Metadata Considerations for Deposits

Prepared by the DPN Preservation Metadata Standards Working Group
Completed July 27, 2017

The purpose of this document is to provide an overview of the types of metadata that may be encountered and/or considered while preparing deposits for the Digital Preservation Network (DPN). This resource addresses three areas: DPN-specific metadata, DuraCloud-specific metadata, core descriptive metadata records, and the significant properties of content.

DPN-Specific Metadata

All deposits into DPN require that digital materials are packaged according to DPN's BagIt specification. The two primary, bag-level metadata records required are the bag-info.txt and the dpn-info.txt files. All of the metadata described in this section of the document is created by DPN when it packages your information for preservation storage. The following is intended to inform you of the kinds of information being captured by DPN.

bag-info.txt
The fields in the **bag-info.txt** file are required to exist; however, the content of those fields may be left vacant. The information collected in this file is as follows:

Source-Organization
Organization-Address
Contact-Name
Contact-Phone
Contact-Email
Bagging-Date
Bag-Size
Bag-Group-Identifier
Bag-Count

Information regarding the content of these fields can be found in the BagIt File Packaging Format Specification Document (V. 0.97) .

dpn-info.txt
The fields in the **dpn-info.txt** file are specific to DPN deposits. The fields, along with the description of values, are as follows:

DPN-Object-ID: *Unique ID generated by Ingest Node.*
Local-ID: *Local identifier from originating repository.*
Ingest-Node-Name: *Name of the ingest node or source repository*

Ingest-Node-Address:
Ingest-Node-Contact-Name:
Ingest-Node-Contact-Email:
Version-Number: *Sequential positive integer*
First-Version-Object-ID: *Object-ID of the first version of the item*
Interpretive-Object-ID: *DPN UUID of Interpretive bag for this object*
Rights-Object-ID: *Reference to DPN and repository agreements*
Bag-Type: *data | interpretive | rights # Bags will be only one of these three types of objects.*

All fields must contain a value except for First-Version-Object-ID, Interpretive-Object-ID ("rights" and "interpretive" only), and Rights-Object-ID ("rights" and "interpretive" only).

In addition to these metadata records, the DPN architecture has a documented PREMIS implementation, which enables people and systems to monitor assets at a technical level. DPN implements PREMIS through its auditing functions, which are documented here.

Optional Tag Directories & Bag Structure
Bags in DPN deposits typically follow the basic BagIt structure detailed in the specification:

```
<DPN-Object-ID>/
        |  bagit.txt
        |  manifest-sha256.txt
        |  bag-info.txt
        |  tagmanifest-sha256.txt
        \--- data/
                |  [payload files]
          \--- dpn-tags/
                |  dpn-info.txt
        \--- [optional tag directories]/
                |  [optional node tag files]
```

The specification accepts optional tag directories at the “top level” of the DPN SIP to contain specific types of metadata records. If you are creating bags for submission, you may consider this option when thinking about how to best preserve your locally-created metadata. Metadata files may also be included with the digital objects located in the “data” directory.

DuraCloud-Specific Metadata

For the most part, DuraCloud does not place restrictions on metadata; instead, they indicate that local policies should be used to define metadata approaches. The metadata captured/created by DuraCloud at the vault-level is primarily created to facilitate processing and application functions.

Each snapshot contains four DuraCloud-created files: two checksum files (md5, sha265), a content properties file (that stores value pairs associated with individual files; value pairs can be entered in the DurAdmin interface), and a collection-snapshot file that contains the user-entered description of the snapshot along with other technical and administrative metadata.

When using DuraCloud, you will not be providing bagged data; bagging happens within the system. To indicate a certain file being a metadata file for the snapshot, you might use the DurAdmin interface to associate a value pair with that specific file. You might also indicate the name of this file in the description you enter prior to submitting your deposit.

Core Descriptive Metadata Records

The DPN Preservation Metadata Standards Working Group has examined minimal metadata records from a wide variety of member institutions to find the common metadata schemas in use by members. What has resulted is the "core record," or the minimum level of information needed in order to understand digital assets at a later date. We recommend that you select a well-documented metadata schema that is widely adopted and both human and machine-readable. Below are the recommended fields, along with definitions of each field. Additionally, mappings to a few commonly used schemas are provided.

Recommended Core Record

Field Name	Field Definition	Simple DC	Qualified DC	MODS	VRA Core	PB Core	EAD
Title	The name of the resource being described	title	title	titleinfo/title	title	pbcoretitle	<unittitle> <title> <titleproper> <subtitle>
Creator	The name of the person(s) or organization(s) with primary responsibility for creating the content	creator	creator	name [with optional attributes and child elements]	agent[@role] with agent[@name]	pbcreator/creator and role	<name> <origination> <persname> <origination> <corpname> <origination> <famname>
Date	Date information significant to an event in the lifecycle of the original content (creation, publication, issued date)	date	date dateSubmitted issued created	dateIssued; datecreated	date[@type]	pbcoreAssetDate	<publicationstmt><date> <unitdate>
Description	Summary description of the content, such as an abstract	description	description	abstract	description	pbcoreDescription	<abstract> <scopecontent> <notestmt><note> <physdesc>
Rights Statement	Information about rights held in and over the resource	rights	rights	accessCondition [@type="use and reproduction"]	rights	pbcoreRightsSummary/rightsSummary pbcoreRightsSummary/rightsLink	<userestrict>

Access Rights	Information about who can access the resource or an indication of its security status	rights	accessRights	accessCondition [@type="restriction on access"]	rights/note?	pbcoreRights Summary/rightsSummary pbcoreRights Summary/rightsLink	<accessrestrict>
Identifier	Unique identifier for a digital object (either a local identifier from your organization or a formal standard identifier issued and maintained by an external organization)	identifier	identifier	identifier[@type="local"]	location[@type="repository"]/refid[@type]	pbcoreIdentifier/source	<unitid> <eadid>
Format (original)	Format of the original item represented in the digital surrogate	type	type	typeOfResource	worktype	pbcoreAssetType	<physdescstructured> <unittype>
Format (digital)	Format of the digital file or digital surrogate	format	format	internetMediaType		pbcoreinstantiationDigital	

Significant Properties

In order for digital files to be usable and accessible in the long-term, it is important to recognize the importance of significant properties and to ensure that the properties of your digital materials are being documented in some form. Often, this metadata is created through automation tools such as FITS, or programs like Exiftool, MediaInfo, and FFmpeg. Currently, some member institutions are producing FITS files and including them with their DPN deposit, either as a separate xml file, or as an extension to a MODS file.

Below we have listed several content types, along with examples of common significant properties. This list is not extensive; we recommend using software to capture this information about your files.

General Significant Properties

FileSize (of object, not original)
Format
MIMETYPE
PUID
CreatingApplicationName
CreatingApplicationVersion

Video Significant Properties

Duration
frameRate
Color Space

Audio Significant Properties

Duration
channels
bitDepth

Image Significant Properties

colorSpace
imageHeight
imageWidth
byteOrder

Text Significant Properties

charset
language

Other Considerations

While we have covered the basics in this document, metadata beyond the scope of this resource may be required for certain types of materials. For instance, if you were preserving software, you might document the system environment it was intended for, the programs needed to render the software correctly, etc., for potential emulation scenarios in the future.

You will also want to keep in mind how much your metadata approach relies on external registries. You will want to ensure that any essential contextual information needed to interpret your metadata is included in/linked to your AIP.

Also, PREMIS event metadata is captured during the auditing process in DPN; however, individual actions to files by an information professional prior to ingest is not recorded in any way. This information may be desirable for some organizations.

Ultimately, the decision of how much or how little to include depends on your collections and your institution. If you have potentially challenging materials, please feel free to contact the working group at preservation-metadata@googlegroups.com. We would welcome the opportunity to help you troubleshoot and document this process to benefit the DPN community.