

The source-forge version of DROID is created by authorised National Archives staff and contractors. As it is open source, anyone is free to examine and modify their *own* copy of it; however, the software offered from this web address is the **only official National Archives version**.

> **Note:** This section includes instructions for the installation, technical requirements and configuration of DROID. You may want to consult your IT department for help in setting this up.

### License

DROID is licensed under the permissive BSD open-source license.[13] It allows you to run, distribute and re-use the software as you like, without payment. The only restrictions the license imposes are in the preservation of copyright notices if you redistribute the software to others, and that you may not claim that the National Archives endorses any products derived from the software, without express permission.

## Technical requirements

### Cross platform

DROID has been engineered to run on as many platforms as possible. It is written in Java 6, and so should run on any platform which has a Java 6 Runtime Environment Standard Edition (JRE6 SE) available for it.

[13] See www.opensource.org/licenses/bsd-license.php

Note that very early versions of JRE6 SE will not work, as a required piece of code was not present until at least update 4. We recommend you use a recent version of the JRE6 SE (the latest being update 23 at the time of writing).

DROID can run using a JRE already installed on your computer, or it can use an embedded JRE provided as an alternative download. In general, we would recommend using the embedded JRE, as this is the version on which DROID has been most extensively tested. However, by using the embedded JRE, you are foregoing any security updates or bug fixes to the JRE environment which may be subsequently released.

**Memory requirements**

When running, DROID requires at least 512Mb of memory, which should be easily achievable on any computer of the last few years, which typically have between two to four times as much memory available. The maximum memory used by DROID can be configured (see below), including increasing the available memory (which may improve performance). However, if the memory is much lower than this, there is a risk of Out-Of-Memory errors, which may prevent DROID from running or completing a profile underway.

**Working area**

DROID stores settings and temporary files while operating in a working area, to which it must have create, read, write and delete permissions when running. By default, this working area is a folder under the user's home folder, called ".droid", although this can be configured (see below). Wherever the working area of DROID is set to, we recommend that your working area is on a local disk to the computer, not a shared network drive.

Depending on how many files you are profiling, you may require several gigabytes of disk-space to store the working profile. DROID stores profiles in a database, which can grow quite large when millions of files are profiled. This database is stored in a temporary working area, by default stored under the user's home folder. For very large profiles of around 20 million profiled files, we have seen working space requirements of the order of 10 gigabytes.

**Profile storage requirements**

Once files have been profiled, the results must be saved to a separate area from the working area. Unlike the working area, saved file profiles are compressed. To give some context, 20 million profiled files would compress into roughly four gigabyes of profile.

**Security requirements**

DROID requires the following access control permissions:

| | |
|---|---|
| **Installation folder** | Read-only access, once installed. |
| **Working area** | Create, read, write and delete permissions |
| **Files being profiled** | Read-only access |
| **Saved profile results** | Create, read and write permissions |

If you are profiling a large volume of files, which may contain sensitive information, then special care must be taken in configuring the access control security around DROID.

**Read-access to files for DROID**

There may be no single account which has read-only access to all the files you wish to profile, other than administrator-level accounts. In accordance with the principle of least privilege, the user under which DROID is run should have the minimum permissions required to carry out its task. **We do not recommend you run DROID with administrator-level permissions**.

This may mean that you have to create a special account with read-only permissions to all the files you wish to profile. Alternatively, you may profile different areas using different user accounts, and run DROID separately under each of these users, creating several different profiles which can then be moved and recombined for exporting or reporting.

**Access to the working area**

As DROID runs, it stores information about the files it has read in the temporary working area. It does not store any direct information about the content of the file, although it may store a hash signature of the contents, if this has been requested.

If DROID is profiling inside archival files (e.g. zip files), it may unzip the contents into its temporary area while working for further processing. These files are deleted after they have been processed. However, it is possible (e.g. if DROID crashes) for files to be left behind in this temporary area. Access to this area should be limited to the DROID user only.

**Access to saved profile results**

When DROID finishes, the results are usually saved to another file on disk. Just as for the temporary working area, all the information about the files profiled is in this profile file, which

may be sensitive. Profile results should be saved to areas which only authorised users can access – and should be treated at a similar level of security to the areas which were profiled.

### Configuring DROID

There are several options for configuring DROID:

#### Memory

To configure how much memory DROID uses when starting, edit the start-up text scripts you use to launch DROID. These files contain a parameter "`droidMemory`" which is the maximum amount of memory DROID will use, in megabytes. Simply edit this to the number of megabytes of maximum memory you want DROID to use, and save the script.

#### Working areas

DROID stores files in various places, including user settings, log files and temporary working areas. These can be configured using the startup scripts provided, which contain documentation on how to do this, or using system environment variables. Documentation on how to do this is included in the DROID Help.pdf file included with DROID.

#### General options

Most options can be configured through the Tools/Preferences menu in the DROID GUI. If you do not have access to the GUI version, or cannot run the GUI, it saves the configurable options into a file called "`droid.properties`", which can be found in the user's DROID working area. This is a simple text file, conforming to Java property file specifications.[14]

#### Updating signatures

The National Archives regularly publishes updates to the file format signatures DROID uses to identify file formats. DROID can automatically check for updates every time it runs, on a defined schedule, or not at all. It can also be configured to make newer signature files the default, or to simply download them so they are available, but not change the default version in use.

When you download signature files, you should place them in the DROID working area, under one of two folders: `signature_files` and `container_sigs`, depending on

[14] See: http://en.wikipedia.org/wiki/.properties

whether they are signatures for “signature” based identification or “container” based identification (see section 2.2.10 Identification method).

Alternatively, you can manually add a signature file to DROID by using the Tools/Upload Signature option. This allows you to download signatures on an internet connected machine and manually transfer them on to a machine which may not be internet connected for security purposes.