

# Metadata

This document is written for people who have an idea that they need to “do something” with metadata, usually to help manage a dataset collected for or used in research. While there are plenty of metadata resources available on the web, the diversity of styles and target audiences for those resources can very quickly bewilder anybody reasonably new to metadata concepts. This document is intended to provide a simple working-level view of the needs, the issues, the processes around metadata collection and creation, without getting too discipline-specific. It also tries to avoid listing online resources that can date or move location; the web is an ever-evolving pool of resources, good and bad. At the end of this you should be able to find and read other metadata-related documents and see how they apply to your needs.

## Table of Contents

---

1. Remind me, what is Metadata? .....	2
2. Why create and collect Metadata?.....	3
3. Dealing with Metadata .....	4
3.1 Levels of Metadata .....	4
3.2 Types of Metadata .....	5
3.3 The <i>lingua franca</i> of metadata .....	6
3.4 Metadata Standards .....	9
3.5 Collecting and keeping Metadata .....	10
4. Problems with Metadata .....	12
4.1 It’s not all good? .....	12
4.2 Avoiding common mistakes.....	13
5. Getting help with it all.....	14
6. Appendix: State-based eResearch Organisations .....	16



# 1. Remind me, what is Metadata?

---

Every guide to metadata starts with its own definition of metadata, and there are plenty of them out there, from Wikipedia to any number of individual and organisational blogs. However, many of these draw from highly experienced people in the libraries and archives on the one side, and the informatics specialists on the other, and the language used to convey meaning can become very dense.

This guide takes a more pragmatic view to metadata. While generally ‘meta-data’ is summarised as ‘data about data’, what does that actually mean?

- *Metadata can actually be applied to anything.* It is possible to describe a file on a computer in exactly the same way as one would describe a piece of art on a wall, a person on a job, or a place on a map. The only differences are in the content of the metadata, and how it refers to the “thing” being described.
- *Metadata is data.* It is snippets of information that have particular meaning, in relation to some other piece of information (the raw data). It can be treated and managed like any other form of information, in terms of how it is created, managed, stored, and so on.
- *Metadata generally has little value on its own.* Metadata is information that adds value to other information. A piece of metadata like a place or person’s name is only useful when it is applied to something like a photograph or a clinical sample. There are, though, counter-examples, like gene sequence annotations and text transcripts of audio, where the metadata does have its own value, and can be seen as useful data in its own right. It’s not always obvious when this might happen. A set of whaling records (metadata about whale kills, in the 18<sup>th</sup> century) ended up becoming input for a project on the changing size of the Antarctic ice sheet in the 20<sup>th</sup> century.



# 1. Why create and collect Metadata?

---

The reason that metadata is created or collected for any data is that it enables and enhances the use of that data. When you are planning the processes for collecting metadata, think about how the data will be used, across the entire research lifecycle.

- *Finding data:* The first step will be to find data that is relevant for a project, through some kind of search. Data formats such as text can be indexed and searched themselves. However, knowing that a document contains your search text doesn't tell you anything else about the document, like a simple Google search. Furthermore, most searches today are still done through text, which limits searches for other formats like audio, images and video. That implies there is metadata that can be indexed and searched, which may include geographical questions ("what data is there for this region?"), times and dates ("what data is there for this period?"), right through to very discipline-specific terms ("who has a gene sequence of a mouse with this particular hereditary feature?") and administrative issues ("who is allowed to read this data?").
- *Using data:* To make use of any dataset, researchers need to understand how the data is structured, what it describes, how it was collected and so on. This means there needs to be a description of the processes that led to the data, what has been done to it since then, and how the data should be read. This tends to become discipline-specific, but there are common processes, for example with particular file formats. One example might be images, where metadata might include the image size, the use of any light filters, the camera model, the name of the photographer, and so on, on top of the 'finding' data described above, like time and date.
- *Preserving and re-using data:* There are many examples in academia where data collected for one project becomes useful in another project, typically in the same broad discipline, but sometimes well outside of the discipline that collected it. The needs for finding and using data remain the same but now stretch further into the future, and require a higher level of trust in the data. Astronomers for example are extremely careful in calibrating optical images under different filters that allow only certain colours through. If that information, about the filters used, is separated from the image, the data loses a lot of its value. Similarly, if a photo of a crime-scene has any modifications made, and those changes are not properly recorded, the data loses a lot of its credibility, and value.

A very important aspect is that good metadata helps you to manage people risks. People are fallible, and forget things, or leave projects, and any knowledge they have about some data should not live just in their heads or their fragile notebooks.



## 2. Dealing with Metadata

---

This section outlines the major components you need to consider when planning your metadata collection and management. It can never be a complete checklist of everything that is needed, but it should trigger the planning, along with the known uses of the data as described earlier.

### 2.1 Levels of Metadata

---

Metadata is “attached” in some fashion to the raw data. Sometimes the raw data can be collected into groups and share a lot of common metadata. For example, all photos taken on a certain day, at a certain place, for a certain project, will share the same location, date and project metadata. However, individually their content is quite independent. This leads to the idea of “item-level” metadata, which describes individual pieces, and “collection-level” metadata which describes all of the items in a collection that have some level of commonality.

There are benefits in dealing with “collection-level” data over “item-level”, such as less effort required and less risk of errors, but it limits discovering individual items within a larger collection; for example, finding a library can be easy, but then finding a specific book requires more information. In many cases it can be useful to describe a “collection” very broadly so it can be easily found, but then searching within a collection may require more specialist information.

If you build a big enough index, you can identify a book and its library in the one search, but it is generally much harder to do that for complex scientific data, where the search terms depend on the discipline. A search engine for geology looks very different to one for social sciences. But, a researcher looking at indigenous land use within a region may want to search for both at the same time, so searching for collections of data relevant to a region, regardless of discipline, might be a crucial first step.

This leads to a perennial question: what is a collection? There is no bound on the number of answers to that. Some individual items may belong to multiple different collections. A herbarium catalogue, for example, could be split into species, or geographical coverage, or a host of other slices, all of which have valid meanings for particular searches. A Google search returns a “collection” of pages with common search text, but each of those pages may individually turn up in other searches (other “collections”). This is one of those questions where it best to think about the use of the data, how it will be found and accessed. At some point, aggregating every item into a single “collection” (‘everything I have ever measured’) does not add value to a search. Conversely, stripping every item out individually makes the complexity of managing it all much harder. If it is a problem in your application it may be useful to present multiple “collections” of the same underlying data, and let users find their own way into it. There is no good single answer.

Apart from “collection” and “item” levels there is often a further level, described as “sub-item” level. While many things can be described as a single object (e.g. a photo, a statue) sometimes they can be made up of component parts: a book has chapters, a documentary has scenes, a musical score has movements, a relational database has tables, and so on. To find a particular component requires not just the “item-level” metadata but also its “sub-item” or “component” level metadata. Obviously this increases the effort required, but it allows for much better searching and can save a lot of time later.



*Example:*

*Every major TV station maintains a video archive of the recorded materials that they have broadcast (such as cartoons and documentaries), or their own recordings of materials that were initially broadcast live (such as news, sports and current affairs). To support effective use of the archive, the recordings are usually described at several levels:*

- *At the “collection” level they usually identify the type of program – news, current affairs, documentaries, cartoons, soap-operas, comedies and so on, rather than say the date of broadcast which may happen more than once, and doesn’t tell you anything about the show.*
- *At the “item” level they identify the particular episode of a program. For ongoing series they might be given a set of labels like ‘Series 10, Episode 5’ of ‘The Simpsons’, which also has a title of ‘When you dish upon a star’. For live shows such as News it may be labelled as ‘7pm News, Sydney, 1 January 1962’.*
- *At the “sub-item” item level of a News episode they will identify a particular story, so for example the ‘7pm News, Sydney, 1 January 1962’ item will have a reference to a story on ‘Independence for Western Samoa’ as being the second story, five minutes and 14 seconds from the start.*

*This hierarchical structure allows users to find particular materials directly, rather than scanning manually through a large number of complete recordings within even a single archive. At the very top a registry of collections might identify who has a video archive and what kind of collections they hold, but finding the right video clip, and the right point within that clip, requires ever finer information granularity and more specialised search tools. You could take it even further and identify individual scenes within a story. In this example it could help someone to find some stock footage of Western Samoa that could be re-used in another story or film, or footage of an individual being interviewed that could be of interest to a genealogist. More work early on in describing the data will (hopefully) reduce work later.*

## 2.2 Types of Metadata

---

Another way to think about metadata is to look at what useful functions it performs. There are many types of searches and other activities that support data use. Associated with those are different types of metadata. There are many lists of different metadata types you can find around the web, and the purists will argue which ones are right. Again, pragmatically it is very useful to keep these metadata types in mind during your planning, to ensure you have the coverage you need, and so any of the lists can be helpful.

For researchers, it is useful to break the types of metadata down, very simply, into

- *Scientific metadata*, and
- All the others.

*Scientific metadata* is all the information that is very specific to the study, and is needed to use and interpret the data collected. It can be very discipline specific, and usually requires some knowledge of the domain.

The other metadata types are broadly administrative, and deal less with the science and more with the process of how the data came to be and how it is managed. The most common types that people



identify include (with some example questions):

- *Provenance metadata*: This relates to the origin of the data, and ranges from the human to the highly technical.
  - E.g. Where did the data come from? Why was it collected? Who collected it, when and where? What instruments/technologies were used to collect the data, and how were they set up? What has been done to the data since it was collected?
- *Rights and access metadata*: This provides information about access and usage rules.
  - E.g. Who is allowed to view, edit or otherwise modify the data, or the metadata, and under what conditions? Who has some kind of authority over the data? Who has the authority to change the rules? Are there costs associated with access? Who has accessed the data, and what have they in turn done with it? Under what licence is the data being made available?
- *Structural metadata*: This provides fundamental information for a person or a computer to read the data.
  - E.g. How is the data set up? What formats, and versions of formats, are used? How is the database configured? How does it relate to other data?
- *Preservation metadata*: This builds on the history from the Provenance and Rights metadata, but also includes information to help build a sense of trust in the data, and allow for the data to be used long into the future.
  - E.g. Is the data authentic, authoritative, and original? Has there been any restructuring, e.g. due to software and file-formats changing? What software has been used to access the data in the past?

Clearly all of these administrative types overlap in a variety of ways, and the names change depending on who writes the list. Some may require new processes to be set up as part of the data collection work. While not every dataset will have or need every type, they do however identify characteristics about the data that need to be thought about, as early as possible, if the data is going to be truly trustworthy and re-usable into the future. It is always wiser to record too much information about the data than too little.

## 2.3 The *lingua franca* of metadata

---

As mentioned earlier, when you get into the world of metadata, you find a lot of information written by highly educated experts who use a rich language that seems all their own. To be able to plan for, collect and apply metadata does not require an in-depth knowledge of all of this technical language. Some concepts, though, are worth learning to start with and will greatly help you to find and make the best use of metadata frameworks.

### 2.3.1 Standards

To be truly useful, metadata needs to have agreement. This can be as trivial as agreeing on language, spelling, on date formats and even which calendar to use, on place names, and so on. If you are using dates differently to everyone else it will make it harder for people to search your data compared to other data. This implies a common language for metadata, based on agreed-upon standards.



### 2.3.2 Schemas

The first key component of metadata is the concept of a **schema**. Metadata schemas are essentially just a plan, an overall structure for your metadata, which describes how all of your metadata is set up. It is quite likely that your overall metadata plan will build on not just one but several schemas at the same time, some of them with long and illustrious histories of wide adoption and others that may be a little more experimental and newly developed. A schema specifies the content and format of metadata and how it should be organised.

The well-established schemas tend to address things that have been managed and used for a long time, like dates and places, common file formats, and processes such as preservation, and it is best to use these. New schemas emerge all the time to support disciplines more specifically, as well as new types of technologies and instruments, and as people trial these they can quickly evolve and eventually settle down.

One of the benefits of most good schemas is that they are usually extensible, i.e. they can be added to or fine-tuned as people identify issues with them. Obviously though, this can also cause problems when some people use such extended schemas and others don't. As long as you are aware of the possibility, it can usually be managed.

### 2.3.3 Namespaces

A schema is a single plan for some set of metadata elements, values, terms, concepts, etc. When you have multiple schemas in use you can have a situation where there is some overlap on common terms.

*Example: the term "Title" appears in many schemas. However, the way a "Title" is defined or recorded may be different depending on the schema. There are multiple ways that TV show episodes can be "titled", based on the series name, the broadcast series and episode numbers, e.g. as before "The Simpsons", "Series Ten, Episode Five", or free form text names, in this case "When you dish upon a star". A generic schema might label all of these as just a "Title" and leave it up to you to disentangle them. A more video-specific schema might have a SeriesTitle ("The Simpsons"), an EpisodeSeriesNumber ("10") an EpisodeNumber ("5") and an EpisodeTitle ("When you dish upon a star"). This makes it much easier to make lists and carry out searches.*

To avoid the potential confusion that can arise, schemas also define a **namespace** which is effectively a tag that says we are using the term "Title" from this particular schema. That way you can have multiple "Title" fields in your metadata and be able to tell which standard they follow. Namespaces basically tie metadata terms to their specific context.

One of the most common generic schemas is called the Dublin Core, and is the most widely adopted schema for descriptive metadata to date. It achieves this by being extremely generic, with just 15 terms (in the "core", but extensible), including *Title, Date, Type, Format, Creator, Coverage* and *Rights*. Each field can be basically free form text, with some restrictions. Dublin Core is used a lot on web pages, with the "DC" namespace – so you will often see things like "DC.Title" as a metadata tag inside a webpage's html headers. If you are working with video you may use Dublin Core and the MPEG-7 standard for video archives and so get the "DC.Title" and the "MPEG7.Title" fields. Each has their own rules how you can use them.

### 2.3.4 Metadata content

Within each schema are a set of metadata fields, sometimes called a parameter, with a label like "Title", and then its value. For simple text fields it is usually up to the metadata creator to put in something reasonable. However, many fields have additional structure within them, and to make comparisons easier they need to follow agreed standard structures. A good example is times and dates. There are an



unlimited number of ways people can convey a time and date, sometimes in quite an ambiguous fashion – a date like 2/3/11 can be interpreted quite differently depending on your background and country. If you want to do a comparison, like “is this date earlier or later than that date?”, then you need to be able to read the time and date in an unambiguous fashion, and ideally in a way that is easy for computers to read. This leads to more detailed concepts you will need to understand.

### 2.3.5 Taxonomies

A **taxonomy** is, roughly speaking, a structure that provides a classification of terms within it, and how they relate to each other. Many people are familiar with the taxonomy of species, from the top level “kingdom” (plants, animals, prokaryote, fungi, etc) down to “genus” and finally individual “species”.

*Example:*

*In the case of times and dates we have a common language (taxonomy) that identifies **year, month, day, hour, minute, and seconds** as the standard terms, with well-established hierarchical relationships between each of them. To uniquely identify a time and date we also need to specify a **time-zone**, and which **calendar** we are using – not everyone uses the western Gregorian calendar.*

- *There are additional terms in common use, such as the day of the week and the week-of-the-year, both of which can be calculated from the date and so don't add extra information but may be useful to store to make searches quicker, say if you wanted to search for content from “every Monday” for example.*
- *Other special event names, like “New Year's Eve” or “Midnight”, are usually derived from the date and/or time, and can act like shorthand.*

*There is a well established standard (ISO 8601) for representing a time and date in an unambiguous fashion, and computers cope very well with it. Bear in mind though that it has the Gregorian calendar specified in the standard, which only came into use in stages in various countries during the 16<sup>th</sup> century and later, so if you're not using that you'll have to flag it somewhere. Astronomers tend to avoid this issue by using “Julian day” (JD) numbers for dates, which is just a count of the number of days (whole and fractions) since noon 1 January, 4713 BC, and include the year zero which most calendars do not (where 1BC was followed by 1AD). Either, or neither, approach may work for you.*

### 2.3.6 Ontologies

Having specified a taxonomy, which provides the metadata content values, it is then useful to specify what each of the terms can actually mean and what values they can sensibly have. This is described as the **ontology** of a system. The term comes from philosophy into the nature of being, so web searches for ontologies often turn up rather diverse and interesting sources. However, in the metadata world, with a focus on information management, it can be broadly thought of as the definitive list of values or other characteristics something can have. A dictionary, or a vocabulary, is an ontology of the words of a language. A gazetteer is an ontology of placenames.

*Example:*

*Using the times and dates example from above, the date ontology says that months can be January, February, ..., December, or shortened to their first three letters, or they can be numbered 1-12, but no other text or numeric values are valid. Days can be any whole number 1-31 - with some constraints for shorter months, leap years, etc. Time has a similar ontology; hours are 0-24 (or 0-12 with an “am” or “pm” flag), minutes are 0-59, seconds are 0-59, fractions are permitted below seconds, and if the hour is 24 then it is only so for an instant, and is equivalent to the next day at 0 hour.*



When creating metadata records for research data it is very useful to put some effort into both the taxonomies and ontologies, to limit hard-to-read free-form data entry and to provide some basic sanity checking. For example, entries in fields that are classified as 'dates' can be quickly verified to ensure that they are both valid and not in the future. There are many taxonomies and ontologies already out there for most disciplines, and schemas usually specify which ones they require or recommend.

### 2.3.7 A changing world

One of the benefits of adopting specific schemas, and their associated taxonomies and ontologies, is that the information can be read almost unambiguously. There may be a time where a new standard, a new schema, etc. becomes relevant to you, or you want to link two differently-managed datasets, and metadata needs to be translated from one standard/schema to another. This is often called **cross-walking** or **mapping**, and computers are very good at it, but only if the information they receive is already well structured. Picking a standard and sticking closely to it early is very important.

## 2.4 Metadata Standards

---

### 2.4.1 Plan from the users perspective

Hopefully it is clear by now how important it is to adopt standards as far as you can for all of your metadata needs. You need to start with setting out a metadata plan based on your needs for users who will access and use the data, as well as the people who will help with managing and preserving the data. That will in turn identify the kind of metadata and content you absolutely need. Then you can select which schemas, and associated taxonomies and ontologies, hold those terms and are the most appropriate.

### 2.4.2 Finding and choosing a metadata standard

When you are searching you need to keep in mind that some schemas are very generic and others are very discipline specific, and all points in between. Some are targeted just at file formats. For example there are several schema standards for text documents, audio, images, and video files. By their nature these are either very generic descriptions (like title, author, and performer) or deal with provenance (like camera settings, image size, date, and location) and are rarely sufficient on their own. But because they have been around for a long time, they tend to be well settled and widely used, and so may get you a long way along.

As mentioned earlier, one of the most common generic schemas is called the Dublin Core, and it is the most widely adopted schema for descriptive metadata to date. It is extremely generic and easy to use—but is so generic that everyone can use it in quite different ways as the description becomes more specialised. So the benefits of using Dublin Core over having no metadata at all are immense, but once you start you will find it is rather limiting.

More discipline-specific schemas provide a richer and more-targeted structure and vocabulary, and require a deeper understanding of the project needs. Finding these can be as simple as searching for "*discipline* metadata schema", from a very high level to start with (e.g. "agriculture metadata schema") and then getting steadily more specific (e.g. "crop yield metadata schema"), but you then need to assess which one(s) suits your needs.

For certain applications there are discipline-standards bodies, industry groups, or other higher agencies like government departments and UN agencies, which recommend the use of particular schemas and standards. Astronomy, for example, has the International Virtual Observatory Alliance, the biodiversity community has Darwin Core, Agriculture has AgMES through the UN/FAO. The governments of Australia and New Zealand have formed ANZLIC to help set metadata standards for geospatial data



across their jurisdictions including the Australian Government Locator Service (AGLS) record keeping standard.

### 2.4.3 Adopt, Adapt, or Act?

Finding good metadata schemas to use is a large part of the battle, but is not the only part. As with all good information standards it is rare that a single standard is ever seen as good enough, and even at discipline-specialisation levels there are often competing or overlapping standards. Your choices often boil down to a simple set of actions.

- If there is just one obvious standard, and it meets your needs, use it.
- If there are several obvious standards that meet your needs, ask as many of your colleagues as you can (locally and internationally) which ones they use, and pick the most common one. Or if there is a government agency you work with, pick what they use. Keep in mind that well-structured metadata can be mapped or cross-walked from one standard to another if you have to.
- If there's a standard that is close to what you need, but is a bit short, it can often be extended. This requires indentifying the 'owners' of the standard and then working with them. The level of formality for any standard varies immensely, from formal international organisations down to a student in a lab somewhere. You may find others have found similar shortcoming and are already working on them.
- If there is absolutely no standard you can use, check again; it's a rare situation nowadays. If there is still no standard to be found, then you may have to develop one. This is not as scary as it sounds. By definition it becomes the standard for everyone else. However, it takes a fair bit of work and you should bring together as many interested people in your discipline as you can. Developing a standard is beyond the scope of this document, but help and good examples are out there.

## 2.5 Collecting and keeping Metadata

---

### 2.5.1 Collecting it

Having determined what metadata you need to collect, you then need to decide how to collect it, and what to do with it.

A lot of metadata is created for you during the collection process. Many scientific instruments generate metadata alongside the data itself. An obvious example is digital cameras, where a lot of provenance metadata is written at the same time as the photo is taken. Some cameras have GPS or other location information, as well as a clock and calendar, so the location, time and date can be automatically captured. Many digital video cameras and audio recorders do the same. Bigger instruments such as microscopes, particle accelerators and telescopes have extremely rich, and complex, metadata that is generated during an observation.

A good rule to follow is if you can collect metadata automatically, do so. It avoids data entry errors later on and reduces the workload. All you need to do is ensure the device is set up the right way before it is used. If you want to be really careful, you can also set up an automated process to sanity-check the metadata when the data comes in.

Other metadata, though, requires human participation to create, and the process to do that depends a lot on the tools you are using to collect the data. One example is the growing interest in electronic



notebooks, or digital lab books, to replace free-form paper notebooks. It is possible to build tools that present the user with a series of forms, matching the metadata fields and automatically enforcing the right ontologies and data structures for each metadata record. That work can happen as the data is being collected, which is preferable, or can be done later on when the researcher organises the data they have collected over some period of time, e.g. after a field trip.

## 2.5.2 Keeping it

You have metadata coming in, in a nicely organised and planned way, but then the question arises of where to put it. There are a range of options, and not all metadata may end up in the one place. However, searching and managing the metadata is a lot easier if you can pick a considered approach.

### 2.5.2.1 Files and folders

The first place that metadata can, and often does, go is inside the file with the data itself. There are many digital file formats that include a range of metadata fields, and some can be extended to hold almost anything. These include

- Text formats such as Word, HTML, PDF and others
- Image formats such as TIFF and JPEG
- Video formats like MPEG and AVI
- Audio formats like WAV
- Specialist discipline formats like FITS (Astronomy) and HDF (Remote sensing)
- Array data formats like NetCDF

A benefit of storing metadata inside the file is that it moves with the file; the **association** between the data and its metadata is easy to maintain. There are some downsides though. Not every metadata field you want may be able to be added. When you want to search for a particular file the computer has to open every single file for every single query, putting a lot of pressure on the system storing the data, especially as the number of files and the number of queries grows. Also, collection-level metadata is not easily managed. If you write a collection reference into each file and then decide to make a change to it, you have to touch every single affected file.

A slightly different model is to write the metadata into a separate, well-structured file, perhaps XML, and associate that with the data file. One common approach to association is to use the same filename stem, so that for example Image1.tiff is the image and Image1.xml is the metadata. This can improve the performance for searching and metadata modifications, but only slightly. The data is still on the same storage medium and you still have to open the same number of files to make queries or changes. It increases the risk when files are moved that the data and the metadata can get separated. It does, however, give you infinite flexibility for storing any and all kinds of metadata without restriction.

### 2.5.2.2 A more structured approach

Taking this approach a step further, aggregating the metadata in a single file, like a spreadsheet, gives you a lot more flexibility, since you can easily share the spreadsheet, do searches and changes within it and so on. In that case there is only one file that is touched for most queries and changes. But now the association between the metadata and the data has to be more carefully handled. The most common



approach is to store the filename and location for the data within the metadata record. Anytime the file moves, the appropriate field has to be changed, but it means you can find the data from the metadata record and vice-versa.

The best approach for large collections is, surprisingly, to increase the separation between the data and its metadata even further and use two quite different systems. Putting the metadata into a database optimises several things at the same time. It stores the “transactional” metadata into a system that is designed to handle transactions. Searches and record updates are what databases are designed to do. They can have infinitely flexible structures, and indeed metadata schemas are usually written as a database structure. Collection-level changes are also very easy to make, since a database will just flag items that belong to a certain collection and point them all to the collection-level record. Only one record in the database has to be changed for every item in that collection to be changed. The data can stay on a system that is good for storage, and the metadata can go to a system that is good for databases, and the data is only touched when it needs to be.

A word of caution: many databases can also hold the raw data itself, but this is rarely a good idea. It makes the database much larger and slower than it needs to be, and doing things like reading segments or slices out of the data is much easier and faster in a file than in a database object.

Now, designing and running a good database takes some practise. There are plenty of web-friendly and desktop-friendly tools to make it easier. There are desktop databases like *FileMaker* and *Access*, or larger web-manageable databases like *MySQL* with *phpMyAdmin* (amongst many others) that are in common use with lots of good documentation to get you started. These don't lend themselves as easily to very large-scale systems, with many records, many users and complex workflows for data and metadata to go in and out. That requires more high-quality databases, possibly up to commercial platforms which cost to buy and cost to maintain. Fortunately it is possible to start small and work up as you need to. Databases are fairly standardised and moving data from one system to another is usually relatively easy.

There are systems out there that combine databases with file storage systems and know about (some) metadata schemas. These are often called **repositories** and there are quite a few systems available online for free up to large commercial systems. The benefit of these systems is that they are designed from the ground up for the purpose of capturing data and its metadata, managing both, and being searchable and accessible with some level of user-friendliness. The downside can be that many are very generic in order to appeal to the widest possible audience, and may be too constraining for your needs, or may require some development effort to do exactly what you need, or how you want.

## 3. Problems with Metadata

---

### 3.1 It's not all good?

---

Every silver lining has a cloud, and metadata is no different.

- *Metadata is only as good as you make it.* While metadata is vital, it is only as good as the information stored within it. Data can become useless if it's associated with poor metadata, and the issue is possibly worse than if it has no metadata at all, because people are more inclined to trust data that appears to be well-described.
- *Managing metadata is expensive.* It takes time and effort to collect and manage metadata, so if it's not done automatically there is an additional cost to a project – which may not have been



budgeted.

- *The world can change under you.* While most people see scientific arguments at a very high level, there can be dissent at the very lowest level too. Classification and naming systems can and do change sometimes. Species of plants in one genus can suddenly find themselves in another, and people who have learned one name for plant suddenly can't find new data under the now-old name. Somebody has to keep an eye on the discipline.
- *All the world does not speak English.* Almost all metadata standards development, indeed most of information technology, is done and documented in English (and American English at that). Support for non-English languages is pretty thin, and support for non-western character sets is even thinner. If you want to store say Arabic or Chinese values in your metadata record you will face problems.
- *All metadata is not text.* An extension of the language problem is that most systems assume all metadata can be input with just a keyboard. This does not help with metadata forms such as musical scores, or images of physical objects. The most common approach is to treat scores and images as other data objects (with their own metadata needs) and then build a reference from say the audio-recording metadata record to the musical-score metadata record.
- *Managing associations takes effort.* Once you go down the path of separating the metadata from the data, you need to maintain the association between them. That means when files move, records have to be updated somewhere. There are technologies like *persistent identifiers* (handles, purl, Digital Object Identifiers, etc.) which can help immensely if set up correctly, but what they do is delegate the problem to another system that requires some maintenance.
- *Metadata owners may not be data owners.* As collaborations grow, and as data is made more accessible, it is quite possible to have shared responsibility for metadata and the data with which it is associated. In certain circumstances the ownership and responsibility may become split, and changes in one can affect the value of the other – for example maintaining a global list of data collections for your discipline is difficult when those collections can come and go.

All of these problems can be managed, if known about up front. It may take more development or collection effort, and thus costs, than were planned. It is prudent to consider the issues early during the planning phase and decide what to actually worry about.

## 3.2 Avoiding common mistakes

---

There are some traps to avoid, based on many people's experiences over many years. Keep in mind that the problems have been around since the earliest libraries, many millennia ago; these are not just digital-age problems. The traps include:

- *Not planning:* You need to know upfront why you are doing this work, who will benefit, and how. You can have the world's highest quality and quantity of metadata but if it's missing the one field your researcher is looking for, it's useless to them.
- *Planning too much:* Don't plan yourself into inaction. Taking concrete steps can help inform planning, and as mentioned, if you start carefully and methodically you can make changes and fix earlier work.
- *Not involving everyone:* Building any research data collection will have a lot of interested stakeholders. They need to be involved in the planning and deployment.
- *Picking tools too soon:* You need a plan to understand the requirements, and you need to understand the requirements, before you can test and pick tools. Many tools have interesting



limitations or designs, and you can lock yourself into something unfortunate.

- *Not automating everything:* People are both fallible and busy. If there is a way of automating metadata collection, then it is worthwhile in the long run to make the effort to do so. An extension of this is to have an automated review process, so that incoming metadata is sanity checked as appropriate.
- *Not following standards:* The whole premise of metadata is enabling people to find and use data. If your project is the only one to use a particular approach then it creates a barrier to others. Where there are multiple choices, try to identify the most popular one.
- *Being inflexible:* Once you've developed your plan and picked your standards you will have a lot of incentive to stick to it. However, the real world of information systems and standards and of research practices is very fluid and will probably always remain so. Be prepared for potentially major changes in the future. They may not happen for a long time, but they will. Fortunately if you follow good practices to start with, change can be managed quite well.
- *Not asking for help:* The development and application of metadata has been around for a long time, and a body of great knowledge has been built up. Much of that is available online now, along with a lot of other less-great knowledge. Nothing however beats talking with people with skills who can translate their knowledge to your problems, and who can identify other good resources.

## 4. Getting help with it all

---

Australia, like much of the rest of the world, has invested a lot in building skills in data management and making them available to the research community. While the web is a great source, it can be an overwhelming and rapidly evolving one, and getting people to help you directly can greatly accelerate your planning and deployment process.

The first line of (formal) support is available in most institutions where the libraries and archives groups often have staff dedicated to the role. Even where they don't, the skill-set in information management in libraries is closely aligned with research data management. Sometimes those skills are also set up in the central IT departments, and a few universities now have eResearch support groups, centres or programs.

After your institutional support, many universities in Australia are members or subscribers of state-based eResearch support organisations. Every state has a slightly different model for support, but contacting them will usually lead to help being made available. The appendix lists the current organisations.

At a broader level the Australian National Data Service (ANDS) has been established to provide institutional support nationally, with a range of materials (like this guide), regular activities and events, and especially people deployed across Australia. ANDS has a very broad, cross-disciplinary and cross-institutional view of the issues surrounding data management and metadata issues, and works with similar organisations in other countries. In many areas they work directly with institutions as well as the state-based eResearch bodies.

From a different perspective, the standards bodies and government agencies who work in your area of interest often have support mechanisms, or at least skilled people you can contact.



Finally, the best people to talk to are those in your discipline who have already done the same kind of thing. They may be in your institution or across the globe. They are ideally placed to answer the widest range of questions, and are often keen to share their insights – especially as they may learn something for their own project from you.

Written for ANDS by [Intersect](#), June 2011



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Australia License](#)



## 5. Appendix: State-based eResearch Organisations

---

The eResearch organisations in each state work together at some level, so contacting any one of them is a good start. Each comes from a different background, each has different institutional memberships, and each has a different engagement model with researchers. All however are keen to help, so do not hesitate to get in touch with any of them:

- Queensland: Queensland Cyber-Infrastructure Facility (QCIF) – [www.qcif.edu.au](http://www.qcif.edu.au)
- New South Wales: Intersect – [www.intersect.org.au](http://www.intersect.org.au)
- Victoria: has at least two
  - Victorian Partnership for Advanced Computing (VPAC) – [www.vpac.org](http://www.vpac.org)
  - Victorian eResearch Strategic Initiative – [www.versi.edu.au](http://www.versi.edu.au)
- Tasmania: Tasmanian Partnership for Advanced Computing (TPAC) – [www.tpac.org.au](http://www.tpac.org.au)
- South Australia: eResearch South Australia (eRSA) – [www.ersa.edu.au](http://www.ersa.edu.au)
- Western Australia: iVEC – [www.ivec.org](http://www.ivec.org)
- Northern Territory: currently does not have any similar organisations, but any of the others will be happy to assist.