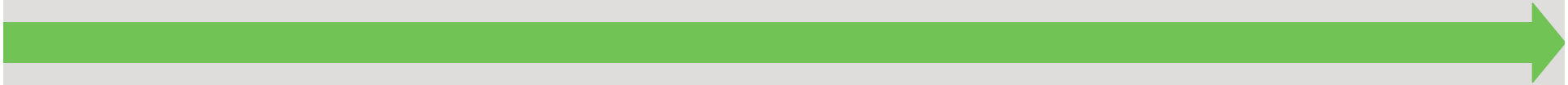
Content Transfer:
Getting Data Moved Around the Network

(Some Days) It is Better to Give than to Receive:
Coordinating File Transfers at the
Library of Congress

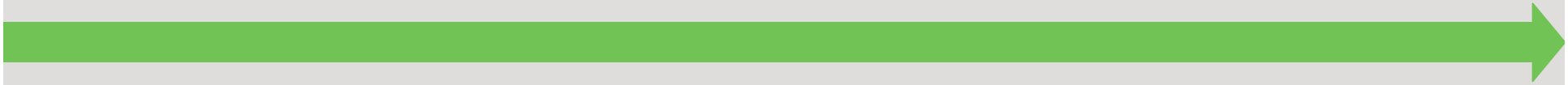
Leslie Johnston
Library of Congress
Office of Strategic Initiatives

DLF Fall Forum 2008



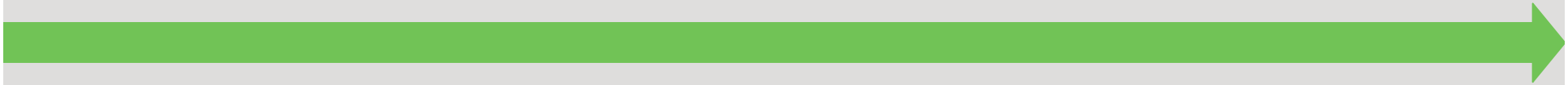
Content Transfer:
Getting Data Moved Around the Network

- Experiences:
  - So far during 2008 the Library of Congress has received 30 Tb from NDIIPP partners and 20 Tb in web crawls from the Internet Archive.
  - Hard drives and network transfers (rsync and ftp), ranging from 100 Gb to over 2 Tb in a single data package.
  - All forms of content, some to be dark archived for preservation, and some to be made publicly available.



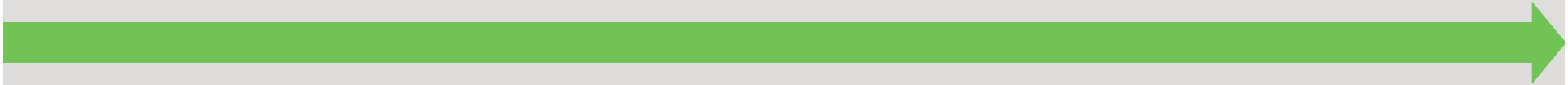
Content Transfer:
Getting Data Moved Around the Network

- In addition to supporting the preservation efforts of the NDIIPP partners, the transfer of partner content has informed the Library's own preservation efforts, as we begin to understand what we need to know about files and what events in their life cycle we need to record and track.
- The transfer of content is also informing work in developing tools and workflows for the transport and inventory of files for partners and in the Library's own environment.



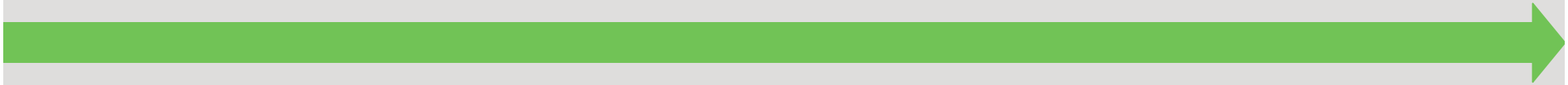
Content Transfer: Getting Data Moved Around the Network

- BagIt: A Packaging Standard for File Transfers
  - Motivating use cases:
    - Transfer of content internally and between preservation partners
    - Long-term storage of content
  - Needs:
    - Minimally self-identifying and self-describing packages
    - Support for error detection and transfer optimization
  - Characteristics:
    - Low overhead
    - Content-type agnostic
    - Supported by off-the-shelf tools (e.g., MD5Deep)



Content Transfer: Getting Data Moved Around the Network

- BagIt
  - Informed by:
    - LC's eDeposit Pilot Project
    - NDIIPP Archive and Ingest Handling Test (AIHT) and ongoing partner transfers
    - Tabata et al., "Enclose-and-Deposit Method," *IWAW '05*
  - Documented at:
- www.ietf.org/internet-drafts/draft-kunze-bagit-01.txt
- www.digitalpreservation.gov/library/resources/tools/docs/bagitspec.pdf



Content Transfer:
Getting Data Moved Around the Network

- BagIt Structure:

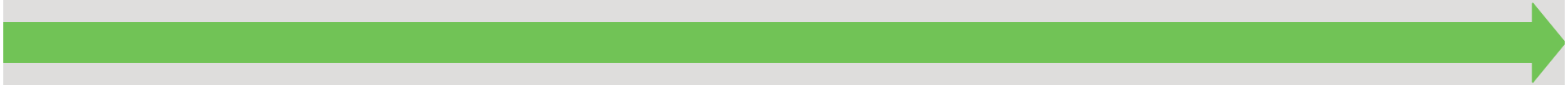
```
<bag_dir>/
bagit.txt
bag-info.txt
manifest-<algorithm>.txt
[optional additional tag files]
data/
[content file hierarchy]
```

- Bag Parts:
  - bagit.txt: Bag signature
  - manifest-<algorithm>.txt: List of content files and fixities

    Example, manifest-md5.txt:

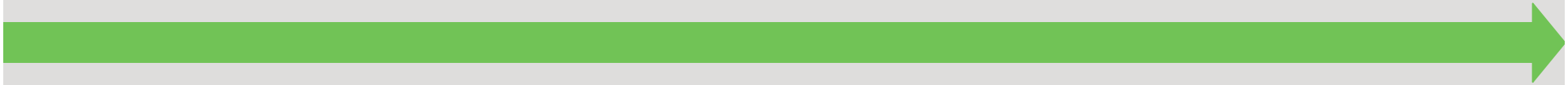
```
49afbd86a1ca9f34b677a3f09655eae9 data/27613-h/images/q172.png
408ad21d50cef31da4df6d9ed81b01a7 data/27613-h/images/q172.txt
```

  - bag-info.txt: Bag contents metadata (optional)
  - fetch.txt: Bag contents included by reference (optional)



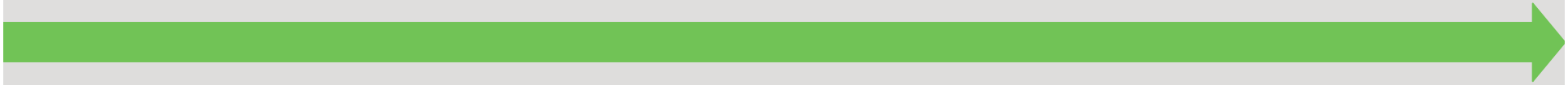
Content Transfer:
Getting Data Moved Around the Network

- Related Tool Development:
  - Parallel Retriever script
    - Multi-threaded package transfer, currently supports rsync and ftp
  - Bag Validator script
    - Validates the package against the Bag format
  - VerifyIt script
    - Verifies package checksums
  - BIL Java Library
    - BagIt Library for application and command line tool development
  - Project-based web work flow tools
  - Bagger Java WebStart Desktop app
    - Graphical desktop tool to create/update/validate Bags
  - Deposit Web App for partners to register transfers, whether shipping a hard drive or initiating a network transfer.
  - Inventory tool to record files and life cycle events.



Content Transfer: Getting Data Moved Around the Network

- What's Next?
  - Continue to receive files from partners and refine work flows based on practical experience.
  - Development of graphical workflow tools for all active projects
    - jBPM is used to instantiate the workflows, with a UI developed with the Spring framework
  - Complete testing of all the tools.
  - Put all tools and services into full production spring/summer 2009.
  - Release of scripts, etc. at Sourceforge



Content Transfer:
Getting Data Moved Around the Network

Questions?

Leslie Johnston, lesliej@loc.gov