Content Transfer: Getting Data Moved Around the Network

Converting Hub and Spoke Preservation Packages to Bagit

Tom Habing, Bill Ingram, & Robert Manaster

University of Illinois Urbana-Champaign

thabing; wingram2; manaster@illinois.edu

Content Transfer:
Getting Data Moved Around the Network

An individual **Hub and Spoke** preservation package consists of a directory containing **content files** and **METS metadata files**.

```
<package directory>/
    mastermets.xml
    echodepmets_0.xml
    [additional echodepmets_n.xml files]
    [one or more content files]
```

Content Transfer:
Getting Data Moved Around the Network

We started out with about **5000** Hub and Spoke preservation packages.

Individual packages ranged in size from **3 KB** to **1.6 GB**. (Although, most were under **100 MB**.)

Packages also varied in the number of content files, ranging from just a few to well over a hundred content files.

Content Transfer:
Getting Data Moved Around the Network

For each package, our tool copied all the files into a zip stream—calculating the **CHECKSUM** values along the way.

```
FileOutputStream fos = new FileOutputStream(zipFile);
ZipOutputStream zos = new ZipOutputStream(fos);
CheckedOutputStream chkStrm =
                      new CheckedOutputStream(zos, summer);
 while((len = fis.read(buf)) > 0) {
        chkStrm.write(buf, 0, len);
        packagesize += len;
}
zos.closeEntry();
```

Content Transfer:
Getting Data Moved Around the Network

For each file in the preservation package, the **FILENAME** and **CHECKSUM** values are used to write **manifest-sha1.txt**.

```
manifest.append(summer.getHexEncodedSHA1()
        + "  "
        + "data"
        + File.separator
        + file.getName()
        + System.getProperty("line.separator");
```

Content Transfer:
Getting Data Moved Around the Network

After all the files had been added to the zip stream, **bagit.txt**, **manifest-sha1.txt**, and **package-info.txt** were added as well.

The zip stream was then serialized into a zip file according to the **Bagit spec***.

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

Content Transfer: Getting Data Moved Around the Network

Once all the packages had been converted, we were left with a directory containing **5000 bags**.

Finally, we converted that directory of bags into a huge ***masterbag***, using the same process.

Content Transfer:
Getting Data Moved Around the Network

It took several hours to process all **5000** packages.

The result was a **16 GB *masterbag***, which we then put up on a web server for LC to retrieve.

We also created a **CHECKSUM** value for the ***masterbag*** to ensure transmission accuracy.

Content Transfer:
Getting Data Moved Around the Network

Questions?