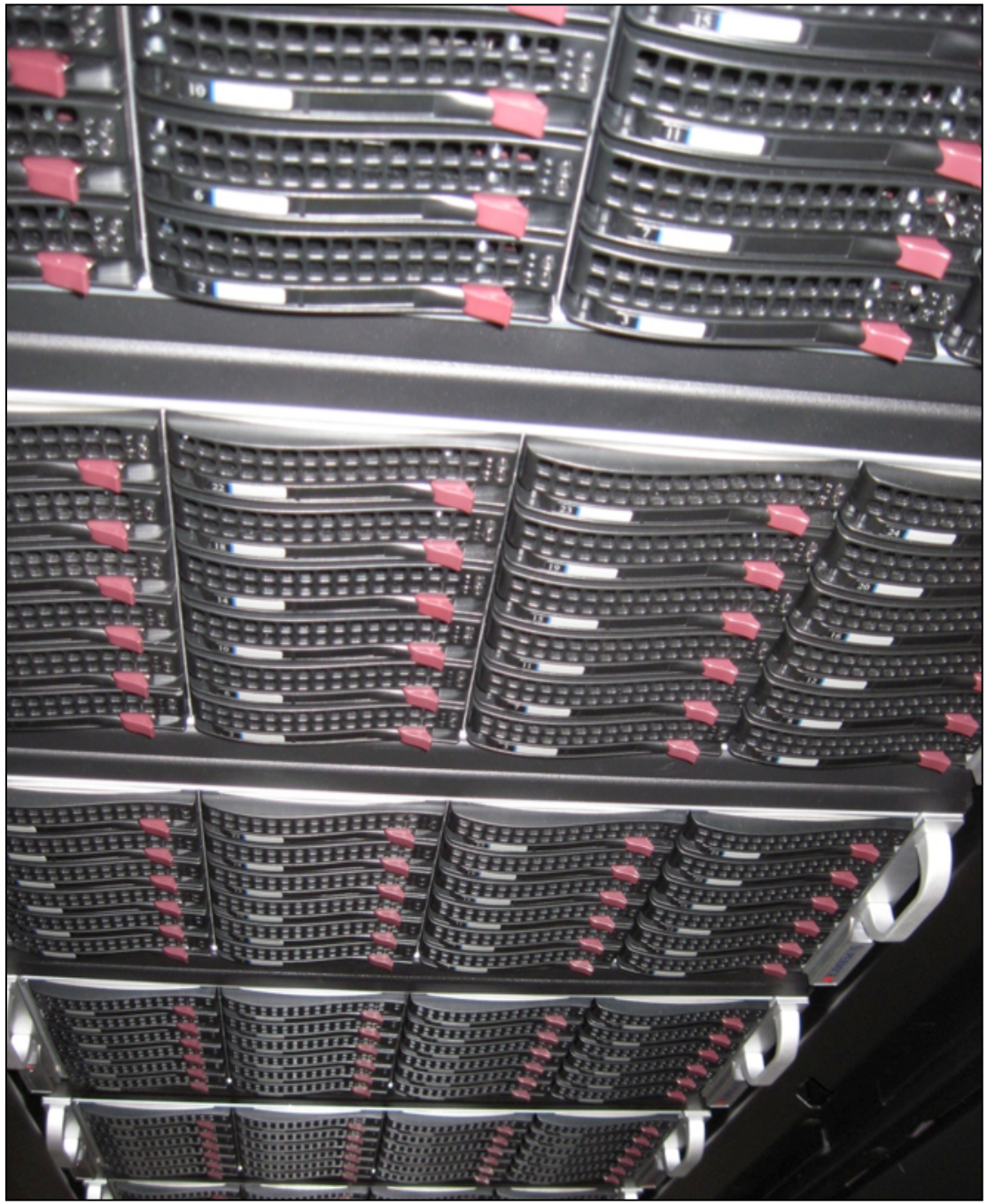


BHL Cluster

Clustered and distributed storage

with
commodity hardware
and open source software

Phil Cryer
BHL Developer, Systems Analyst
BHL Europe Technical Board Meeting
25-27 August 2010, NHM London

BHL data, on our cluster

BHL Cluster

BHL's first cluster in Woods Hole

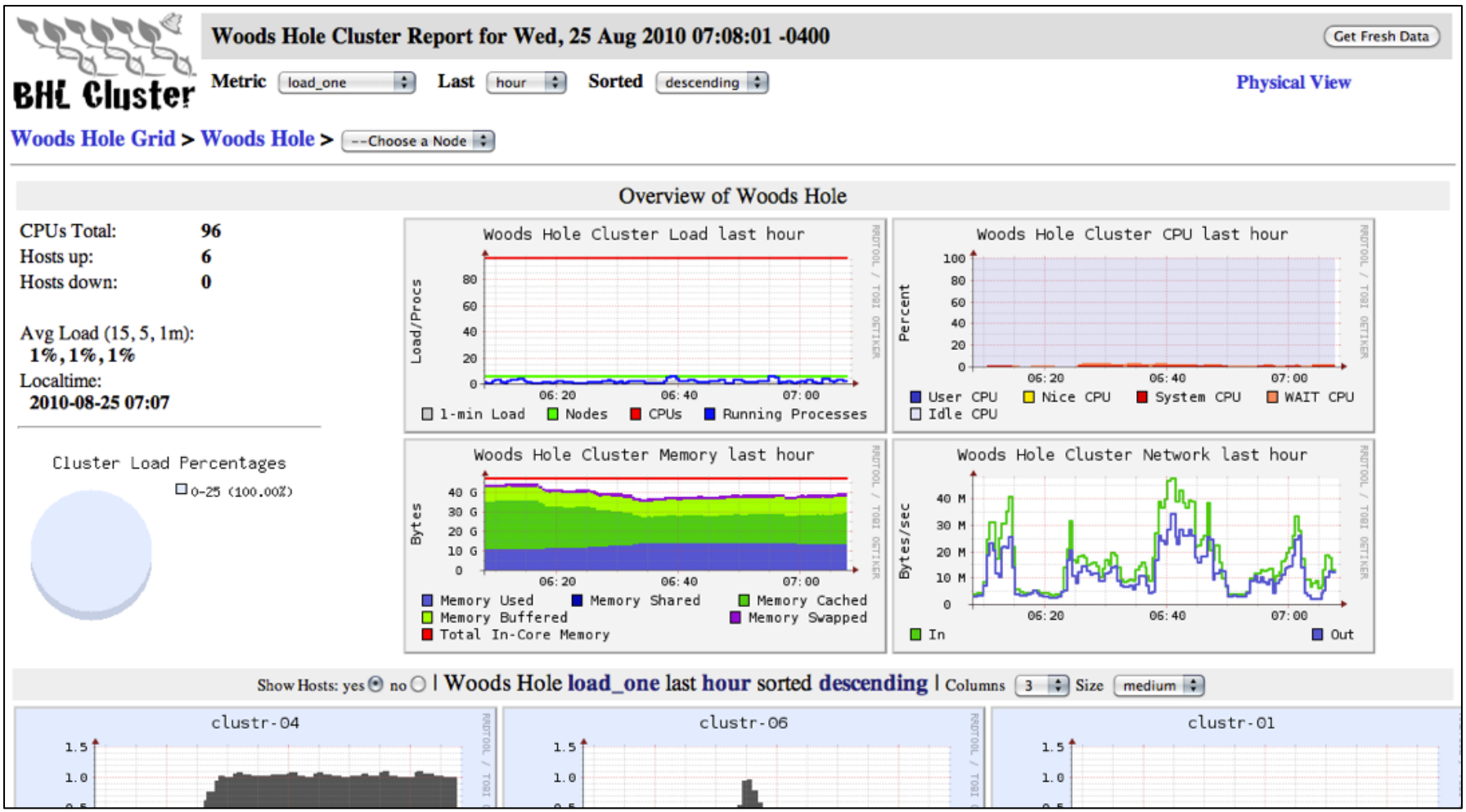
- **Hardware -** commodity servers
  - (6) six 4U sized cabinets
  - (24) twenty-four 1.5TB hard drives in each cabinet

BHL data, on our cluster

BHL's first cluster in Woods Hole

- **Hardware** - commodity servers
  - (6) six 4U sized cabinets
  - (24) twenty-four 1.5TB hard drives in each cabinet
- **Software** - open source software
  - operating system is Debian GNU/Linux (squeeze)
  - filesystem - ext4
    - supports filesystems up to 1 EB (1000 PB) and max file size of 16 TB
  - clustered file system - GlusterFS (3.0.4)
    - all drives run in a networked/RAID1 setup
    - all files are replicated and redundantly copied across the cluster
    - New: Acquia is using GlusterFS for their Drupal SaaS implementation
  - monitoring - Monit, Ganglia for alerts and reporting

BHL data, on our cluster



http://whbhl01.ubio.org/ganglia

BHL data, on our cluster

BHL Cluster

BHL's first cluster in Woods Hole

- **Hardware** - commodity servers
  - (6) six 4U sized cabinets
  - (24) twenty-four 1.5TB hard drives in each cabinet
- **Software** - open source software
  - operating system is Debian GNU/Linux (squeeze)
  - filesystem - ext4
    - supports filesystems up to 1 EB (1000 PB) and max file size of 16 TB
  - clustered file system - GlusterFS (3.0.4)
    - all drives run in a networked/RAID1 setup
    - all files are replicated and redundantly copied across the cluster
    - New: Acquia is using GlusterFS for their Drupal SaaS implementation
  - monitoring - Monit, Ganglia for alerts and reporting
- **Capacity** - cluster has 97TB of replicated/distributed storage
  - currently using 66TB of data for 78492 books
  - a full record for a book can be 24MB - 3GB

Files from a record


```
# ls -lh /mnt/glusterfs/www/a/actasocietatissc26suom
total 649M
-rwxr-xr-x 1 www-data www-data  19M 2009-07-10 01:55 actasocietatissc26suom_abbyy.gz
-rwxr-xr-x 1 www-data www-data  28M 2009-07-10 06:53 actasocietatissc26suom_bw.pdf
-rwxr-xr-x 1 www-data www-data 1.3K 2009-06-12 10:21 actasocietatissc26suom_dc.xml
-rwxr-xr-x 1 www-data www-data  18M 2009-07-10 03:05 actasocietatissc26suom.djvu
-rwxr-xr-x 1 www-data www-data 1.3M 2009-07-10 06:54 actasocietatissc26suom_djvu.txt
-rwxr-xr-x 1 www-data www-data  14M 2009-07-10 02:08 actasocietatissc26suom_djvu.xml
-rwxr-xr-x 1 www-data www-data 4.4K 2009-12-14 04:42 actasocietatissc26suom_files.xml
-rwxr-xr-x 1 www-data www-data  20M 2009-07-09 18:57 actasocietatissc26suom_flippy.zip
-rwxr-xr-x 1 www-data www-data 285K 2009-07-09 18:52 actasocietatissc26suom.gif
-rwxr-xr-x 1 www-data www-data 193M 2009-07-09 18:51 actasocietatissc26suom_jp2.zip
-rwxr-xr-x 1 www-data www-data 5.7K 2009-06-12 10:21 actasocietatissc26suom_marc.xml
-rwxr-xr-x 1 www-data www-data 2.0K 2009-06-12 10:21 actasocietatissc26suom_meta.mrc
-rwxr-xr-x 1 www-data www-data  416 2009-06-12 10:21 actasocietatissc26suom_metasource.xml
-rwxr-xr-x 1 www-data www-data 2.2K 2009-12-01 12:20 actasocietatissc26suom_meta.xml
-rwxr-xr-x 1 www-data www-data 279K 2009-12-14 04:42 actasocietatissc26suom_names.xml
-rwxr-xr-x 1 www-data www-data 324M 2009-07-09 13:28 actasocietatissc26suom_orig_jp2.tar
-rwxr-xr-x 1 www-data www-data  34M 2009-07-10 04:35 actasocietatissc26suom.pdf
-rwxr-xr-x 1 www-data www-data 365K 2009-07-09 13:28 actasocietatissc26suom_scandata.xml
```

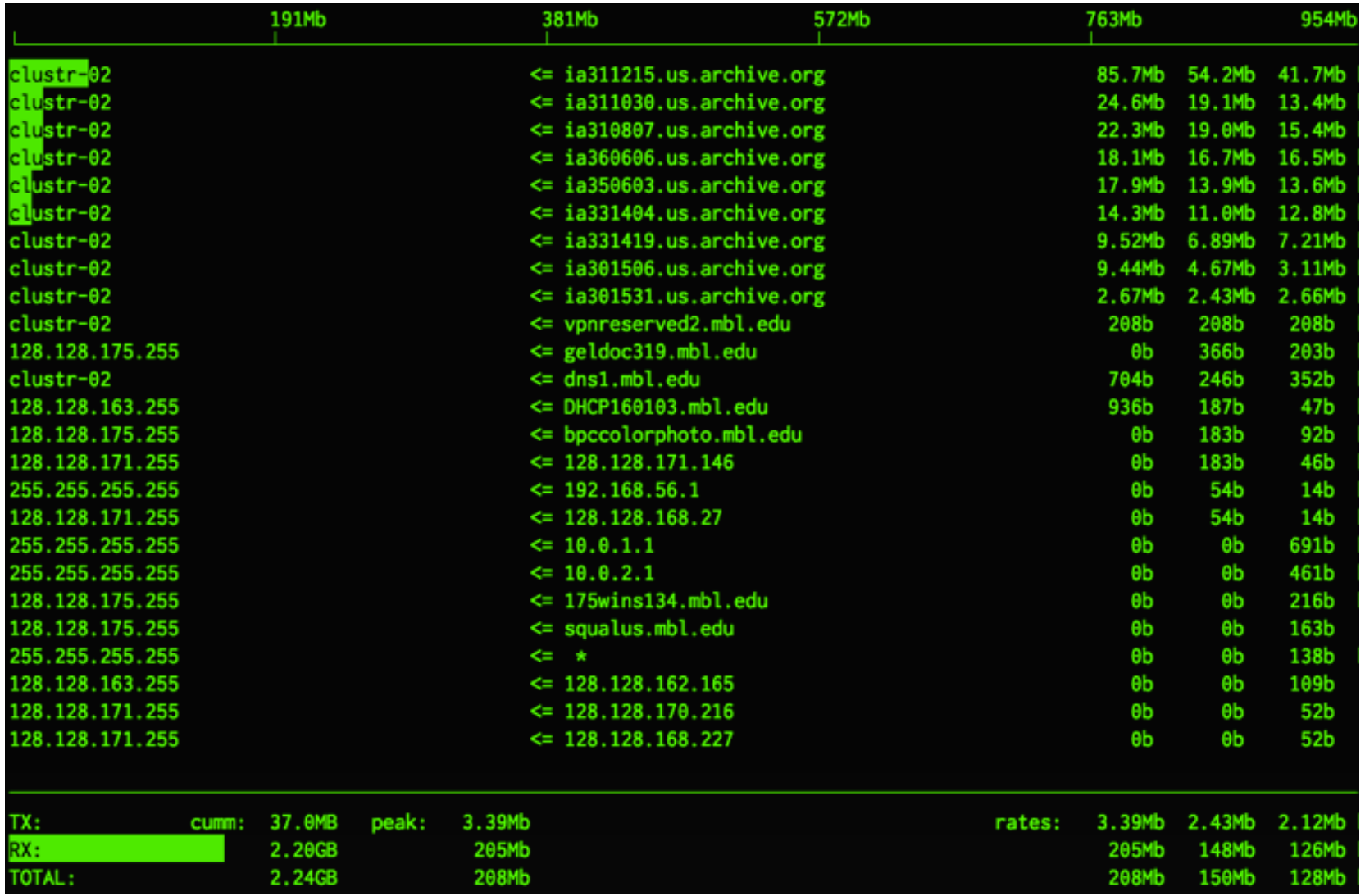
Initial file population

Populating a cluster with our data at the Internet Archive

- Looked at many options
  - ship a pre-populated server (Sun Thumper with 48TB capacity)
  - shipping individual external hard-drives
  - download the files on our own

Initial file population

BHL Cluster



Initial file population

Populating a cluster with our data at the Internet Archive

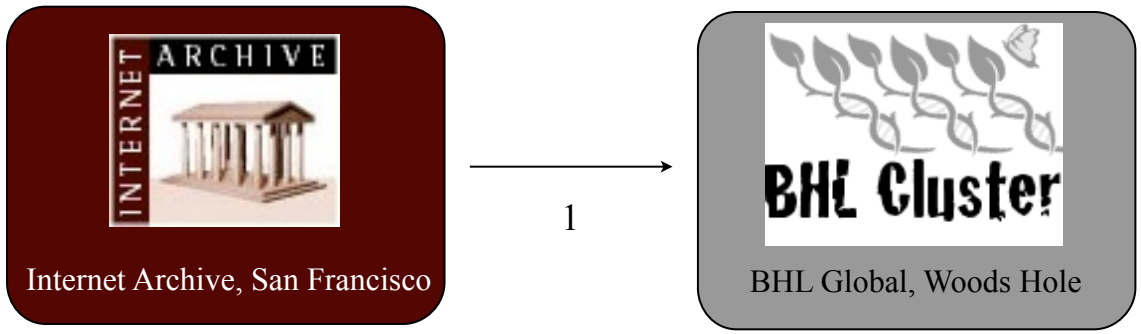
- Looked at many options
  - ship a pre-populated server (Sun Thumper with 48TB capacity)
  - shipping individual external hard-drives
  - download the files on our own

- Path of least resistance, we wrote a script and used the Internet2 connection at the Marine Biology Laboratory (Woods Hole) to download directly to the first cluster
  - knew it would take forever to download (but it took longer)
  - needed space to download files (cluster buildout)
  - networking issues in Woods Hole (overloaded local router)
  - file verification (checksums that don't...)

- **Lessons learned** - would we do it again? Probably not.

- Current propagation method
  - initial distribution - mailing external drives (1, 5)
  - syncing of the changes for future content (smaller bites)

Code: grabbyd

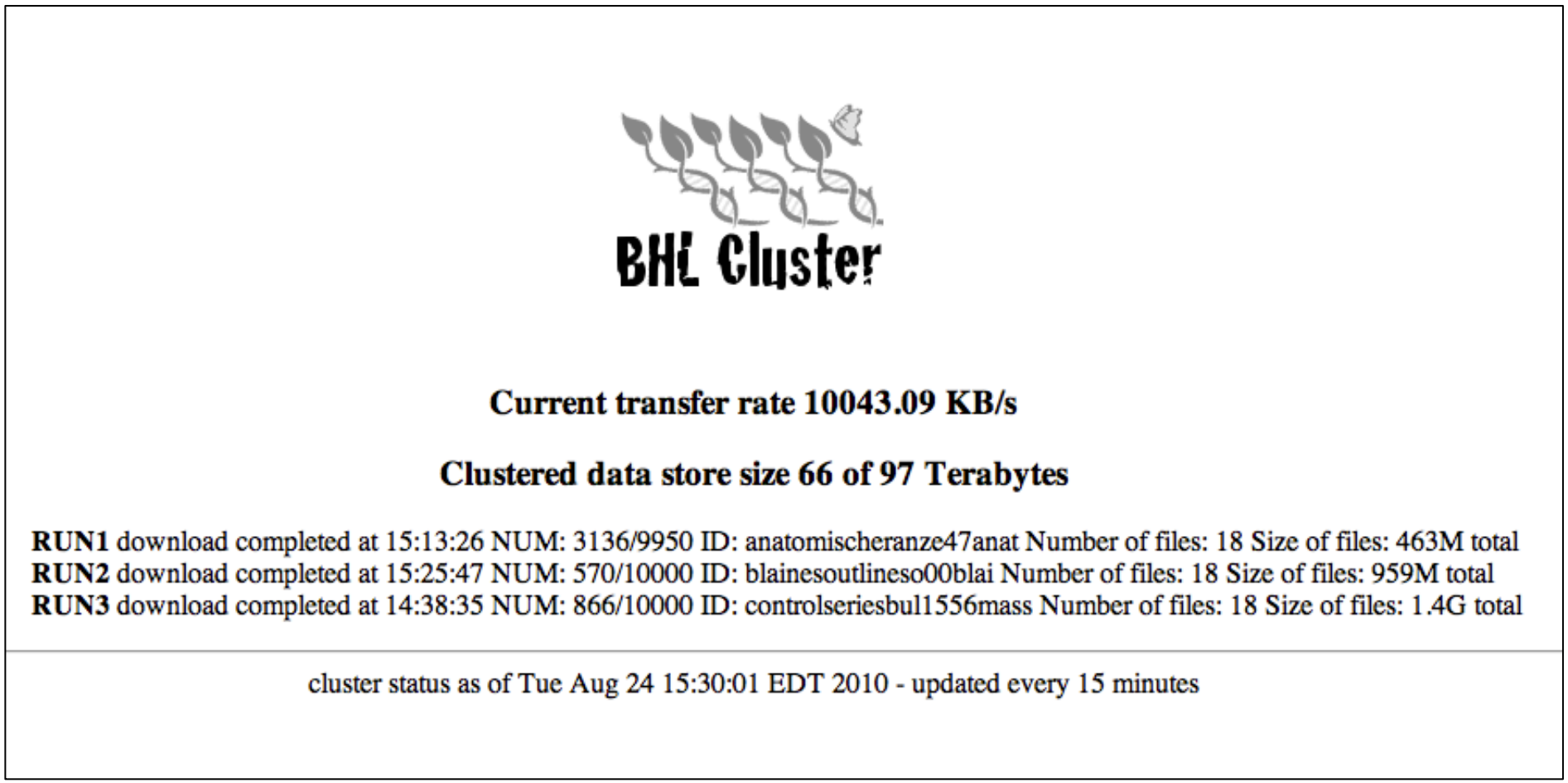


Automated process to continuously download the latest BHL data

- Uses **subversion** to get an updated list of new BHL content as IA identifiers
  http://code.google.com/p/bhl-bits/source/browse/#svn/trunk/iaidentifiers
- An enhanced version of the original download script to transfer the data
  - **grabbyd -** a script that parses the latest iaidentifiers list, determines the IDs of the new data and downloads the data to the cluster
  - Will provide detailed reporting with status pages and/or another method (webapp, email, RSS, XML, etc)

Code available (open sourced, BSD licensed):
[1] http://code.google.com/p/bhl-bits/source/browse/trunk/utilities/grabby/grabbyd

Code: grabbyd + reporting



http://cluster.biodiversitylibrary.org/

Replication|Replication

BHL Cluster

Why do we need replication?

- First BHL stored everything at the Internet Archive in San Francisco
  - no backup or safety net
  - limited in what we could do with, and serve, our data
- Now with our first BHL cluster, we gain
  - redundancy - will be able to serve from the cluster and fall back to IA if needed
  - analytics - the files are ‘local’ to parse through, discover new relationships
  - serving options - geo-location, eventually will be able to serve from closest server

Replication|Replication

Why do we need replication?

- First BHL stored everything at the Internet Archive in San Francisco
  - no backup or safety net
  - limited in what we could do with, and serve, our data
- Now with our first BHL cluster, we gain
  - redundancy - will be able to serve from the cluster and fall back to IA if needed
  - analytics - the files are 'local' to parse through, discover new relationships
  - serving options - geo-location, eventually will be able to serve from closest server
- Next - share the data with everyone
  - Europe
  - Australia
  - China
  - etc...
- Provide safe harbor
  - lots of copies...

Code: bhl-sync

Open source **Dropbox** model

- uses and implements many open source projects
  - **inotify -** a subsystem within the Linux kernel that extends the filesystem to notice changes to the filesystem and report them to applications (in the kernel since 2.6.13 (2005))
  - **lsyncd -** an open source project that provides a wrapper into inotify
  - **OpenSSH -** secure file transfer
  - **rsync -** long term, proven syncing subsystem

Code: bhl-sync

Open source **Dropbox** model

- uses and implements many open source projects
  - **inotify** - a subsystem within the Linux kernel that extends the filesystem to notice changes to the filesystem and report them to applications (in the kernel since 2.6.13 (2005))
  - **lsyncd** - an open source project that provides a wrapper into inotify
  - **OpenSSH** - secure file transfer
  - **rsync** - long term, proven syncing subsystem

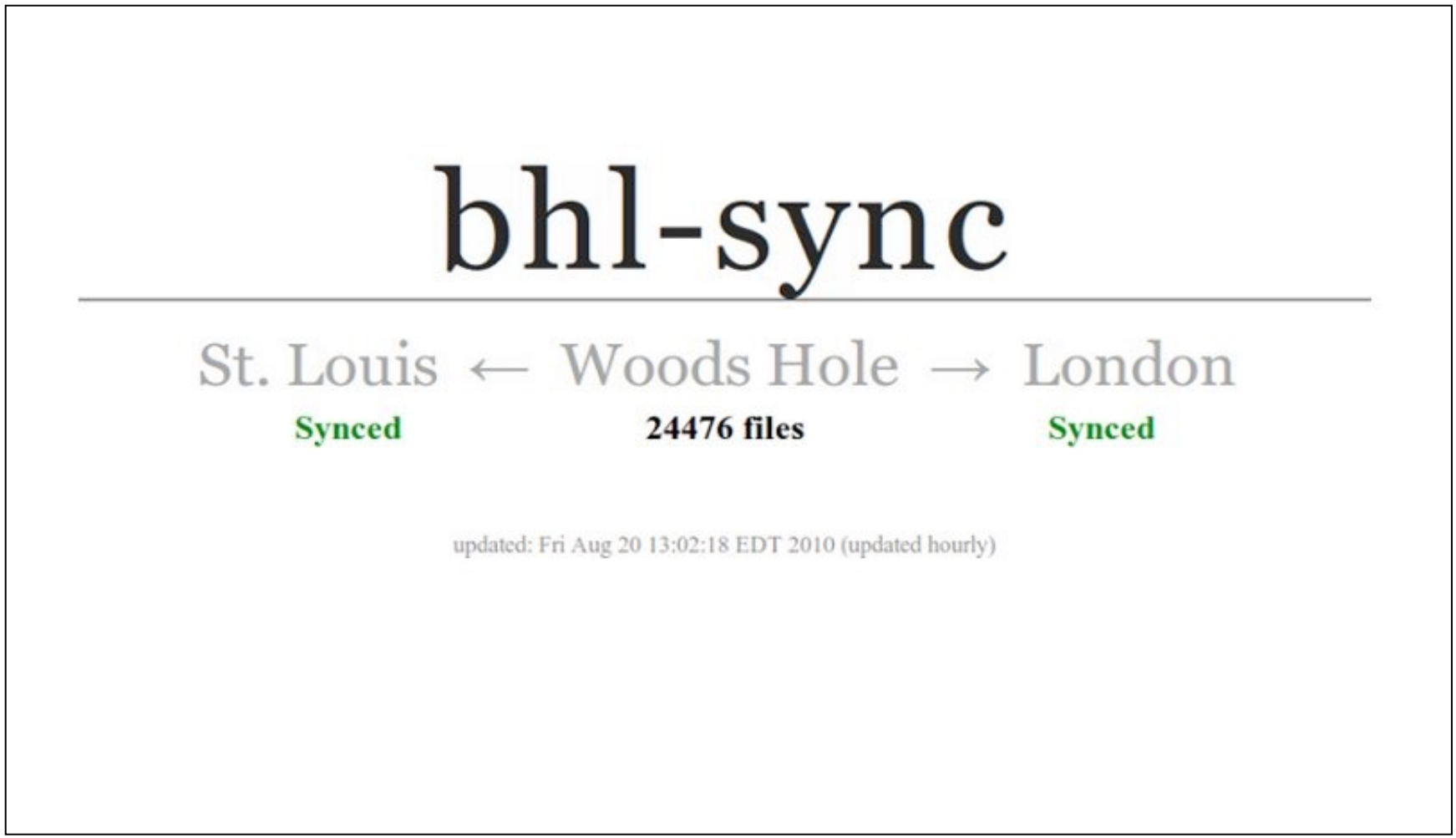
What does bhl-sync do?

- runs lsyncd as a daemon that notices kernel events and kicks off rync over OpenSSH to mirror data to designated remote servers
- the only requirement on the remote system is a secure login for a normal user (using a key based OpenSSH) keeping the process neutral and not requiring any other specific technologies (OS, applications, filesystem) on the remote system (cross-platform)
- want to mirror BHL? it's now possible (you just need a lot of storage)

Code available (open sourced, BSD licensed):
http://code.google.com/p/bhl-bits/source/browse/trunk/utilities/bhl-sync.sh

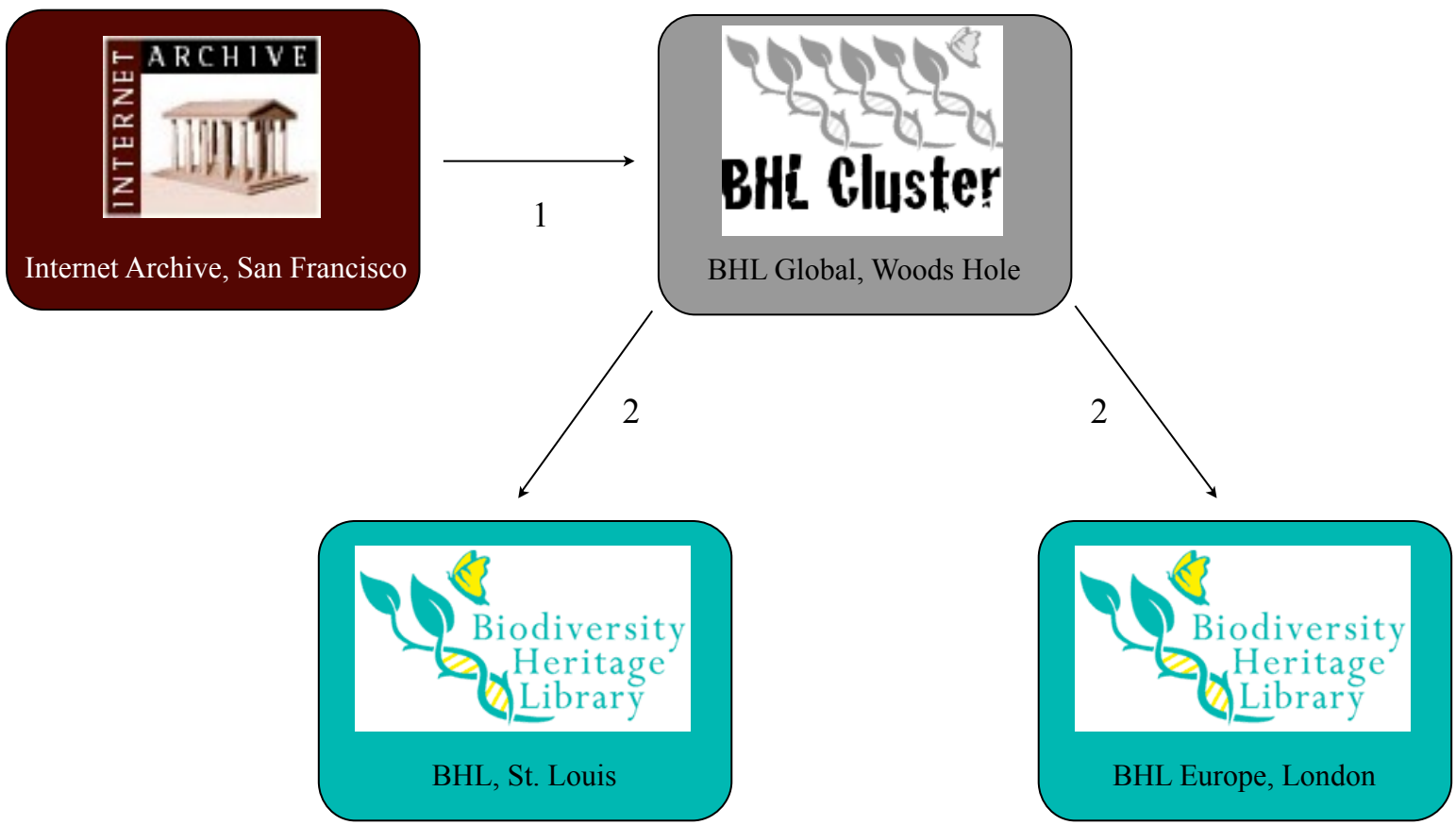
Code: bhl-sync + status

BHL Cluster



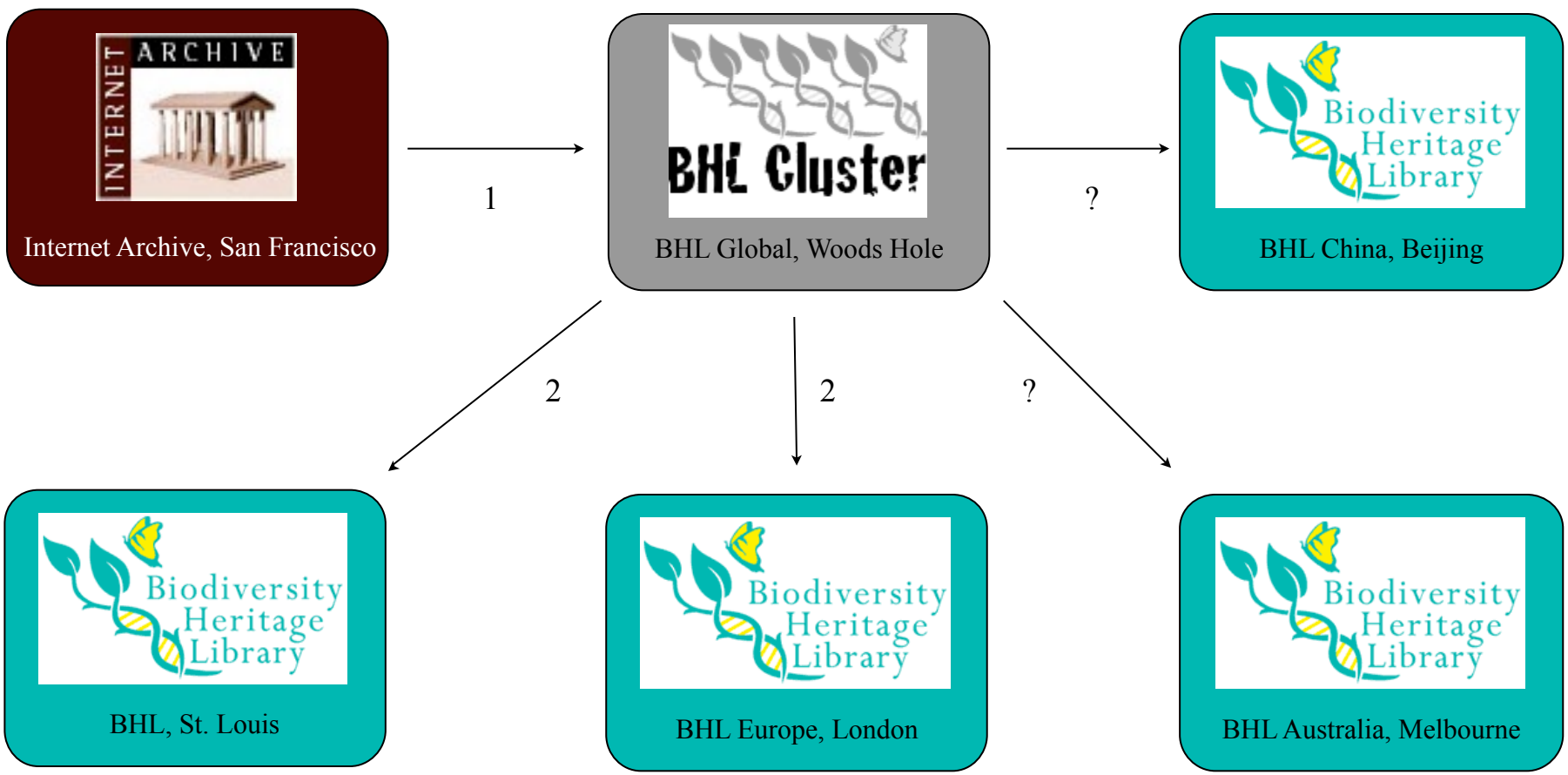
http://bit.ly/09-bhl-sync

BHL content distribution



Code available (open sourced, BSD licensed):
[1] http://code.google.com/p/bhl-bits/source/browse/trunk/utilities/grabby/grabbyd
[2] http://code.google.com/p/bhl-bits/source/browse/trunk/utilities/bhl-sync.sh

BHL content distribution



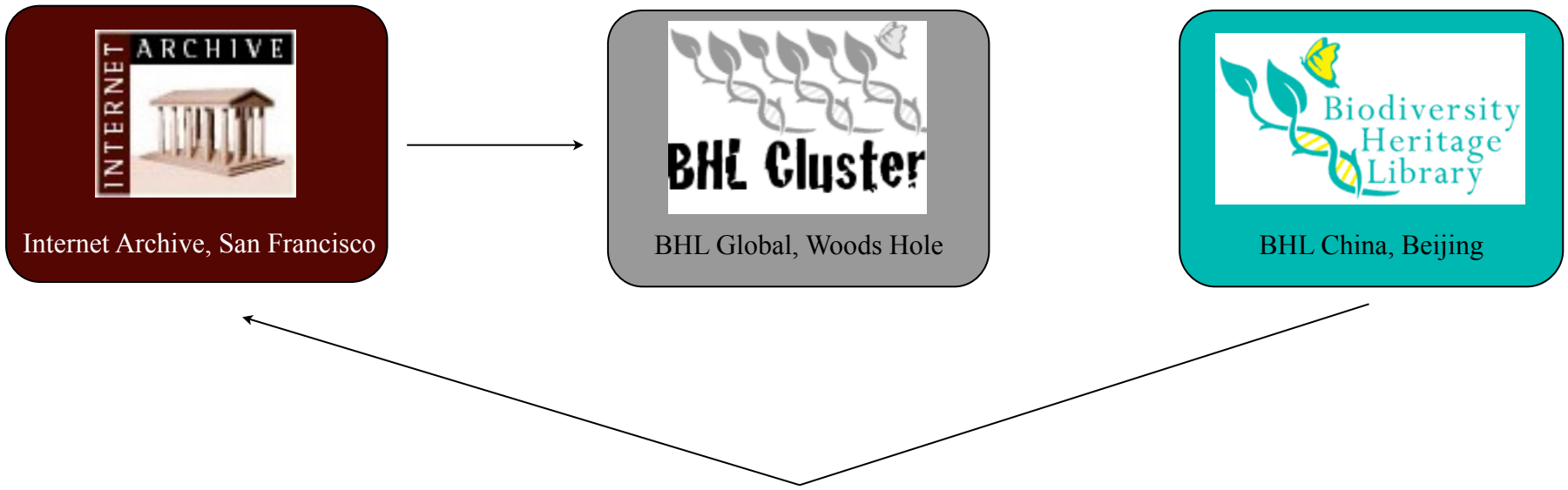
Code available (open sourced, BSD licensed):
[1] http://code.google.com/p/bhl-bits/source/browse/trunk/utilities/grabby/grabbyd
[2] http://code.google.com/p/bhl-bits/source/browse/trunk/utilities/bhl-sync.sh

Other replication challenges

BHL Cluster

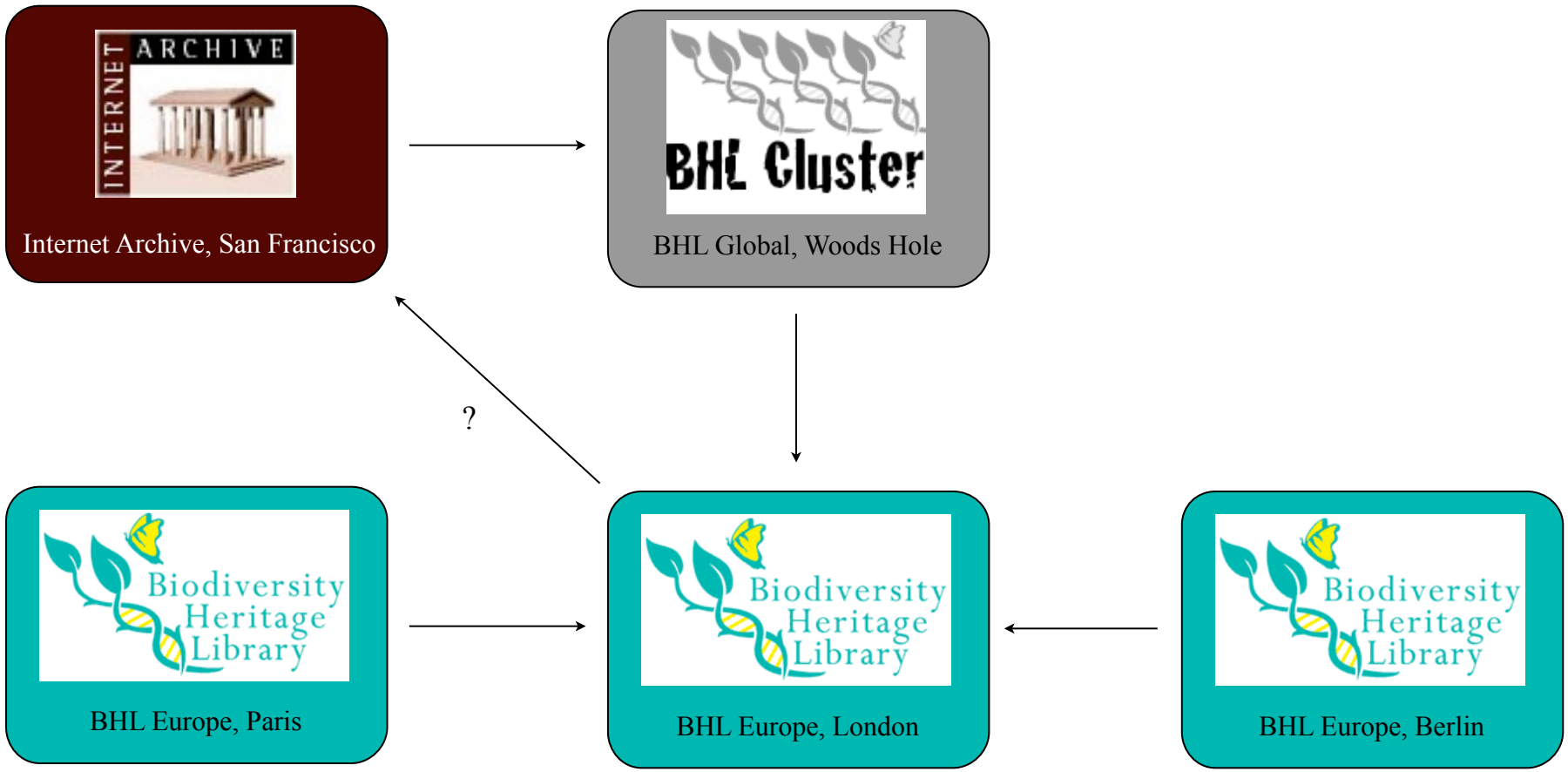
- Deleting content - "going dark"
  - this can be data that is removed from search indexes, but still retrievable via URI
  - or deleted data not available (requires a separate sync process)
- New content coming in from other sources
  - Localization of content - maybe it all can't be shared?
  - National nodes consideration

BHL content + local data



Content sourced from China, scanned by Internet Archive, replicated into BHL Global

BHL content + regional data



Content sourced from BHL Europe partners may, or may not, be passed back to Internet Archive and BHL Global

Fedora-commons integration

Integrated digital repository-centered platform

- Enables storage, access and management of virtually any kind of digital content
- can be a base for software developers to build tools and front ends on for sharing, reuse and displaying data online
- Is free, community supported, open source software

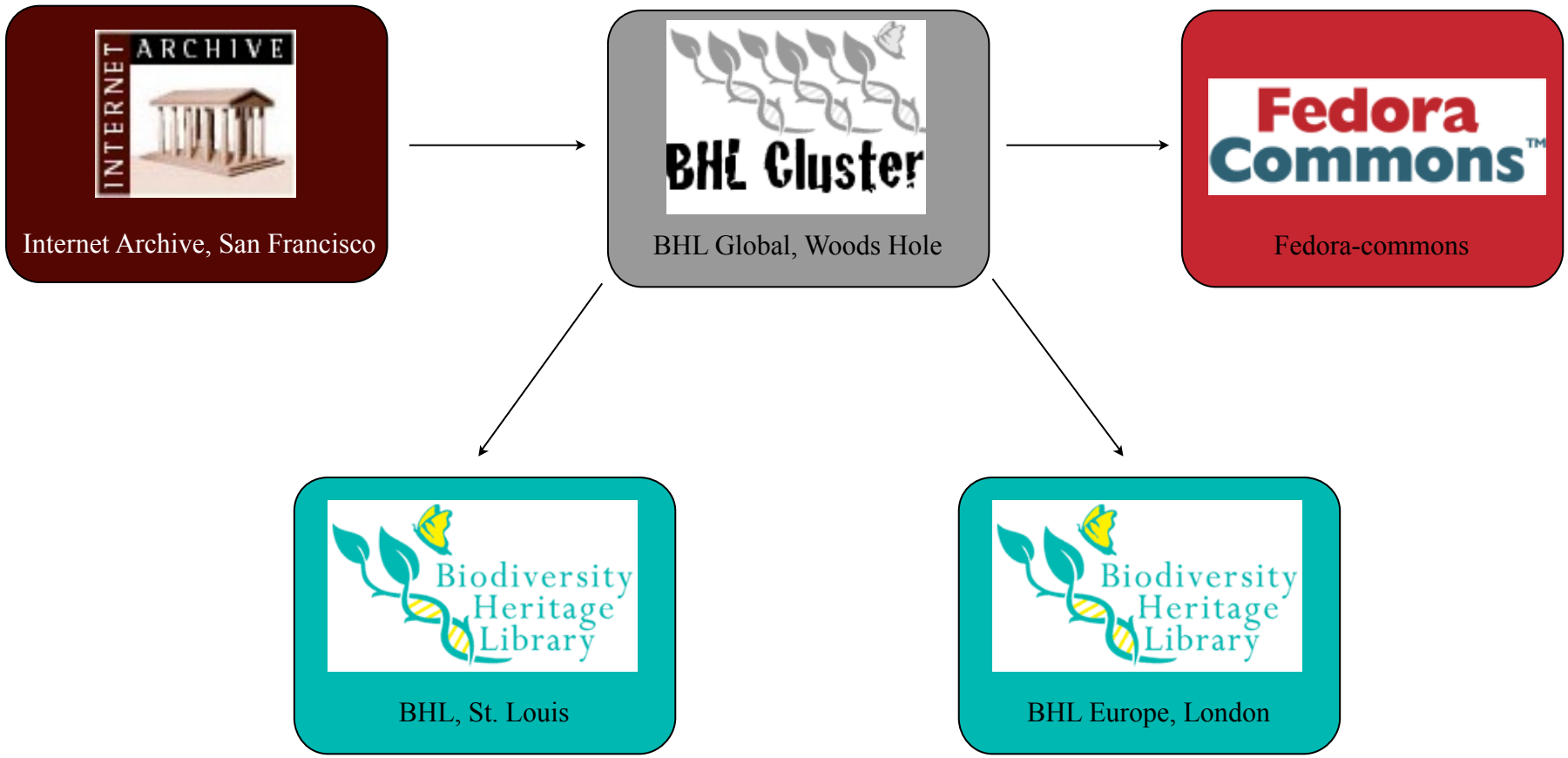
Fedora-commons integration

Integrated digital repository-centered platform

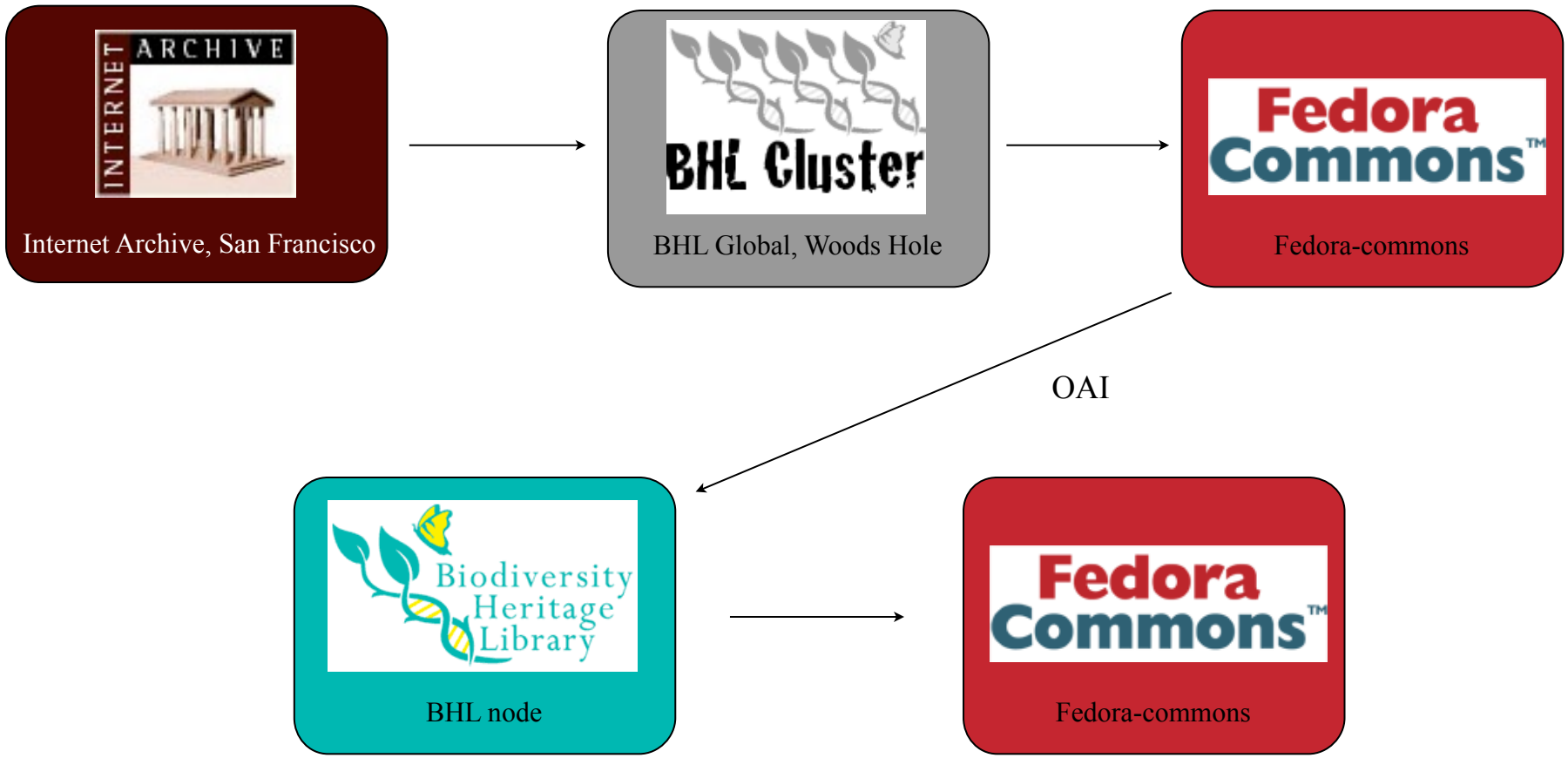
- Enables storage, access and management of virtually any kind of digital content
- can be a base for software developers to build tools and front ends on for sharing, reuse and displaying data online
- Is free, community supported, open source software

- Creates and maintains a persistent, stable, digital archive
  - provides backup, redundancy and disaster recovery
  - complements (doesn't replace or put any demands upon) existing architecture by incorporating open standards
  - stores data in a neutral manner, allowing for an independent disaster recovery option
  - shares data via OAI, REST based interface

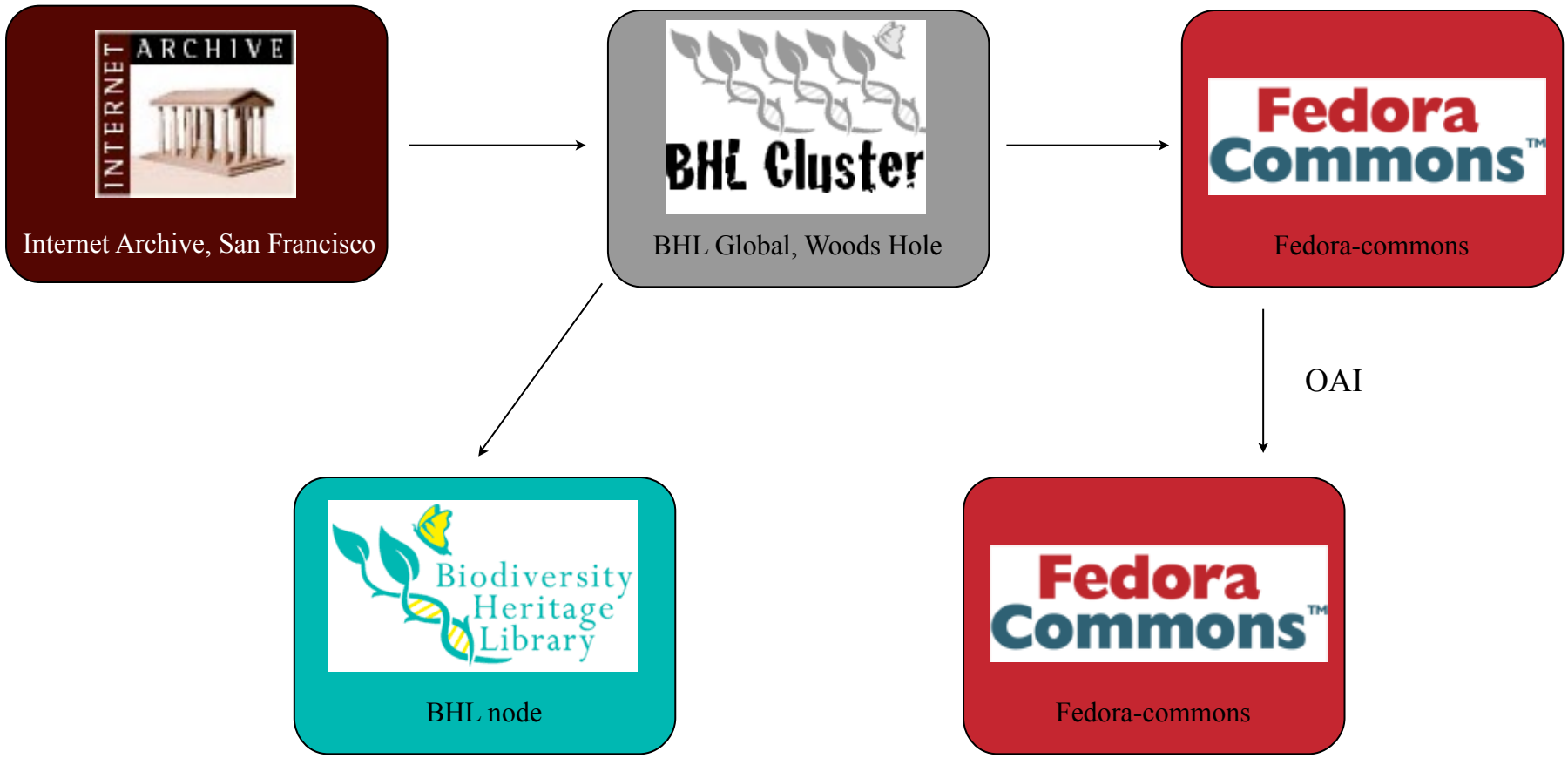
BHL content distribution



BHL content distribution



BHL content distribution



Thanks + questions

Thanks to Adrian Smales, Chris Sleep (NMH), Chris Freeland, Tom Garnett (BHL) and Cathy Norton, Anthony Goddard, Woods Hole networking admins (MBL) for their work and support of this project.

BHL Cluster

email phil.cryer@mobot.org
skype phil.cryer
twitter @fak3r

slides available on slideshare