Implementing the IIIF Content Search API at Stanford

Mark A. Matienzo, Stanford University Libraries
@anarchivist / https://matienzo.org/presentations
2018 IIIF Conference, 24 May 2018

Project context: Virtual Tribunals

- Collaborative initiative between Stanford Libraries and WSD Handa Center for Human Rights and International Justice
- Technical work part of a larger grant project to provide access to documents from Special Panels for Serious Crimes in East Timor
- Situated in how Digital Library Systems and Services and Stanford Libraries approaches projects and infrastructure

Scope of work

- Provide “search within” a particular SDR object’s pre-existing text (e.g. OCR or transcription), including phrase searching
- Support subset of OCR formats: ALTO, plain text
- Implement parts of Content Search API specifications needed to support use case

`content_search`: the application

- https://github.com/sul-dlss/content_search
- Dependencies: Ruby, Rails, Solr, Redis
- Tightly integrated with SDR infrastructure

Indexing process

- Identify new/changed objects (or object for one-off indexing)
- Fetch object structure
- Identify applicable files (e.g. OCR/transcription)
- Read files, transform content, and index to Solr

Text transformation for indexing

- Text and word boundaries extracted from source file
- Each **`alto:TextBlock`** treated as single value in a multivalued field
- Each **`alto:TextLine`** delimited by **`\n`** in each field
- Concatenate each word with its boundaries using a symbol to generate a Lucene payload: ☞


```
{
  "id": "kx307gd4524/kx307gd4524_6/EastTimor_CE-SPSC_Final_Decisions_2001_17-
2001_Cipriano_da_Costa_Decision_Withdrawal_0005.xml",
  "druid": "kx307gd4524",
  "resource_id": "kx307gd4524_6",
  "filename": "EastTimor_CE-SPSC_Final_Decisions_2001_17-
2001_Cipriano_da_Costa_Decision_Withdrawal_0005.xml",
  "ocrtext": [
    "UNITED 536.00,188.00,232.29,44.00 NATIONS 807.00,188.00,271.00,44.00",
    "NATIONS 1412.00,186.00,264.92,46.00 UNIES 1714.77,186.00,189.23,46.00",
    "TRIBUNAL 593.00,312.00,346.21,42.20 DISTRITAL 982.48,312.00,389.48,42.20 DE
1415.24,312.00,86.55,42.20 DILI 1545.07,312.00,173.10,42.20\nPENGADILAN
852.66,396.40,432.76,42.20 DISTRIK 1328.69,396.40,302.93,42.20 DILI
1674.90,396.40,173.10,42.20\nDISTRICT 895.93,480.80,346.21,42.20 COURT
1285.41,480.80,216.38,42.20 OF 1545.07,480.80,86.55,42.20 DILI 1674.90,480.80,173.10,42.20",
    "SECÇÃO 691.00,610.00,512.30,73.00 CRIMES 1242.70,610.00,236.44,73.00 GRAVES
1518.56,610.00,236.44,73.00",
    "SERIOUS 937.00,702.00,286.00,47.00 CRIMES 1263.86,702.00,245.14,47.00",
    "A 322.00,921.00,24.70,51.33 retirada 371.40,921.00,197.59,51.33 da
593.68,921.00,49.40,51.33 acusação 667.78,921.00,197.59,51.33 pode 890.07,921.00,98.79,51.33
ocorrer 1013.56,921.00,172.89,51.33 através 1211.15,921.00,172.89,51.33 de
1408.74,921.00,49.40,51.33 um 1482.84,921.00,49.40,51.33 requerimento
1556.93,921.00,296.38,51.33 preliminar 1878.01,921.00,246.99,51.33\ntal
322.00,1023.67,74.10,51.33 como 420.79,1023.67,98.79,51.33 resulta
544.29,1023.67,172.89,51.33 do 741.88,1023.67,49.40,51.33 artigo 815.97,1023.67,148.19,51.33
27° 988.86,1023.67,74.10,51.33 do 1087.66,1023.67,49.40,51.33 citado
1161.75,1023.67,148.19,51.33 Reg. 1334.64,1023.67,98.79,51.33 2000/30.
1458.14,1023.67,197.59,51.33",
```

Metadata and access integration

- Manifests dynamically generated from our delivery systems
- Incorporation of content search services into manifests triggered by structural metadata
- Any changes to objects (addition of new OCR resources) will lead to transparent updates

Example

- homepage
- manifest
  - search **`service`**
    - autocomplete **`service`**

  - **`rendering`**: source PDF and searchable PDF (not from `content_search`)
  - **`Canvas`**: no text annotations or **`seeAlso`**s (yet)

Search API flow

- Search Solr for each word for terms or phrase and get the matching pages with hit highlighting
- Extract text preceding and following matches to return **`before`** and **`after`**
- Transform each hit highlight into an **`Annotation`** that is on a **`Canvas`** URI fragment identified by word boundaries
- Create the hits and response

Autocomplete API flow

- Search Solr (with custom suggester) for requested string
- Remove duplicate matches
- Sort by occurence (“weight”) and then by length
- Gather the top 5 results
- Build the response

Strengths

- Both simple word and phrase matching
- Ability to provide surrounding text to put in context
- A potential good start for a more generic Content Search API implementation?

Limitations

- Tightly coupled to Stanford's infrastructure
- Can't effectively phrase search across pages/canvases (no multi-hit annotations)
- Only intended for searching OCR text
- No support for **`motivation`**, **`date`**, and **`user`** queries
- No persistent or dereferenceable annotations, even for text resources

Challenges

- Identifying things beyond scope of the specification that are supported (e.g. phrase searching)
- Authentication for restricted text resources
- Client behavior
  - UV expects Content Search API 0.9 responses
  - Adding support for multiword autocomplete: https://github.com/UniversalViewer/universalviewer/pull/552
- User experience and import to viewer behavior

Thank you!

Mark A. Matienzo
@anarchivist
matienzo.org/presentations
2018 IIIF Conference
24 May 2018

Virtual Tribunals Team

- Cathy Aster
- Chris Beer
- Gary Geisler
- Darren Hardy
- Jessie Keck
- Kris Kasianovitz
- Mark Matienzo
- Jack Reed
- Penelope Van Tuyl
- Camille Villa
- Drew Winget