Identifying and retrieving digital objects: A Study of the Handle System

Taarik Hassanmahomed
System and Network Engineering
June 30, 2010
CineGrid.org

CineGrid's Mission:
To build an interdisciplinary community that is focused on the research, development, and demonstration of networked collaborative tools to enable the production, use and exchange of very-high-quality digital media over photonic networks.

Source:Cinegrid.org
AMPAS

- One of the members is AMPAS – the Academy of Motion Picture Arts and Sciences.
- They are just like the rest of the community looking into new way to manage their data explosion.
  - 500 movies every year
  - metadata storage frame by frame
  - over 25 million object per movie
- A candidate for managing this is the Handle System.
Research question

• How can the Handle System help in storage, search, retrieval and preservation of digital content more efficient and reliable within AMPAS/CineGrid in particular?
Overview

- Identifiers
- Metadata
- Handle System
- Applicability in CineGrid/AMPAS
Overview

- Identifiers
- Metadata
- Handle System
- Applicability in CineGrid/AMPAS
Identifiers

- Namespaces
  - Uniqueness and Persistence
- Locating object
  - Broadcast, Home-base, P2P, Hierarchical
- Scalability
  - Distribution and Replication
- Preservation
  - Loss of meaning and loss of provenance or authenticity
Identifier using metadata

<table>
<thead>
<tr>
<th>ISAN</th>
<th>ISAN 0000-0000-D07A-0090-Q-0000-0000-X</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenURL</td>
<td><a href="http://resolver.x.com/cgi?genre=book&amp;isbn=0236218310">http://resolver.x.com/cgi?genre=book&amp;isbn=0236218310</a></td>
</tr>
<tr>
<td>PURL</td>
<td><a href="http://purl.oclc.org/oclc/rsch/metadataAll">http://purl.oclc.org/oclc/rsch/metadataAll</a></td>
</tr>
</tbody>
</table>
Overview

- Identifiers
- **Metadata**
- Handle System
- Applicability in CineGrid/AMPAS
Metadata

“An element of metadata describes an information resource, or helps provide access to an information resource. A collection of such metadata elements may describe one or many information resources “

W. Cathro (1997)

- Metadata is an important part in understanding the semantics of digital content
- Power lies in choosing the right set of element.
Metadata categories

- Metadata itself can be categorised in various ways:
  - General vs. Specialistic
  - Minimalistic vs. Rich
  - Hierarchical vs. Linear
  - Embedded vs. Detached
  - and much more
Metadata schema examples

- Dublic Core,
  - 15 core elements, interoperable
- IEEE LOM,
  - nine categories, sub elements
- MPEG7,
  - Video, audio, generic features, multimedia description
- HTML
  - DESCRIPTION and KEYWORDS
- Conclusion there is no single best solution
Overview

- Identifiers
- Metadata
- Handle System
- Applicability in CineGrid/AMPAS
Handle System

- CineGrid.AMPAS\NiceMovie4K
Handle System properties

- Globally unique identifiers
- Handle name persistence
- Multiple instances and attributes of an object
- Extensible namespace
- Distributed service model
- Secured name service
- Distributed administration service
HS Architecture

GHR : Global Handle Service
LHR : Local Handle Service

GHR
LHS-CineGrid
LHS-UvA
LHS-AMPAS

Site3
SiteN
Prague
Ams

Contain same group of handles
Distributed over different number of servers

Replicas

PRIMARY SITE

1 2 3 4 N
Overview

- Identifiers
- Metadata
- Handle System
- Applicability in CineGrid/AMPAS
Handle metadata access

- Web proxy
  - GHS proxy at hdl.handle.net
  - local LHS proxy (port 8000)
  - Any other non local LHS proxy
- Java tool from handle.net
  - Admin tool
  - Java command line tool
- Custom client Java/C application
  - OpenHandle (open source)
    - http://code.google.com/p/openhandle/
Web proxy resolution

- "wget" handles with the web proxy hdl.handle.net
- **Recursive resolution**

<table>
<thead>
<tr>
<th>URL</th>
<th>time req-resp</th>
<th>Ping</th>
</tr>
</thead>
<tbody>
<tr>
<td>hdl.handle.net</td>
<td></td>
<td></td>
</tr>
<tr>
<td>glow.handle.net</td>
<td>0.2945 seconds!</td>
<td>0.0975 seconds</td>
</tr>
<tr>
<td>Local Proxy</td>
<td>0.0103 seconds</td>
<td>same server</td>
</tr>
</tbody>
</table>
Java tool resolution

- Direct request with the Java command line
- **Iterative resolution**

<table>
<thead>
<tr>
<th>time req-resp</th>
<th>GHS*</th>
<th>location</th>
<th>ISP</th>
<th>Ping</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6102 seconds!</td>
<td>glow.handle.net</td>
<td>US, United States</td>
<td>PSI</td>
<td>0.0963 seconds</td>
</tr>
<tr>
<td>0.6251 seconds!</td>
<td>macmini1.handle.net</td>
<td>Reston, VA 20191</td>
<td>CNRI</td>
<td>0.0978 seconds</td>
</tr>
<tr>
<td>0.5997 seconds!</td>
<td>hercules.handle.net</td>
<td>Reston, VA 20191</td>
<td>CNRI</td>
<td>0.0989 seconds</td>
</tr>
<tr>
<td>0.5826 seconds!</td>
<td>Crossref.org</td>
<td>Lynnfield, MA 0194</td>
<td>Verizon</td>
<td>0.0987 seconds</td>
</tr>
<tr>
<td>0.7634 seconds!</td>
<td>China</td>
<td>Beijing, 22</td>
<td>CNIC</td>
<td>0.1103 seconds</td>
</tr>
</tbody>
</table>

1. request handle resolution from random server
2. random reply with address of primary or mirror
3. request handle resolution
4. respond with handle data
Failover test

- Turning of the primary to see how many times handles from the primary are requested.
  - The command line tool (50% of the time)

<table>
<thead>
<tr>
<th>action</th>
<th>GHS</th>
<th>LHS</th>
<th>LHS</th>
<th>time req-resp</th>
</tr>
</thead>
<tbody>
<tr>
<td>primary down</td>
<td>Random</td>
<td>primary (3x)</td>
<td>mirror</td>
<td>6.5711 seconds!</td>
</tr>
</tbody>
</table>

- The web proxy (10% of the time)

<table>
<thead>
<tr>
<th>action</th>
<th>GHS</th>
<th>time req-resp</th>
</tr>
</thead>
<tbody>
<tr>
<td>primary down</td>
<td>glow.handle.net</td>
<td>10.2958 seconds!</td>
</tr>
</tbody>
</table>
Resolution recommendations

- There is just one recommendation and that is to use a local caching server which is one of the optional component of the Handle System.
  - It caches handle data, service information of any LHS and allows re-use of information obtained from earlier queries, reducing traffic between Handle System clients and servers.
Handle metadata access

- Custom client Java/C application
  - OpenHandle v0.21 (open source)
    - http://code.google.com/p/openhandle/

- Repopulate a copy of the Cinegrid Amsterdam portal with handles
Big Buck Bunny

Author: Blender Foundation
Created: ago
Tags:
Preview images: [4k tiff][4k jpeg]
Description:
[c] copyright Blender Foundation | http://www.bigbuckbunny.org

<table>
<thead>
<tr>
<th>Codec</th>
<th>Format</th>
<th>Location</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpeg4</td>
<td>1920x1080</td>
<td>cgdevil</td>
<td>1.1 GB</td>
</tr>
</tbody>
</table>
Architecture
Architecture

User

Portal

Display

Disk

Stream server

bvmpplayer

iRODS

VLC

metadata server

XML

RDF

iRODS

SQL DB

metadata client

XML RPC

Streaming node #1

Streaming node #n
New architecture

[Diagram showing a new architecture for a system involving components such as User, Display, Portal, Disk, stream server, bvplayer, iRODS, VLC, and interactions between HTML, Python script, JSON/XML, and Java.]

Python script

JavaScript

XML

JSON

OpenHandle

JAVA

HDl
Example of Population

Template used:

After choosing handle:

Big Buck Bunny

(c) copyright Blender Foundation | http://www.bigbuckbunny.org
Selecting multiple handles

**Big Buck Bunny**

- **Author:** [Blender Foundation](http://www.bigbuckbunny.org)
- **Created:** 1970-01-01 01:33:28 ago
- **Tags:**
  - Preview images: [4k tiff] [4k jpeg] [4k small] [4k normal]
- **Description:**
  
  (c) copyright Blender Foundation | http://www.bigbuckbunny.org

<table>
<thead>
<tr>
<th>Codec</th>
<th>Format</th>
<th>Location</th>
<th>Size</th>
<th>Framerate</th>
<th>Duration</th>
<th>Filename</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpeg4</td>
<td>1920x1080</td>
<td>cgdevil</td>
<td>1135094932 bytes</td>
<td>30 frames</td>
<td>3600 minutes</td>
<td>big_buck_bunny_1080p.mp4v</td>
</tr>
</tbody>
</table>

**7 Bridges**

- **Author:** [CineGrid](http://www.cinegrid.org)
- **Created:** 1970-01-01 01:33:27 ago
- **Tags:**
  - Preview images: [4k tiff] [4k jpeg] [4k small] [4k normal]
- **Description:**
  
  (c) A boat ride on the canals of Amsterdam.

<table>
<thead>
<tr>
<th>Codec</th>
<th>Format</th>
<th>Location</th>
<th>Size</th>
<th>Framerate</th>
<th>Duration</th>
<th>Filename</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpeg4</td>
<td>1920x1080</td>
<td>cgdevil</td>
<td>214748364 bytes</td>
<td>30 frames</td>
<td>138 minutes</td>
<td>hollandfestival07.7bridges.1080p.30fps.dxt.bmv</td>
</tr>
</tbody>
</table>
# Handle system metadata

<table>
<thead>
<tr>
<th>Handle:INDEX</th>
<th>#</th>
<th>TYPE</th>
<th>HANDLE DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>10677/7_Bridges</td>
<td>1</td>
<td>TITLE</td>
<td>7 Bridges</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>AUTHOR</td>
<td>CineGrid</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>DESC</td>
<td>(c) A boat ride on the canals of Amsterdam.</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>CREATED</td>
<td>&quot;1970-01-01 01:33:27&quot;</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>URL</td>
<td><a href="http://cinegrid.uvalight.nl/portal">http://cinegrid.uvalight.nl/portal</a></td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>IMAGES</td>
<td><a href="http://cinegrid.uvalight.nl/images/bridge.png">http://cinegrid.uvalight.nl/images/bridge.png</a></td>
</tr>
</tbody>
</table>
CineGrid Handle Access

- CineGrid community consists of various members all over the globe and therefore content all over the world.
- Handles are not hard linked to any digital content.
- Use the whole prefix: CineGrid\LHS
  - Use a member site as the primary, other members can then choose to become a mirror or leave the replication to the primary.
- Make use of subprefixes: CineGrid.AMPAS\LHS
  - Every member can be a primary and have administration and storage near by.
Conclusion

- Handle system is fairly good for storage, retrieval and preservation of metadata and the location of digital content.
- Combination of the LHS and GHS with all its characteristics ensures that digital content can continue to grow and still be available without much loss in performance.
- Still it can be seen as a form of middleware, which require application like OpenHandle to get its full potential.
More info?

- Handle System website “http://www.handle.net/”
- RFC3650 “http://www.handle.net/rfc/rfc3650.html”
- RFC3651 “http://www.handle.net/rfc/rfc3651.html”
- CineGrid.org website “http://www.cinegrid.org/”
- CineGrid Ams website “http://cinegrid.uvalight.nl/portal/”
- AMPAS “http://www.oscars.org/”

Questions?