





## File Format Conversion

This guidance relates to:

-  Stage 1: Plan for action
-  Stage 2: Define your digital continuity requirements
-  **Stage 3: Assess and manage risks to digital continuity**
-  Stage 4: Maintain digital continuity

This guidance has been produced by the Digital Continuity Project and is available from [www.nationalarchives.gov.uk/dc-guidance](http://www.nationalarchives.gov.uk/dc-guidance)

© Crown copyright 2011

You may re-use this document (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence. To view this licence, visit

<http://www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm>

or write to the Information Policy Team, The National Archives, Kew, Richmond, Surrey, TW9 4DU; or email: [psi@nationalarchives.gsi.gov.uk](mailto:psi@nationalarchives.gsi.gov.uk) .

Any enquiries regarding the content of this document should be sent to

[digitalcontinuity@nationalarchives.gsi.gov.uk](mailto:digitalcontinuity@nationalarchives.gsi.gov.uk)

---

## CONTENTS

<b>1. Introduction</b>	<b>5</b>
1.1 What is the purpose of this guidance?	5
1.2 Who is this guidance for?	6
<b>2. Why convert file formats?</b>	<b>7</b>
2.1 Replacing a format	7
2.1.1 Software support	8
2.1.2 Standardising your formats	9
2.1.3 Moving from a locked-in format	9
2.1.4 Preserving your information	11
2.2 Creating additional versions	11
2.2.1 Sharing or publishing information	12
2.2.2 Using your information in new contexts	13
2.2.3 Aggregating information from different sources	13
<b>3. When to convert file formats</b>	<b>14</b>
3.1 On-demand conversion	14
3.2 Early and regular conversion	15
3.3 Late conversion	16
<b>4. How to convert formats</b>	<b>18</b>
4.1 Assess your information	18
4.2 Assess your environment	21
4.2.1 Potential naming conflicts	21
4.2.2 Access control issues	21
4.2.3 External references to files	21
4.2.4 File system metadata dependencies	22
4.2.5 Requirements to maintain links with the originals	22
4.3 Select your migration tools	23
4.3.1 Batch conversion	23
4.3.2 Conversion settings	23
4.3.3 Characteristic support	24
4.3.4 Error logging	24
4.3.5 Conversion performance	24
4.3.6 Test your tools	24

4.3.7	Dealing with failures .....	24
4.3.8	Managing the environment.....	24
4.4	Migrate your files.....	25
4.4.1	Quality assurance .....	25
4.4.2	Retention of originals.....	26
<b>5.</b>	<b>Further information .....</b>	<b>27</b>
	<b>Appendix: Format conversion checklist.....</b>	<b>29</b>

## 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. 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.

Information is kept in a variety of file formats. Over the life span of your information you may need to convert it into different file formats, either as an addition to its original format, or as a replacement. While converting file formats can protect information against digital continuity loss, the conversion itself brings risks and this process must be carefully planned and carried out to maintain the digital continuity of your information.

### 1.1 What is the purpose of this guidance?

This guidance will help you to understand the process of file format conversion, and will help you to understand:

- why you should convert file formats
- when to convert file formats
- how to convert file formats.

This guidance will not go into detail on how to understand your current formats or which new formats to convert to. However, a related piece of guidance, *Evaluating Your File Formats*<sup>1</sup> suggests a process for assessing which formats you should work with to meet your usability needs. While it may seem to be obvious which file formats are in current use, there is often a large amount of legacy information encoded in older or non-standard formats. The National

---

<sup>1</sup> See *Evaluating Your File Formats* [nationalarchives.gov.uk/documents/information-management/evaluating-file-formats.pdf](https://nationalarchives.gov.uk/documents/information-management/evaluating-file-formats.pdf)

Archives has created a freely available file format identification tool, DROID, to help you audit your file formats.<sup>2</sup>

This guidance will not tell you how to convert specific formats into other specific formats; there are too many formats and combinations of formats to allow this. It will give you the steps you should go through in performing a file format conversion process, and flag up areas of potential risk that you should consider.

This guidance forms part of a suite of guidance<sup>3</sup> that The National Archives has delivered as part of a digital continuity service for government, in consultation with central government departments. Assessing and converting your file formats is part of Stage 3 of the four-stage process of managing digital continuity.<sup>4</sup>

Format conversion may be an action you take to help maintain access and use of your information, mitigate risks that arise from technological or economic obsolescence,<sup>5</sup> or to facilitate the re-use and greater interoperability of your information.

## 1.2 Who is this guidance for?

This guidance is primarily aimed at information management and IT professionals charged with responsibility for assuring access to digital information stored in files. It will also be useful for anyone undertaking a file format conversion process.

See more on the roles and responsibilities that your organisation will require to ensure the digital continuity of your information in *Managing Digital Continuity*.<sup>6</sup>

---

<sup>2</sup> Download DROID at <http://droid.sourceforge.net/> and read *How to Use DROID and How to Interpret the Results* [nationalarchives.gov.uk/documents/information-management/droid-how-to-use-it-and-interpret-results.pdf](http://nationalarchives.gov.uk/documents/information-management/droid-how-to-use-it-and-interpret-results.pdf)

<sup>3</sup> For more information and guidance, visit [nationalarchives.gov.uk/digitalcontinuity](http://nationalarchives.gov.uk/digitalcontinuity)

<sup>4</sup> See *Managing Digital Continuity* [nationalarchives.gov.uk/documents/information-management/managing-digital-continuity.pdf](http://nationalarchives.gov.uk/documents/information-management/managing-digital-continuity.pdf)

<sup>5</sup> Technological obsolescence arises when technology to read or access information in a file format is no longer available. Economic obsolescence arises when the cost of maintaining (or re-acquiring) technology to access information in a file format is prohibitive.

<sup>6</sup> See more on roles and responsibilities in *Managing Digital Continuity* [nationalarchives.gov.uk/documents/information-management/managing-digital-continuity.pdf](http://nationalarchives.gov.uk/documents/information-management/managing-digital-continuity.pdf)

## 2. Why convert file formats?

The information your organisation needs will be held in a variety of formats. Each individual piece of information may itself be held in several formats so that it can meet a number of usage requirements.

When converting file formats you will either:

- 1) need to replace one format with another. For example, this may be due to changes to the software tools that are used in your organisation, a move away from legacy formats that are at risk of obsolescence, or changes to the standard format your organisation uses to publish online.
- 2) need to create an additional version in a different file format to meet your usability requirements. For example, a brochure was created in a desktop publishing format, but it must be converted into an additional format which can be published online.

There are a number of different drivers for each type of conversion. You should review your file formats periodically to assess whether any of them hold risks to your information. You can also pro-actively convert file formats to reduce the risk of digital continuity loss.

### 2.1 Replacing a format

Maintaining the digital continuity of your information means ensuring it is complete, available and usable, over time and through change. It means making sure that your business has the information it needs, and that the technology enables the information to be used in the way business needs it to be. To maintain your digital continuity you may need to convert file formats, as file formats naturally age and can become more risky, but also because your business may change how it needs to use its information, or because your technology environment changes.

If you do not pro-actively convert your formats, you may find that you are no longer able to access or use your information in the way that you need, or that to do so you are locked in to using particular pieces of software. However, when replacing formats you may be planning to remove support for the older file formats, and potentially deleting the original files altogether, which holds its own risks – so you must make sure your process and testing are comprehensive.

You may want to replace a format because:

- your available software doesn't support your use requirements, or will not in the future (see [section 2.1.1](#))
- you are standardising formats, or your standard formats are changing ([section 2.1.2](#))
- you are moving away from locked-in formats ([section 2.1.3](#))
- you are archiving information for long term preservation ([section 2.1.4](#)).

### 2.1.1 Software support

As software and formats change over time, information stored in older formats becomes at risk of obsolescence. Conversely, newer formats may be unreadable by older software. It is not necessarily a good idea to convert file formats every time you change software, as all conversions carry risk and cost, but it is vital to convert often enough that you do not lose your digital continuity. We address the question of when to convert later in this document, in [section 3](#).

The technology used to access information in your file formats is constantly changing, either as the technology itself changes, or as your organisation changes the components of the technical environment.

- **Evolving formats:** many file formats gradually evolve over time. While each small change may not in itself cause a loss of access, over time the accumulation of changes can mean that current software can no longer access older information, or vice versa. For example, you may find that files saved in older word processing formats no longer render correctly in later versions of the software, or cannot be opened at all.
- **Upgrading software:** when software you are already using is upgraded to a newer version, it can default to saving information in a newer file format, which may be incompatible with other software which needs to access the information, or with earlier versions of the same software. For example, Microsoft Office 2007 defaults to using a completely new set of file formats (OOXML) when saving information. When released, these formats were entirely unsupported on the Mac platform, and also with most other office suites, including earlier versions of Microsoft Office.
- **Removal of support:** it is not uncommon for software vendors to reduce support or entirely remove access to older formats when upgrading their software.
- **Changing software:** a change of software in your organisation can also require the use of entirely new formats, or make information in older formats unreadable, or with some



loss of fidelity. Even if two pieces of software have support for the same format, embedded features or functionality may not actually be transferrable – for example media embedded in a document, or macros embedded within a spreadsheet.

Organisations should not allow the file format of their information to be dictated by the software's selection of default, as the interests of the organisation are not necessarily aligned with the interests of the software vendor. Vendors may change a file format for a variety of reasons – to enable new functionality, to fix issues with earlier versions, to conform to a newer standard, or to create lock-in. Any change could impact adversely on your own organisation, and may introduce more incompatibility, cost and risk, so should be assessed carefully.

An alternative to conversion is to maintain older software for access to older or archived information, although this can be expensive, and you will run the risk of eventual obsolescence and complete loss of access to your information. While it is possible to keep older software running for some time, eventually it stops working on newer platforms. This can be somewhat mitigated by using virtualisation software to run older platforms in virtual machines. However, even where it is technically possible to keep running older software, eventually, the vendor will cease support, and will stop issuing security updates for it, which may pose other risks to your organisation. Finally, the skills in older software become harder to find and the costs of maintaining it rise as time goes on. Therefore, you cannot rely on running older software to access information in older formats indefinitely. You could combine both strategies, with older software available for a defined time, and a conversion policy for information older than this.

### **2.1.2 Standardising your formats**

Over time, organisations will collect information of the same type (e.g. images) in different formats. This may be due to changing software, a deliberate policy to use different formats, or a lack of control over which formats are created by users. In addition, information may be supplied by external bodies or the public in alternative formats. Having information in a variety of different formats increases the cost of managing access to that information and the risk of losing access.

Standardising the formats which you use can help reduce costs and risk, and increase your flexibility to use different software in the future. The goal is to periodically converge the formats you use to a stable set of formats to which you can guarantee access.

### **2.1.3 Moving from a locked-in format**

Some file formats are heavily tied to the software used to create them. This can increase the risk of loss of access to your information, reduce your business flexibility and result in more

expensive software procurement and licensing. You may need to convert from these formats because your organisation is changing the software that you are using, or it may be part of a wider objective to move towards more open information standards.

To better understand locked-in formats, standardisation and interoperability, see *Evaluating Your File Formats*.<sup>7</sup>

Locked-in formats can pose a conversion problem, in that it may be hard to convert them into different formats. This is not generally due to an absence of available conversion software, but rather because locked-in formats tend to have specific features which only the creating software can implement fully.

If you are converting a locked-in format, you must evaluate whether your legacy information uses features which are hard to convert, and whether those features can be safely discarded or replaced by similar (but not identical) features. For example, some complex formats allow mini programming scripts (i.e. “macros”) to be embedded within them. Typically, embedded programming languages do not survive conversion processes well, as they rely on specific interfaces with the specific technology they are embedded in.

You may discover that your existing information uses features of the locked-in format which are not possible to replace. In this case, you have some difficult choices to make:

- Do you accept the lock-in and not change formats?
- Do you accept the loss of certain features of your existing information?
- Do you preserve the appearance of the information in a read-only format (e.g. converting to an image, losing the ability to further edit, change or re-use the information easily)?
- Do you leave your existing information in the locked-in format and continue to maintain support for it, but move to the new format for new information?
- Do you preserve access to the legacy information by maintaining separate or different versions of software?

This reduction of choice (or the necessity to make hard decisions), is of course the essential feature of lock-in, and in itself is a good reason to avoid it, if possible.

---

<sup>7</sup> *Evaluating Your File Formats* [nationalarchives.gov.uk/documents/information-management/evaluating-file-formats.pdf](https://nationalarchives.gov.uk/documents/information-management/evaluating-file-formats.pdf)

#### **2.1.4 Preserving your information**

If you need to preserve access to static information for a long period of time, you should focus on the essential characteristics of the information you wish to preserve, rather than the functionality of the software used to create it in the first place. You may discover that you do not need all of the original functionality, and can instead use a much simpler format for preservation. For example, you can largely divide document formats into dynamic formats which support user-editing, and page-layout oriented documents aimed at printing or electronic publication. For preservation of documents, the latter may be a better format than the former. However, be aware that there are features which may not be preserved in such a conversion process, for example, change history or other embedded metadata.

In general, if you want to maintain your digital continuity over long periods of time, formats which are based on open standards, or are simple, are better than proprietary or complex ones. In addition, formats which are effectively read-only are usually less complex than ones which do support user-editing, as there are typically fewer features to support. Regardless of complexity or standardisation, if a format is widely used, with a lot of information encoded in it, there is a good chance that support for it will remain long into the future. You will need to schedule periodic testing of your preserved files to make sure they are still complete, available and usable in the way you need them to be.

## **2.2 Creating additional versions**

In many cases, rather than converting a file to a new format, you will be creating an additional version of your file in a different format to enable new forms of access and use. This does not mean that the original format becomes redundant, only that more than one format may be required to satisfy all requirements for the same information. However, you should not multiply formats unnecessarily – if a single format can provide all your access requirements, then this is usually the preferred option (see [section 4.2 Assess your environment](#) for more information).

**For example:** your organisation has a selection of stock images which are part of its branding. These images are used in a large number of different situations, and while there is a standard 'pack' of different versions, occasionally a situation requires a new version to be created.

- The 'master' of each image is a high resolution .psd file, which when opened with Adobe Photoshop contains layers, masques and additional functionality specific to the software.
- A variety of .jpeg versions of each image are stored at a range of resolutions and qualities.
- A print company requested the files as .tif files.
- A request is made for a version of the files to be shown on presentation screens, the decision is made to embed the files in a .ppt presentation format (an alternative was to embed them in a video format)

There are a number of reasons why you might want to create additional versions of your files in different formats:

- sharing or publishing information (see [section 2.2.1](#))
- using your information in new ways ([section 2.2.2](#))
- aggregating information from different sources ([section 2.2.3](#)).

### 2.2.1 Sharing or publishing information

Exchanging information includes the sharing or transfer of information, either with the public or with other organisations, as well as the publication of information (probably in a fixed unchanging state).

When you're creating new versions for this purpose, it's important to consider not only how other users want to use the information, but also their technical capabilities. You should not assume that the public have access to business technology at home. They may run entirely different operating systems, word processors or other software, or it may be much older than the technology typically found in a business environment. Likewise, if you're sharing information with other organisations, they may have a different IT infrastructure. You may need to convert your file into an older or more open format with wider support.

There may be additional requirements for your information when you publish or share it. For example if you are publishing datasets via the data.gov website there are specific requirements

for the format and form.<sup>8</sup> If you are publishing for the public there may be accessibility requirements which you must meet – for example, compliance with the Disability Discrimination Act.

### **2.2.2 Using your information in new contexts**

It is quite common to find that you need to use or make your information available in different contexts. For example, you may need to convert images to smaller and more web-friendly formats to publish online. Changing technologies and usage patterns can also demand conversion. For example, you may need to make information available on mobile platforms such as smartphones.

### **2.2.3 Aggregating information from different sources**

New information is frequently assembled from a variety of different sources. In many cases, the information must be transformed from its original format into a common format to allow the information to be aggregated.

For example, you may find raw data in databases, spreadsheets, XML documents and even web pages. All of these sources of information are incompatible with each other; hence interchange formats must be used. Your choice of formats will obviously depend on the software you use to assemble the information, and the features which you must preserve in the source data.

---

<sup>8</sup> Dataset format guidance <http://data.gov.uk/blog/guidance-very-basic-standard-file-format-data>

### 3. When to convert file formats

If you have determined that you need to convert information from one file format to another, you must also decide when the conversion takes place.

There are three basic strategies for when to convert file formats. The strategy you choose will be largely dictated by your driver for format conversion, but it may also depend on your technical environment or other business needs:

- on-demand conversion (see [section 3.1](#))
- early conversion ([section 3.2](#))
- late conversion ([section 3.3](#)).

Early and late conversions are really just variations on batch conversion processes, but with different risks and costs attached due to the timing of the conversion. These strategies shade into one another; the extreme ends are explained to demonstrate the different trade-offs involved. On-demand conversion is a completely different strategy, relying on servers to perform a conversion dynamically. There is no one “right” strategy to use – only by assessing your business needs can you determine the appropriate balance of risk, cost and benefit.

#### 3.1 On-demand conversion

On-demand conversion is the immediate conversion of a file to another format on receiving a request for the file in that format. It typically operates on a single file at a time, although batch conversions can also happen on demand. This process may be automated (see the example in the paragraph below) or it may require an individual to convert files manually upon request. This strategy may be applied to replacing formats, but the strategy is most often applied to create additional versions of files in different formats as the need arises.

For example, your website may offer documentation to users in a variety of formats (e.g. PDF, DOC and ODT). However, you do not store each document in each format. Instead, the underlying file is converted by the web server on receiving the request for the file in a different format. This strategy has both benefits and downsides:

Benefits	Downsides
<ul style="list-style-type: none"> <li>• You do not need to store several copies of each file in each format – only one file, with</li> </ul>	<ul style="list-style-type: none"> <li>• There is almost no possibility of quality assurance on the converted file(s). If you</li> </ul>

<p>the conversion happening dynamically on request. You may, however, store a converted file to speed up any future requests.</p> <ul style="list-style-type: none"> <li>You do not need to convert a large number of files in one go, which may be time consuming.</li> <li>Adding new files to the system is easy, as you have no need to provide them in all required formats upfront.</li> <li>The system can be updated to provide different formats as the need arises, again without having to process all your existing files up front.</li> </ul>	<p>adopt this strategy, you must assure yourself that the conversion process is sufficiently reliable for your requirements.</p> <ul style="list-style-type: none"> <li>The systems you are using may not allow you to issue dynamic requests for files in different formats. For example, if your files are accessed via a network file share, there is no way to intervene an on-demand conversion server.</li> <li>The system will require updating as different source formats are introduced.</li> <li>On-demand conversion may be slow, or place too great a load on your systems depending on the size, complexity and number of conversions.</li> <li>This strategy generally only makes sense for static information. If editing of the data by users is required, then an on-demand format conversion strategy may not work, unless there is a clear master version, and only that version can be changed.</li> </ul>
--	---

### 3.2 Early and regular conversion

Early conversion means that you have decided to convert files to different formats as soon as you can (but not on-demand). Early conversion is a batch-processing strategy, involving converting a body of files in a common format into another which better fits your business requirements and is generally a replacement process. For example, if you have decided to use a newer format provided by some upgraded software, you may convert all your previous files into the new format.

Benefits	Downsides
<ul style="list-style-type: none"> <li>The number of different formats you need to support is greatly reduced, converging your files on to a standardised set of formats. This can mean: <ul style="list-style-type: none"> <li>information is always encoded in a</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>Each file has more frequent conversion and each conversion has an associated cost and risk of information loss.</li> <li>If your original or new formats are fairly</li> </ul>

<p>currently supported format</p> <ul style="list-style-type: none"> <li>○ reduced support, maintenance and software licensing costs</li> <li>○ increased flexibility in choosing alternate software to use</li> <li>○ the risk of file format obsolescence becomes negligible.</li> </ul> <ul style="list-style-type: none"> <li>● You have the opportunity to review information and allow for <a href="#">quality assurance</a> of the files. With frequent conversion, these processes will be streamlined and each conversion will benefit from previous experience.</li> </ul>	<p>new</p> <ul style="list-style-type: none"> <li>○ conversion tools may not be as readily available, may have bugs or fail to deal with complex or unusual files well. This can also impact both the cost and quality of your conversion process.</li> <li>○ the new format may not be as widely supported, so you may also have to create additional formats if you need to share the information with users who have not yet upgraded.</li> </ul> <ul style="list-style-type: none"> <li>● If you need the same information to be accessible in multiple formats, storing all the converted files will take more space than using on-demand conversion.</li> </ul>
--	---

### 3.3 Late conversion

Late conversion means you have decided to defer conversion to the last sensible moment. Obviously, the definition of “last sensible moment” will vary based on your own assessment of the risks and benefits involved in your own environment.

For example, following a risk assessment of the file formats in use in your organisation, you may find that you have a large amount of legacy information recorded in ten different file formats, some of which are not accessible any more using current software. Some of this information may not be needed for active business use; hence a preservation strategy is employed. However, some of the information is still occasionally required, so a different format is selected for this.

Benefits	Downsides
<ul style="list-style-type: none"> <li>● Each file has less frequent conversion, therefore there is a lower risk of information loss and lower overall costs</li> <li>● If your target format is well established <ul style="list-style-type: none"> <li>○ there will probably be far more</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>● You will have a greater variety of formats in use in your organisation at any one time. This can: <ul style="list-style-type: none"> <li>○ increase support, maintenance and software licensing costs</li> </ul> </li> </ul>



<p>conversion tools available to use</p> <ul style="list-style-type: none"><li>○ existing conversion tools will probably deal with unusual or complex files better, as there has been time for bugs and edge-cases to be worked out.</li></ul> <ul style="list-style-type: none"><li>• You may be able to discard older information no longer deemed useful to the business, avoiding the need to convert at all.</li></ul>	<ul style="list-style-type: none"><li>○ reduce your flexibility to choose different software.</li><li>○ prevent older information from being usable in newer contexts.</li></ul> <ul style="list-style-type: none"><li>• You will probably have to convert a greater number of files and a greater variety of formats in one go, making this a larger project to manage and more complicated to quality assess.</li><li>• You may misjudge the “last sensible moment” and find that converting some information is now economically or technically unfeasible.</li><li>• If you need the same information to be accessible in multiple formats, storing all the converted files will take more space than using on-demand conversion.</li></ul>
---	---

## 4. How to convert formats

Any major format conversion project should be managed using your organisation's change management processes<sup>9</sup> including making appropriate impact assessments, risk analysis, quality assurance and communications. You will need to work alongside a number of different people in your organisation, including the relevant Information Asset Owners (IAOs) and primary users of the information so that you understand their requirements, and that they understand the changes.

This section presents a simple methodology for converting files from one format to another. It will give you the steps you should go through in performing a file format conversion process, and flag up areas of potential risk that you should consider.

Assuming you have already understood your drivers for conversion, and chosen when you need to convert your files, you should follow the following four steps to convert your files:

- assess your information (see [section 4.1](#))
- assess your environment ([section 4.2](#))
- select your migration tools ([section 4.3](#))
- migrate your files ([section 4.4](#)).

### 4.1 Assess your information

When assessing your information, you need to consider your business requirements – that is how you need to be able to find, open, work with, understand and trust your information.<sup>10</sup>

These requirements may not be immediately obvious and you should liaise with the owner and principle users of the information to ensure all their requirements are met. This will help inform whether the information contained in the formats you are migrating from have particular characteristics that you want to ensure remain unchanged. Some conversion processes only change the format of the underlying information, but many conversion processes will alter some aspect of the information as well. In general, very simple types of information can survive a conversion process without change, but complex information will be altered in some way.

---

<sup>9</sup> *Digital Continuity for Change Managers* [nationalarchives.gov.uk/documents/information-management/digital-continuity-for-change-managers.pdf](https://nationalarchives.gov.uk/documents/information-management/digital-continuity-for-change-managers.pdf)

<sup>10</sup> See *Identifying Information Assets and Business Requirements* for more information [nationalarchives.gov.uk/documents/information-management/identify-information-assets.pdf](https://nationalarchives.gov.uk/documents/information-management/identify-information-assets.pdf)

For example, you may be converting from one document format to another. It is possible that while the text of the document remains unchanged, the pagination, colours, styles and fonts used within it will be altered in conversion.

Before conversion, identify the key characteristics of your information that must survive conversion without (or with little) change. You should be aware that features which you do not regard as essential may in fact be essential because of the way in which they have been used. For example, while you may not regard the colours in a document to be important, users may have annotated minutes using green to indicate things that are complete, and red for things that are unfinished. Or the pagination of a document may change, breaking page references embedded in the document, rendering a contract unusable. It is important to review your information to determine whether changes to an aspect of your information can subtly affect the meaning of it.

There are often some less obvious characteristics you also should consider, typically related to complex or hidden functionality in the format. Below is a non-exhaustive list of a few of them:

Characteristic	Example	Factors to consider
Embedded metadata	Many formats allow various pieces of descriptive metadata to be embedded in them. For example, documents recording the author of the document, and photographs recording the geographic location at which it was taken and the camera settings used.	<ul style="list-style-type: none"> <li>• Whether any embedded metadata is required in the converted files.</li> <li>• Whether your conversion tools will move this information across.</li> </ul>
Embedded objects	Many complex formats allow other files or formats to be embedded within them. For example, documents may contain embedded images or spreadsheets, or presentations may contain videos.	<ul style="list-style-type: none"> <li>• Not all conversion tools will be able to deal with all kinds of embedded objects.</li> <li>• You must test files with embedded objects to quality assure that the conversion process will work for them.</li> </ul>
Scripts and macros	Some formats can contain mini-	<ul style="list-style-type: none"> <li>• If you need the support of</li> </ul>

	<p>programming languages. For example, documents often have a macro feature which allows common tasks to be automated. In general, scripts and macros do not survive conversion processes, unless the conversion is from one version of a format to another of the same format. Occasionally, another format will provide the same support for the same embedded scripts or macros, or provide equivalent ones, but this is rare.</p>	<p>scripts and macros in your files, you may need to rewrite these manually for the newer format.</p>
Digital signatures	<p>Some files allow digital signatures to be embedded within them (or you may have digital signatures in external systems relating to those files). Digital signatures validate that a file was signed by an authorised user, using strong cryptography over all of the information in the file to prove the assertion.</p>	<ul style="list-style-type: none"> <li>Any converted file, being different to the original, will lose this digital signature (or the signature will no longer be valid), and you will need to produce a new digital signature for it.</li> </ul>

You must make sure that your new format supports the required capabilities and that the conversion process will maintain the characteristics through the transfer. If your new format does not support the capabilities you may need to re-evaluate your choice of format, or whether to migrate at all. The process for assessing file formats is described in another document: *Evaluating your File Formats*.<sup>11</sup>

<sup>11</sup> See *Evaluating Your File Formats* [nationalarchives.gov.uk/documents/information-management/evaluating-file-formats.pdf](https://nationalarchives.gov.uk/documents/information-management/evaluating-file-formats.pdf)

## 4.2 Assess your environment

Before proceeding with any file format conversion process, you should assess whether there are any dependencies on or from your wider information environment which may cause problems once your files are converted. You need to determine whether there are:

- potential naming conflicts (see [section 4.2.1](#))
- access control issues ([section 4.2.2](#))
- external references to the files which are to be converted ([section 4.2.3](#))
- file system metadata dependencies ([section 4.2.4](#))
- requirements to maintain links with the originals ([section 4.2.5](#)).

### 4.2.1 Potential naming conflicts

When converting files from one format to another, there are two principle issues you must be aware of:

- the file name (including the file extension) is changing – this will break any external references or links to the file (see [section 4.2.3](#)). You should also make sure that the new name does not overwrite another file accidentally.
- the file name (including the file extension) is NOT changing – this may lead to the original file being over-written. However, even if you are planning to delete the original file you should not do so until full testing is performed. You may need to move the original files, or create the new versions in another location (which may in turn, break external references).

### 4.2.2 Access control issues

If the files had access restrictions on them before the conversion, these restrictions will likely need to be applied to the converted files as well. These restrictions may have been applied on a file-by-file basis, or based on the location of the file(s) so you must make sure they still apply to your new format version.

### 4.2.3 External references to files

You may find that files have references to them, for example on intranets or web servers. In some cases, other files make reference to files by filename – for example, Office documents can make a link to an external spreadsheet. You may also find file references inside databases or other storage systems. These references are usually made using the file name and location of the file. However, some systems (e.g. Content Addressed Systems) use a “hash” of the content of the file to reference it. After conversion, the hash value of each file will be different.

If an external system, or other file, makes reference to the original file on the basis of its file name, location or content hash, then after the conversion these links will be broken. This means that as part of the conversion process for these files, you must also rewrite or add the external references to them, and/or change the final location of the migrated files once conversion is complete.

#### **4.2.4 File system metadata dependencies**

When files are converted, dates and times on the converted files (and the files being converted) can change. Many file systems store metadata such as the creation date, the last accessed date, and the last modified date. When creating new files in a conversion process, all of these dates will be set to the time of conversion.

You may be relying on these dates to support retention planning, or other systems may adapt their behaviour based on these dates. For example, a web server may offer files for download which have been changed in the last month – clearly if all files are migrated then this system would offer every single file at once.

Various tools exist which can alter file system metadata – to enable you to ‘restore’ the original metadata (for example by altering the ‘last modified’ date on a file back to the original date, rather than the date of the format conversion). However, this is a difficult decision to make, as clearly by altering the metadata in this way, you are modifying the genuine history of these new files, which may have other impacts on future decision making.

In addition, there may be other file-system metadata properties which apply to the original files – whether they are read-only, or hidden, or some other file-system-specific setting. You must decide whether you replicate the original file system metadata settings on the converted file.

#### **4.2.5 Requirements to maintain links with the originals**

Once any conversion process has ended, you will need to be able to find both the original file and the converted file. At the very least, you will need to do this for quality assurance purposes in the short term. You may also want to retain the file in both formats so it may be repurposed for use in different environments.

For example, if you encounter a problem with a conversion, it may be necessary to go back to the original and re-convert it with different tools or options. Alternatively, if you are given a

smaller image suitable for web publication, you may need to go back to the full size image for print purposes.

You must consider how to maintain links between the original and the converted files. There are many ways to achieve this. For example, you may write out the converted files in the same logical structure as the originals, or by recording a simple database of the names and locations of the originals and the converted files (assuming their location will not change after conversion).

### **4.3 Select your migration tools**

There are a vast number of tools available to convert files from one format to another. Some are proprietary, some are freeware and some are open-source. However, the coverage of formats can be patchy. For widely-used formats, such as images, there may be many choices, but for niche or older formats your choices may be highly limited. For formats with poor support, you may have to perform two conversions, using an intermediate format to bridge the gap between the format you have and the one you would like. In some cases, you may have to commission bespoke software to perform a conversion, particularly if your file formats are themselves bespoke. Only by evaluating your information, formats and environment can you decide which tools may work for you.

There are also companies which will perform file format conversion as a service, using their own tools to do the conversion and managing the process for you. These services are available as part of the Digital Continuity Framework.<sup>12</sup>

#### **4.3.1 Batch conversion**

Some conversion tools are aimed at single file conversion, in that they literally take one file and produce another in the new format. If you want to convert multiple files, you will either need to use a tool which lets you convert entire folders or sub-folders, or you may need to write scripts to automate the batch processing of multiple files.

#### **4.3.2 Conversion settings**

There are often many ways to convert a file from one format to another with the same tool. You should check that the settings you use are performing the conversion with the appropriate features enabled or quality settings. For example, when converting video from one format to

---

<sup>12</sup> See more on the Digital Continuity Framework  
[www.buyingsolutions.gov.uk/frameworks/contract\\_details.html?contract\\_id=1190](http://www.buyingsolutions.gov.uk/frameworks/contract_details.html?contract_id=1190)

another, you may find that the default conversion settings are using a very high compression (and thus lowering quality).

#### **4.3.3 Characteristic support**

It is important to assess whether the tool fully supports the essential characteristics and metadata you are trying to convert, not merely that it converts from your source format to your target format. Complex formats can often be written out in different ways, which may impact on whether the characteristics you are trying to preserve will survive the conversion process.

#### **4.3.4 Error logging**

You should check how your migration tools report errors with conversion, and have strategies to deal with this error output. Some tools may simply silently fail to convert a file; others may produce an error log, or write information out to the console.

#### **4.3.5 Conversion performance**

Some tools may be noticeably slower than others when converting complex formats, particularly large media files. You should assess whether the performance of a tool will allow you to convert the number and volume of files you have with the time and resource available.

#### **4.3.6 Test your tools**

Once you have found a set of migration tools, it is essential to test them on a representative sample of your information to assure yourself that they can perform the conversion to an acceptable level of quality.

#### **4.3.7 Dealing with failures**

You must pay particular attention to conversions which are not successful. It may be that the majority of files convert successfully, but certain files with particular characteristics do not. In these cases, you may use another tool on these files, or accept a degraded quality of conversion. You must also determine if it is possible to detect failed conversions automatically, or if you can identify files which will fail before attempting a conversion.

#### **4.3.8 Managing the environment**

If you have identified your need for references in external systems, updates to file system metadata, or access control changes, you must consider how to integrate these updates with running your migration tool(s). You must also consider how to record the results of migration, linking the migrated files back to their originals.



Changes rarely happen in isolation. If you make a change to one piece of information or technology there may be extensive follow-on effects on other information and technology. Ideally your organisation will have documentation for the interactions of its technology and information (for instance, in an Information Asset Register or Configuration Management Database<sup>13</sup>). You can use this to understand the various impacts which arise from the migration.

## **4.4 Migrate your files**

Once you have gained an understanding of your information and environment, and selected your formats and tools, you are ready to begin converting your files. Before undertaking a large-scale conversion, you should run a pilot conversion on test systems, to provide assurance that all the files and systems are updated correctly.

### **4.4.1 Quality assurance**

Even after initial tool testing, unless you are choosing to use an on-demand conversion strategy, you should define further quality assurance processes to make sure that essential information is not being lost in conversion. The quality assurance criteria should be agreed in advance with the Information Asset Owner and primary users. You must assess whether the properties and functionality which had to remain the same have done so, and that the elements which were planned to change have changed as planned.

Manual checking of each file may be impractical, so spot-checking on a representative sample of your converted files may be the best option, opening both the original and the converted file for a direct comparison. The end users of the information should be included in this process as they may spot subtle problems which non-users would not.

An automatic method of checking the results of conversion is to use some of the invariant characteristics of your information, which you stated should be unchanged by the conversion process. Using metadata extraction tools, you can compare the before and after values on the original and converted file to provide assurance that the conversion process has been successful. This information could be recorded in the same database you use to maintain links between the originals and the converted files.

---

<sup>13</sup> See *Identifying Information Assets and Business Requirements* for more on information assets and the development of an IAR [nationalarchives.gov.uk/documents/information-management/identify-information-assets.pdf](http://nationalarchives.gov.uk/documents/information-management/identify-information-assets.pdf)

**For example:** you could extract the dimensions of an image and ensure they are the same before and after conversion, examine the audio-length (not file size) of converted audio files, or compare embedded metadata (e.g. the author) of a document.

If you are interested in performing automatic characteristic checking, then you must also select your metadata extraction tools and test them on the originals and converted files. You may have to use different metadata extraction tools for the source format and the target format, and convert their output to a common form to facilitate comparison. You will probably have to create custom software or scripts to carry out such a process, which can make this process cost-prohibitive. However, there may be situations in which the conversion of a large body of essential information makes it essential for assurance purposes.

If you have had to rewrite or update external references to the converted files, or alter file system metadata, or access controls, you must also test that these are correct after conversion.

#### **4.4.2 Retention of originals**

Even if the converted files are intended to fully replace the originals, you should keep hold of originals for a defined period, to guard against accidental functionality or information loss. This remains true even if your quality assurance processes have shown that the conversion is fully successful. This acts as mitigation against the risk that previously unknown business requirements may become evident after the conversion process, requirements which mandate that different characteristics of the information should have been preserved. This should be balanced with conflicting drivers around reducing the amount of data you store.

**For example:** a conversion of a large body of documents to PDF format was undertaken which was deemed successful. However, at a later date, when these documents were loaded into a search engine, it was discovered that the search engine could not index them. This is because PDF files can store converted documents in several ways – in some cases as high quality images of the source documents, in others as actual text. Since the business requirement to search had not been properly understood prior to conversion, no tests for this had been carried out, and the assumption had therefore been made that the conversion was successful. All of the original documents had to be converted to PDF again, this time using a different tool to preserve the underlying text instead of rendering them as images of text.

It is impossible to say how long the originals should be retained for, as this will depend on the criticality of the information, your drivers for conversion, your organisation's risk appetite, confidence in the conversion process, and the costs of retaining the originals and maintaining links between them and the converted files.

## 5. Further information

### Digital continuity guidance

All Digital Continuity Project guidance is available at [nationalarchives.gov.uk/information-management/our-services/dc-guidance.htm](https://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](https://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.

*Risk Assessment Handbook* [nationalarchives.gov.uk/documents/information-management/Risk-Assessment-Handbook.pdf](https://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.

### File profiling tool

The National Archives offers a file format identification tool (DROID) which can help you understand the format, volume and ages of the information you hold. Reports generated from DROID may also help you to identify opportunities for disposing of redundant information or to identify possible mitigating actions. Find out more about DROID and download our guidance

*DROID: How to Use It and How to Interpret the Results* on our website:

[nationalarchives.gov.uk/droid](https://nationalarchives.gov.uk/droid). You can download DROID directly here:

<http://droid.sourceforge.net/>.

### Conversion services procurement

There is a Data Migration Services lot on the Digital Continuity Framework, from which you can procure expertise in data migration projects. See more on the Digital Continuity Framework:

[www.buyingsolutions.gov.uk/frameworks/contract\\_details.html?contract\\_id=1190](https://www.buyingsolutions.gov.uk/frameworks/contract_details.html?contract_id=1190)



## Appendix: Format conversion checklist

### Why are you converting formats?

- Replacing a format
  - because of a software change
  - to standardise formats
  - to move away from a locked-in format
  - for long term preservation
- Creating a new format
  - to share or publish information
  - to use in a new context
  - to aggregate information from different sources

### When are you converting formats?

- On-demand
- Early and often
- Late conversion

### How to convert formats

- Assess your information, have you:
  - confirmed the business requirements for the usability of the information
  - confirmed the characteristics which must be maintained
- Assess your environment, have you considered requirements for:
  - naming strategy
  - access controls
  - external references to files
  - file system metadata dependencies
  - maintaining links with originals
- Selecting migration tools – have you considered: batch conversion
  - conversion settings
  - characteristic support
  - error logging
- Migration process - do you have a strategy for:
  - quality assurance criteria and a test plan
  - whether the original format is being deleted and how long to retain it for