# Digital Preservation—a Focus on Metadata

By: Emily Reed, Digital Records Archivist
Mercy Heritage Center, Sisters of Mercy of the Americas
ereed@sistersofmercy.org

Digital Preservation is the active management over time to ensure ongoing access.

**Some key words:**

**Electronic records**— Electronic records are records understood through machine use. For example: a digital file must be read through a software program and a vinyl record must be heard through a turntable. Electronic records can be analog information formats or digital information formats. Magnetic tape, old IBM computer punch cards, and vinyl records are examples of analog formatted assets. CDs, websites, and multimedia files are examples of digitally formatted assets. Electronic records are often called "machine-readable".

**Textual records**—Textual records, like paper and photographic prints, are not electronic records because they can be understood without machine help. Textual records are "human-readable".

**Digital records**—Digital records are digital information format, machine-readable records. Not all electronic records are digital, but all digital records are electronic.

**Born-digital records**—Born-digital records are a specific term referring to digital records that were created digitally, like a PowerPoint or an email. A VHS that was transferred into a digital file is not a born-digital record. Therefore all born-digital records are digital records, but not all digital records are born-digital.

**Audio-visual records**—Audio-visual records refer to any record that is not textual (paper), like audio, movies, and film. Audio-visual records can be digital, but they are not always.

Digital Preservation is often called "digital stewardship" because one must actively manage digital records. A person can place a piece of paper in an attic, forget about it, and probably come back to it in 300 years and it'll still be readable. One cannot place and forget a digital record, it will likely not be readable in 10 years.

Here are some reasons digital records need special attention:

1. Due to their immaterial nature, it is more difficult to ensure security and authenticity. You can tell if a painting is an original, but can you tell if a digital court document is an original?
2. Electronic records' physical containers are often not as durable as paper records. Hard drives and CD are only designed to last 10 years. And film will self-decay after 30 to 50 years.
3. There is no single system that will manage all digital records. So each digital preservation program is unique and designed that way by the archivist.
4. Accessibility of digital records is threatened by obsolescence. Even if a digital record's physical container is preserved (let's say a CD). If that CD holds files formatted in Windows Word 1992

and you can only find computers that can process 1995 files and higher, then those files are gone. Or if you have a Betamax tape and can’t find any players, then there’s no way you would be able to transfer that into a digital format.

5. Because digital files are composed of bit streams, any sort of change in the record will make it inaccessible. Bit streams are liable to become corrupt over time and use. If you rip a corner off a piece a paper you’ll still be able to read and understand the text. However, if you remove some bits from a digital file, the whole thing will be unreadable and inaccessible.

Digital preservation hopes to deal with the frailty of digital records in a couple different strategies:

1. **Data Redundancy**— making many copies. According to ISO standards, archival repositories should have three different geographical locations for digital storage. That means three copies. This is only helpful in the short term, because of obsolescence. Other plans must be in place so that file formats and media can be updated.
2. **Emulation**—Emulation involves using a program that imitates the original obsolete hardware or software to render the digital object (make it accessible). In emulation the original bit stream is saved and used. Emulation can be pricey, and you can never guarantee that you can create the perfect presentation of the original digital object. Emulation is meant to save the context and content of the digital object (which aligns more closely with archival theory).
3. **Migration**—Migration takes the original bit stream and changes it over into a new, current file format. You might lose the original formatting, but the content will still be accessible. This is the more popular, and in my opinion, the easier method.
4. **Create Metadata**—Gather as much information about the object as soon as possible when it is created. This can include basic descriptive information and information about the file format of the object. This method helps place the digital object in context—this is important because often the original context of a digital file must be abandoned. Metadata can also help track what was done to preserve the digital object through its life cycle, it can link digital objects with other objects, and help you locate the object.

## Metadata

As noted, metadata is a way to ensure the preservation of a digital object. If you think about it, you have interacted with metadata continuously over your life. A citation is metadata, a footnote is metadata, an inventory is metadata. **Metadata is data about data**. It is structured information that describes, explains, locates, and makes easier to retrieve, use or manage an information resource.

Importantly, metadata helps one locate and authenticate a resource (aka digital file) because it will hold information on the digital object that you can’t know by solely looking at it. In some ways, metadata is like a little finding aid for digital objects.

In the archives world, there are three types of metadata:

1. **Descriptive**—used to discover the object. Identifies things like: author, abstract, keywords, title
2. **Administrative**—used for managing and preserving objects in the repository. Identifies things like: file type, technical information, rights-management, and preservation information

3. **Structural**—used for storage of objects in the repository and for presentation. Identifies things like: number of pages, formation of chapters, how often a resource was printed (was it a serial?)

There are different ways to structure your metadata when you create it, this is called metadata schema (sometimes called schemes also). A metadata schema follows a specific metadata standard. There are a number of different standards designed to highlight different things. Some schema is very basic, some is more in-depth, some is designed as only structural metadata, some are better for audio records, etc. You will quickly learn it is hard to standardize anything in the archives world. The schemas have various fields you fill out to the best of your ability—sometimes you can't fill them all out if you don't have the information.

In its final form, metadata is written in an XML wrapper so that it can be linked, embedded, or encapsulated with its corresponding digital file and be hosted online. But that doesn't mean you can't initially write and store your metadata in a spreadsheet or text document until you get a generator, XML editor, or learn to code XML. Just be sure to store your text metadata document in the same folder as your digital file so that they are connected in at least this way.

The most widely accepted and used metadata in the archives field is DublinCore. Here are the DublinCore fields, they are designed to be limited, simple, but get the most essential information you can out of it.

- Contributor
- Coverage
- Creator
- Date
- Description
- Format
- Identifier
- Language
- Publisher
- Relation
- Rights
- Source
- Subject
- Title
- Type