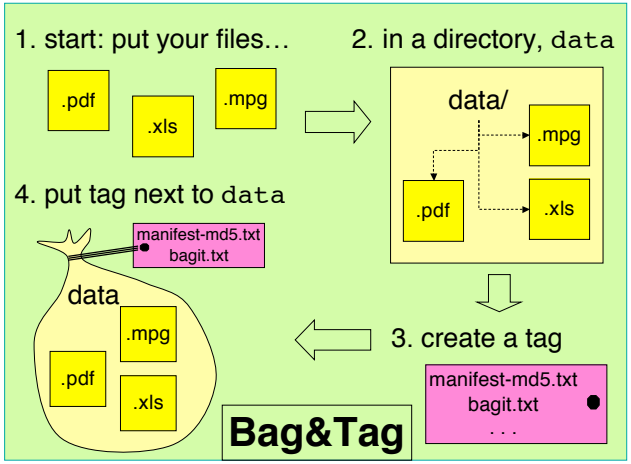
The BagIt file package format

John Kunze and Stephen Abrams, California Digital Library (CDL)

Replicas for a rainy day

- Need a data package format that can carry *any* data
- Suitable for disk-based and network-based transfer
- Main purpose is to move data from one digital library or archive to another for safe-keeping
- Not important whether the receiver provide access to or even understand meaning of the sender's content

Fig. 1. Sleeping better at night is more likely for the archivist who has replicas safely tucked away at other memory organizations.



A BagIt bag directory listing

- A BagIt *bag* is a special directory (or Windows folder)
- The directory name is your choice, but inside the top level BagIt reserves names for required files:
  - `data` : a subdirectory where you can put anything
  - `bagit.txt` : a 2-line file declaring this is a bag
  - `manifest-md5.txt` : a list of files present

```
MyFirstBag/
|   manifest-md5.txt
|   bagit.txt
|   bag-info.txt
\---data/
    |   my.pdf
    |   my.xsl
    |   my.mpg
```

Fig. 2. Payload directory, `data`,holds anything you want. "Tag" files (purple) describe the bag itself, including list of files and checksums in the manifest (algorithms other than md5 possible) and an optional "bag-info.txt" metadata file.

Holey bags! for efficiency

- A common optional tag file lists "holes" to be filled
- Payload is incomplete until these files are fetched
- File `fetch.txt` lists URLs for receiver to grab
- Benefits include (network transfers only):
  - No need to stage extra data copy at either end
  - Cheap parallelism from fetching URLs in batches

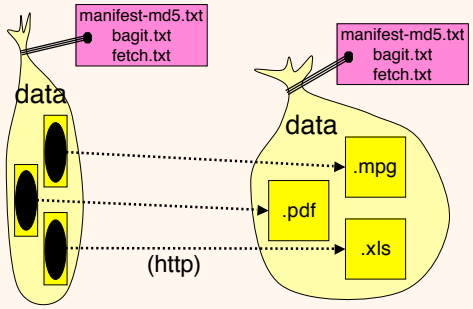
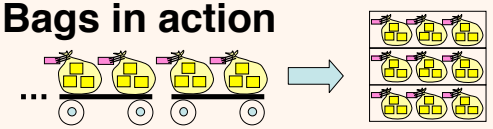


Fig. 3. On the left is a bag received with missing files, or "holes". The "fetch.txt" file lists URLs and corresponding payload filenames that the receiver must fetch before declaring the bag on the right complete.

Summary

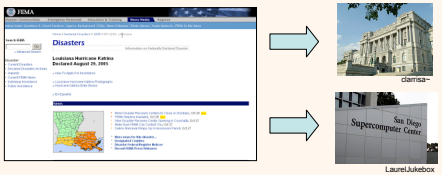
BagIt is a file hierarchy, suitable for disk- or network-based transfer, and maybe bag return, with just enough structure to safely enclose a manifest, checksums, tag info, and an arbitrary payload.

Bags in action



Library of Congress (LC) grantees send bags

- For example, bagged web crawls funded by LC are sent by CDL to LC, with replica sent to San Diego



BagIt credits and details

Authors from LC and CDL: Andy Boyko, John Kunze, Justin Littman, Liz Madden, Brian Vargas

Many thanks to: Stephen Abrams, Mike Ashenfelder, Scott Fisher, Erik Hetzner, Keith Johnson, David Loy, Tracy Seneca, Mark Phillips, Adam Turoff, Jim Tuttle

BagIt specification:

http://www.ietf.org/internet-drafts/draft-kunze-bagit-03.txt

http://www.cdlib.org/inside/diglib/bagit/bagitspec.html

BagIt was informed by

- Enclose-and-Deposit method, Tabata & Sugimoto,
- LC's eDeposit Pilot and NDIIPP AIHT
- ARC/WARC aggregate file format

For further information

Please contact *jak@ucop.edu* or *stephen.abrams@ucop.edu*

For information on CDL's Preservation Program, see

http://www.cdlib.org/programs/digital_preservation.html