Mbooks @ UM

Archive and Access
Basic archive work

• Considerable attention paid to TRAC requirements and standards (e.g., OAIS) in implementation
• Designed for fast ingest and registration of content, including validation (more later)
• Security is paramount
• Scalable and robust commodity hardware (e.g., Isilon’s storage and blade servers) with replication
• Essentially complete compliance with TRAC required elements; review by DRAMBORA
<table>
<thead>
<tr>
<th>#</th>
<th>Format</th>
<th>Author</th>
<th>Title</th>
<th>Year</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Book</td>
<td>Welsh, Charles</td>
<td>Chauffeur chaff, or, Automobilia, edited by Charles Welsh.</td>
<td>1905</td>
<td>- MBooks Online (Full Text)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Anecdotes, stories, bon-mots, also a history of the automobile.</td>
<td></td>
<td>Buhr Shelving Facility - Ask at any library</td>
</tr>
<tr>
<td>2</td>
<td>Book</td>
<td>Giacosa, D.</td>
<td>Forty years of design with Fiat.</td>
<td>1979</td>
<td>Transportation Research Inst. TRI7111.</td>
</tr>
<tr>
<td>3</td>
<td>Serial</td>
<td></td>
<td>Moteurs.</td>
<td></td>
<td>Buhr Shelving Facility - Ask at any library</td>
</tr>
</tbody>
</table>
sequence #13 - 1 matching term
... the various parts, be able to adjust and clean them, and know by the sounds which issue from the engine and the gearing what is happening in the dark recesses whether the monster has indigestion...

sequence #16 - 1 matching term
... the world is open to the amateur chauffeur just because some clever fellow has invented a little engine to run for hours without tiring, given some common spirit to vaporize, and an electric sp...

sequence #25 - 5 matching terms
...AUTOMOBILIA horse-power of the engine and the nature of the work the car will be put to. The engine power must be such that with...
...
... AUTOMOBILIA horse-power of the engine and the nature of the work the car will be put to. The engine power must be such that with a full load the car will be able to do a journey over give a...
...
... if average speed, and will not fail on an ordinary hill. Makers of repute always proportion their engine power to the work the car will
AUTOMOBILIA

horse-power of the engine and the nature of the work the car will be put to. The engine power must be such that with a full load the car will be able to do a journey over give-and-take roads at a fair average speed, and will not fail on an ordinary hill. Makers of repute always proportion their engine power to the work the car will be put to, and this matter may be safely left to them. As a rule manufacturers understake the horse-power of their cars, but in a badly made car this is no guarantee that the engine will do its work under full load, as the horse-power stated is the gross power given by the engine, and not that which is delivered at the road wheels. In the transmission much power is
AUTOMOBILIA

horse-power of the engine and the nature of the work the car will be put to. The engine
power must be such that with a full load the car will be able to do a journey over give
and-take roads at a fair average speed, and will not fail on an ordinary hill. Makers of
repute always proportion their engine power to the work the car will be put to, and this
matter may be safely left to them. As a
rule manufacturers understimate the horse
power of their cars, but in a badly made car
this is no guarantee that the engine will do
its work under full load, as the horse-power
stated is the gross power given by the engine,
and not that which is delivered at the road
wheels. In the transmission much power is
lost, especially in poorly made vehicles.
Generally speaking, one can only get two
seater cars at prices up to $1,000, although
there are several notable exceptions to that
rule. The horse-power runs to about eight
as the maximum, and a good average figure
is six. Thus, the man who will not spend
Architecture overview

Four basic pieces:

- Mirlyn, the library catalog
- Page image and metadata repository
- Rights and GeoIP databases
- Pageturner
Library catalog — Mirlyn

• During scanning, tracking information is stored in Mirlyn:
  – Manual scanning of barcode or batch process for a call number range
  – Triggers unavailability message online
  – Metadata extracted and made available
  – Daily list of barcodes for returned items used to remove unavailable status
Library catalog — Mirlyn

• When images are added to repository:
  – Barcodes used to add to the item record call-number-2 field
  – Supplies info to rights database
  – Feeds identifiers to staff involved in copyright determination work
  – Record updated in 006, 007, and 533 fields
  – NB: No 856 (a virtual 856 field or link to detailed holdings is generated in the display)
  – Also, later, supplies metadata to the pageturner
METS Object

• Why METS?
  – Can serve as an Archival Information Package and a Dissemination Information Package
  – Designed to record the relationship between pieces of complex digital objects
  – Can be created automatically as texts are loaded or reloaded
METS Object

• What’s there?
  – metsHdr with an ID and CREATEDATE
  – dmdSec with a URL
  – Two techMD referencing notes files
  – Two fileGrps (images and OCR)
  – Physical structMap tying together the files with any metadata (page numbers or features)
  – PREMIS metadata
Page image and metadata repository

• Objectives:
  – A guiding principle: store archival images, create deliverables on demand
  – Incorporate TDR-like practices

• Simple filesystem layout adapted from DLXS
  – One directory per volume, all files (most in a zip file) inside
  – Example: /l1/obj/bc/39015/1/2/3/39015123456789/
  – Use of a namespace allows for conflicting identifiers
  – Currently, several namespaces, including that for the Google-digitized content, UM’s own digitization efforts, and now also for Wisconsin files
Page image and metadata repository

• Download and ingest
  – Google Return Interface (GRIN) at Google provides volumes over HTTPS
    • TIFF, JP2, UTF-8 OCR, metadata in a package
  – Google Return (Object Oriented) Validation Environment (GROOVE) at UM
    • manages all download and validation
    • Written in Perl; MySQL backend for state tracking
    • available to Google Partners via CVS
Page image and metadata repository

- **Automatic validation in GROOVE**
  - Check barcode check digit using Luhn algorithm
  - Fixity check on JP2, TIFF, UTF8 using MD5
  - Well-formedness and embedded metadata check on JP2, TIFF, UTF8 using JHove
  - Various completeness cross-checks
  - Failures retried, will eventually refer to Google

- **Periodic fixity checks using MD5**
Page image and metadata repository

- Files for qualitative validation extracted
- METS file created (functions as a manifest)
- Persistent identifier (Handle) created
- Feed of identifiers sent to/from Mirlyn via ssh
Rights database

• **What information to store?**
  – Complex and changing table of statuses and means of determination
  – Considered using MARC directly, but need for interaction/updates and “service” pushed for separate DB

• **Approach:** examine bib record, determine authoritative copyright status, store rights attribute, source, reason, and timestamp

• **Stored in MySQL**
Rights database

- **Rights attributes, most currently in use**
  - pd: public domain
  - pdus: public domain for US viewers
  - inc: in copyright
  - und: undetermined (a body of work for cataloging!)
  - nobody (override): no access
  - orph: orphan copyrighted work
  - umall (override): open access for authenticated UM affiliates
  - world (override): open access to world
Rights database

• Each rights attribute must have a reason, incl:
  – bib: bibliographically-derived
  – man: manual access control override
  – ddd: due diligence documented
  – con: contractual agreement on file (not yet in use)

• Source identified
  – e.g., ‘google’ or ‘lit-dlps-dc’ for UM’s local digitization
  – different actions based on each (e.g., hypothetically, full book download on latter)
GeoIP database

- Right attribute ‘pdus’ requires country of origin.
- Based on feed from MaxMind GeoIP database:
  - Update utility and several query APIs included
  - $12/month
Pageturner middleware

• **Ties all the rest together**
  – *User interface*
  – *Access rights determination*
  – *OCR, page image access*
  – *Item search*

• **Gives users the opportunity to provide feedback**
Pageturner interactions

Access GeoIP and Rights dB

• Access repository for METS object and archival image
• Access Mirlyn library catalog for item metadata
• Transform images and cache them
• Wrap in XML, apply XSL to create HTML
Pageturner: page image retrieval

- Rights database
- GeoIP database
- METS XML
- Library catalog metadata
- Archival page image
- Online page image
- XSLT
- XML
- HTML
- Browser
Pageturner: *Item searching*

- Wrap, concatenate and cache OCR files
- Create and cache indexes
- Perform search
- Display results
Pageturner: Item searching

Create and cache indexes
- XPAT full text index
- XML page region index
- Processed in real-time
- Cache indexes

Display results
- Hit counts and KWICs, per page
- No KWICs if access restricted
- OCR Word/Phrase highlighting
Pageturner: *Item searching*

Note: detailed search results are not shown for restricted items.

Sequence #33 - 3 matching terms
Sequence #109 - 2 matching terms
Sequence #255 - 4 matching terms
Sequence #257 - 4 matching terms

Viewing results for: 1 to 4 of 4 pages
What’s Next?

• Building collections
• Continue work on generalizing (e.g., skinnable pageturner)
• Large-scale fulltext search (SOLR)
• Fedora integration
• Facilitating research applications (working with NCSA?)
• SDR work
Contact and URLs

- jpwilkin@umich.edu
- MDP info page with some links
  - http://www.lib.umich.edu/mdp/
- SDR
  - http://tinyurl.com/63vmhy