Distributed Repositories of Medieval Calendars and Crowd-Sourcing of Transcription

Rob Sanderson  
azaroth42@gmail.com  
azaroth@stanford.edu  
t: @azaroth42  
Stanford University

Ben Albritton, Stanford University  
Doug Emery, University of Pennsylvania  
Will Noel, University of Pennsylvania  
Dot Porter, University of Pennsylvania

http://iiif.io/

This research was primarily funded by the Andrew W. Mellon Foundation
Image Repositories

- Increase in digitization
  - Particularly precious, fragile, beautiful objects
  - e.g. Medieval Manuscripts

- Digitized images online
  - Increasingly Open
  - At high resolution

- Easy to capture an image
- Very hard to capture the text

http://gallica.bnf.fr/ark:/12148/btv1b8449691v/
Calendar Pilot

• Ubiquitous in liturgical books
  • e.g. Books of Hours
• Structured and often tabular: Date, Day, Saint / Event

• Content varies slightly
• Variation details give us information about the provenance of the object

• Much easier to transcribe than full text

http://www.e-codices.unifr.ch/en/bge/lat0033
Collaborative Crowd Sourcing?

- Meeting at U. Penn including content providers and scholars

- Plan:
  - Collect transcriptions together
  - Analyze similarities between manuscripts for patterns of provenance

- Manuscripts and images distributed: need a community to collect sufficient data

http://brbl-dl.library.yale.edu/vufind/Record/3446275
Micro Repository Rant: TEI

- Most transcribing done in TEI
- Terrible for this use case:
  - Single XML file
  - Single author
  - Single location
  - Hard to link to images
  - Tries to describe too much
  - Impossible to use once created
- Creating TEI is good for:

http://www.thedigitalwalters.org/Data/WaltersManuscripts/html/W41/
Micro Repository Rant: TEI

- Most transcribing done in TEI

- Terrible for this use case:
  - Single XML file
  - Single author
  - Single location
  - Hard to link to images
  - Tries to describe too much
  - Impossible to use once created

- Creating TEI is good for:
  - The academic exercise of creating TEI

http://www.thedigitalwalters.org/Data/WaltersManuscripts/html/W41/
Requirements

- Distributed image content
  - Consistent, rich API

- Selection of regions
  - Base, not displayed size

- Alignment of text with region
  - Distributed creation
  - Distributed curation
  - Multiple texts per region
  - Styling of the text
  - Some semantics

http://oculus-dev.lib.harvard.edu/manifests/view/drs:5981093
Open Technology: IIIF Image API

Base URL: `{scheme}://{host}{/prefix}/{identifier}`

Image Resource:

`{base}/{region}/{size}/{rotation}/{quality}.{format}`

```
http://www.example.org/image-service/abcd1234/80,15,60,75/pct:80/345/grey.jpg
```

**Order of Implementation**

```
http://iiif.io/api/image/1.1/
```

Distributed Repositories and Crowd-Sourcing Transcription
Open Repositories 2014, Helsinki, Finland, 11\textsuperscript{th} of June 2014
Open Technology: IIIF Image API

```
{
  "@context" : "http://library.stanford.edu/iiif/image-api/1.1/context.json",
  "@id" : "http://iiif.example.com/prefix/object1",
  "width" : 6000,
  "height" : 4000,
  "scale_factors" : [ 1, 2, 4 ],
  "tile_width" : 1024,
  "tile_height" : 1024,
  "formats" : [ "jpg", "png" ],
  "qualities" : [ "native", "grey" ],
  "profile" :
    "http://library.stanford.edu/iiif/image-api/1.1/compliance.html#level0"
}
```

http://iiif.io/api/image/1.1/

Distributed Repositories and Crowd-Sourcing Transcription
Open Repositories 2014, Helsinki, Finland, 11th of June 2014
(Part of the) IIIF Community

- ARTstor
- Bibliothèque Nationale de France
- Bodleian Libraries, Oxford University
- British Library
- C2MRF
- Cambridge University
- Cornell University
- DPLA
- Europeana
- e-codices
- Harvard University
- Johns Hopkins University
- National Library of Denmark
- National Library of Poland
- National Library of New Zealand
- National Library of Norway
- National Library of Wales
- Princeton University
- Stanford University
- Wellcome Trust
- UK National Archives
- Yale University
Distributed Repositories and Crowd-Sourcing Transcription

Open Repositories 2014, Helsinki, Finland, 11th of June 2014

Crowdsourced Box Drawing

Ex libris P. Barnabel
Open Technologies

- Mirador
  - IIIF Community developed viewer
    - Stanford, Harvard, Yale, [LANL]
  - Zooming via Open SeaDragon
    - Princeton, and OSD committers

- JCrop
  - JQuery plugin for drawing little boxes

- MongoDB
  - Store information via REST interface
  - W3C Media Fragment image segments
  - Trivially converted to IIIF Image API requests
Delimit View: Temporary Calendar Images
Open Technology

- Line/Column inspiration from TPEN (IIIF compliant)
- Transcription tool developed at St. Louis
- http://t-pen.org/TPEN/
- Line detection needs correction, no internal columns
Distributed Repositories and Crowd-Sourcing Transcription

Open Repositories 2014, Helsinki, Finland, 11th of June 2014

Open Technologies

• Inspiration from TPEN (IIIF compliant)
• Transcription tool developed at St. Louis
• http://tpen.org
• Line detection flakey, no internal columns
Open Technologies

• Inspiration from TPEN (IIIF compliant)
• Transcription tool developed at St. Louis
• http://t-pen.org/TPEN/
• Line detection flaky, no internal columns
Boring (but Open) Metadata

- Metadata collection to drive the analysis
  - Stored along with the segments
  - Defaults are normally correct
  - Custom extension, not intended for general purpose use

- Convenient to do inline
  - Other metadata could be added
  - Could be done in a different workflow
Transcription View: Temporary Calendar Images

Image View: Temporary Calendar Images / Image: 12148-f4

Annotation List (1)

All (1)

La circumsicion

No Dimensions Given

Letter: text
Transcription View: Temporary Calendar Images

Image View: Temporary Calendar Images / Image: 12148-f4

Annotation List (1)

All (1)
La circumcision

No Dimensions Given

Octaves S. Estiene

b
Transcription View: Temporary Calendar Images

Image View: Temporary Calendar Images / Image: 12148-f4

Annotation List (2)

All (2)

La circuncision

Octaves S. Estiene
<table>
<thead>
<tr>
<th>Page</th>
<th>Letter</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>12148-f4</td>
<td>A</td>
<td>La circuncision</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>Octaves S. Estiene</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>Saincte Geneviefve</td>
</tr>
</tbody>
</table>
Open Technology: IIIF Presentation API

Text/Image Linking is a subset of a larger challenge:

- Non-Text / Image Linking
- Dynamic Images
- No Image to link to
- Multiple Images
- Parts of Images
- Parts of larger texts
- Distributed images, texts and links

Need an indirection layer:

- Solution: align text and image with an abstract Canvas

http://iiif.io/api/presentation/1.0/
Open Technology: IIIF Presentation API

Canvas

target

body

Anno

Distributed Repositories and Crowd-Sourcing Transcription
Open Repositories 2014, Helsinki, Finland, 11th of June 2014
Open Technology: IIIF Presentation API

Canvas

Anno

target

body

iii A La circuncision

Anno

body

target
Open Technology: IIIF Presentation API

Canvas

Anno

body

iii A La circuncision

Anno

body

b Octaves S. Estiene

Anno

body

Canvas

Anno

body

Canvas
Linked Data People...

If you do not want to know the score, look away now!
{ "it's" : "just JSON" }
{ } Are The New <>

{
  "@context" : "http://www.shared-canvas.org/ns/context.json",
  "@id" : "http://www.example.org/iiif/book1/manifest.json",
  "@type" : "sc:Manifest",
  "label" : "Book 1",
  "metadata" : [
    {"label" : "Author", "value" : "Anne Author"},
    {"label" : "Published", "value" : {"@value" : "Paris", "@language" : "en"} } ],
  "attribution" : "Provided by Example Organization",
  "sequences" : [
    {
      "@type" : "sc:Sequence",
      ...
    }
  ]
}
Web Developers...

If you do not want to know the score, look away now!
Web Developers...

<__:it's__>
<__:all__>
<__:Linked_Data__>
Micro Repository Rant 2: RDF Serialization

“RDF/XML was the Semantic Web’s 3 Mile Island incident”

Or … RDF – Not in my back yard!

• Serializing a graph is, admittedly, hard
• RDF/XML is terrible, and too many others
• Web currently uses JSON as convenient transfer syntax
• JSON-LD allows transfer of RDF in syntax that does not require full RDF stack, just a JSON implementation
• … as available in every web browser

• Rob's Conclusion: Require JSON-LD
  • http://json-ld.org/
JSON-LD Context Magic

```json
{ // Canvas resource
    "@context": "http://iiif.io/api/presentation/2/context.json",

    @context provides mapping for JSON keys into RDF.

    "sc": "http://www.shared-canvas.org/ns/",
    "oa": "http://www.w3.org/ns/oa#",
    "service": {
        "@type": "@id",
        "@id": "sioc_svcs:has_service",
    },
    "height": {
        "@type": "xsd:integer",
        "@id": "exif:height",
    },
    "sequences": {
        "@type": "@id",
        "@id": "sc:hasSequences",
        "@container": "@list"
    }
}
```
Open Technologies: REST

- Experimental IIIF REST specification
  - http://iiif.io/api/annex/rest/
  - For both Presentation and Image

- Trivial Python/WSGI handler
  - Processes @context and generates identities
  - Stores in MongoDB (but API is agnostic)
  - Follows IIIF Presentation and Open Annotation
    - http://www.w3.org/community/openannotation/
  - Returns the correct JSON-LD
  - Doesn't fully handle image upload yet
The Future is Now

- IIIF Image API 2.0
  - Request for Comment period open!
  - http://iiif.io/api/image/2.0/

- IIIF Presentation API 2.0
  - Ditto!
  - http://iiif.io/api/presentation/2.0/

Please give us feedback: iiif-discuss@googlegroups.com

- Ongoing work with U.Penn to make a more robust system
Thank You

Rob Sanderson
azaroth42@gmail.com
azaroth@stanford.edu
t: @azaroth42
Stanford University

http://iiif.io/
iiif-discuss@googlegroups.com