DCS, Johns Hopkins University

Item: Instantiating DCS ORE Resource Map Specification in BagIt Packages

Produced by Tim DiLauro, 8/9/2013

1 DCS-TXT-10 Instantiating DCS ORE Resource Map Specification in BagIt Packages

The BagIt specification 0.97[1], henceforth referred to as "BagIt', will be the first implementation for the packaging specification.
- [1] Last retrieved from: https://datatracker.ietf.org/doc/draft-kunze-bagit/ -- Expires April 6, 2013
- We can implement a later version of the BagIt specification, including a later draft, so long as it remains consistent with the discussion below. If not, then we will need to review and possibly modify this specification (currently DCS-TXT-10).
- The Library of Congress has developed a java library and command line tools to process bags. These are actively maintained, so it would probably be worthwhile to explore using these, rather than reinventing the wheel. These are available on Source Forge at http://sourceforge.net/projects/loc-xferutils/.

The BagIt directory structures will be serialized as either TAR, TAR.GZ, or ZIP.

All files to be deposited should be included in the payload directory:
- <bagdir>/data

The algorithm used for the payload and tag file manifests ({tag}manifest-<algortihm>.txt) will be MD5. These manifests contain a complete list of the payload and most of the BagIt tag files, respectively, along with their checksums. Since MD5 will be employed, the files will be named:
- manifest-md5.txt
- tagmanifest-md5.txt

The BagIt specification provides for a number of standard, but optional fields in the bag-info.txt file. The following should be provided:
- Contact-Name: Name of a contact person for this bag.
- Contact-Phone: Phone number for bag contact.
- Contact-Email: Email address for bag contact.
- [NB: The contact information does not need to match contact information in the DCS.]
- External-Identifier: Sender-provided identifier of the package. There should be a one-to-one mapping between the External-Identifier and the Bag-Group-Identifier.
- Bag-Count: Sequence number of a given bag within a bag group. The format is "<n> of <m>", where <n> is the sequence number within the group and <m> is the total number of bags in the group. If the latter is not known, the value of <m> should be '?'. "Bag-Count: 1 of 1" should be specified if the entire package will be captured in a single bag. This mechanism allows a package to be split across multiple bags.
- Bag-Group-Identifier: A sender-provided identifier for a bag group. For DCS packages, the bag group identifier should uniquely identify a package and when combined with the sequence number ('<n>') from the Bag-Count should uniquely identify a bag.
Some tools (e.g., the Library of Congress BagIt libraries and command-line tools) can generate the

following information automatically for certain operations:
- Bag-Size: The size or approximate size of the bag, intended primarily for human consumption.
- Payload-Oxum: The "octetstream sum" of the payload directory, in the format "<octetcount>.<streamcount>" (number-of-bytes.number-of-files).

Additional standard fields are available and additional fields can be added as needed.

The BagIt specification provides tremendous flexibility. In order to reasonably take advantage of the format, it is important to constrain the possible organization and encoding of the information contained in the package. We call this set of constraints a "profile" of the specification. In order for a package to be compliant, it must conform to both its packaging format (BagIt, in this case) specification and the constraints of any applicable profiles.

Unfortunately, the concept of a profile is not part of the BagIt specification; but as mentioned earlier, the spec provides enough flexibility to add one.

Fortunately, there is ongoing work to create tools to support documenting and enforcing the constraints of profiles. We are working with the team developing these to track and influence their work. We are also working with a subgroup of the technical team of the DPN.org project in order to align our work with theirs.

A DCS package can be identified by the following property in bag-info.txt:
- BagIt-Profile-Identifier: <profile-URI>
The profile-URI is meant to be opaque, but would allow more than one profile to be supported over time. The DCS instance would need to know the profile-URI to properly parse the bag(s).
For example:
- BagIt-Profile-Identifier: http://dataconservancy.org/formats/data-conservancy-pkg-X5

DCS packages that include an ORE resource map (ReM), will also contain the following property:
- PKG-BAG-DIR: A string consisting of a valid Unix directory name that, if specified, will be treated as the top-level bag directory name
- PKG-ORE-REM: A URL pointing to a serialization of the package resource map
-- If it is a "file:" scheme URL, no host should be specified and the path will begin with the name of the bag directory, followed by a forward slash, followed by the bag-relative path to the ReM
--- Format: file-url = "file:///" bag-directory-name "/" path-to-file-within-bag
--- In a bag named "ELOKA002.bag", the entry might look like this:
PKG-ORE-REM: file:///ELOKA002.bag/ORE-REM/8CF61A5C-1EB1-4ED3-9AC6-C072CFA2471C-ReM.xml
-- If a package consists of more than one bag, each bag should point to the same package resource map
--- For example, if the first bag of a 10-bag package contains the package ReM, the entry in the third bag might look like this:
PKG-ORE-REM: file:///bag-01-of-10/ORE-REM/94A5657E-DAF6-44C2-B32F-50EB6C9A3761-ReM.xml

[There are a number of ways we could handle packages for which no resource map is provided. We should discuss how we'd like to handle these "relationship-less" packages..]

Unless the resource maps were part of existing data (and thus, part of the package payload), they should not be placed in the package payload. The preferred location for these non-payload ReMs will be the ORE-REM directory in the package base directory.

ID	DCS-TXT-10
Heading	2.2.1.13.4
Item Type	Text
Project	DCS
Name	Instantiating DCS ORE Resource Map Specification in BagIt Packages
Global ID	5497