Something Old, Something New

Applying Linked Data to a Digital Repository

Charles Blair

Digital Library Development Center
University of Chicago Library

THE UNIVERSITY OF CHICAGO Library

THE UNIVERSITY OF CHICAGO

born-digital. retrospectively digitized. untidy archival and mss. collections. tidy digital collections. text, image, audio, audio-visual. simple structure. complex structure.

Workflow

Transferring

Accessioning

Processing

Transferring

preserve the bits. provide basic administrative metadata: who initiated the transfer; what the transfer contains; what constraints (rights and permissions) pertain to the transferred materials.

Accessioning

all accessions (deposits) must belong to a collection. establish a collection for the accession if one does not already exist. assign a NOID (Nice Opaque Identifier) for the accession, create a formal statement of rights and restrictions, including embargoes (e.g., "R-80 or death"); size; preferred citation; abstract. generate technical metadata (FITS). migrate at-risk file formats. record all of this in a relational database.

Processing

archivists mean something specific by processing: arranging the inventory into boxes and folders; creating a finding aid. we will appropriate that term for the library digital repository and map it onto the OAIS reference model, returning to the archival use at the end.

OAIS Reference Model: Information Packages

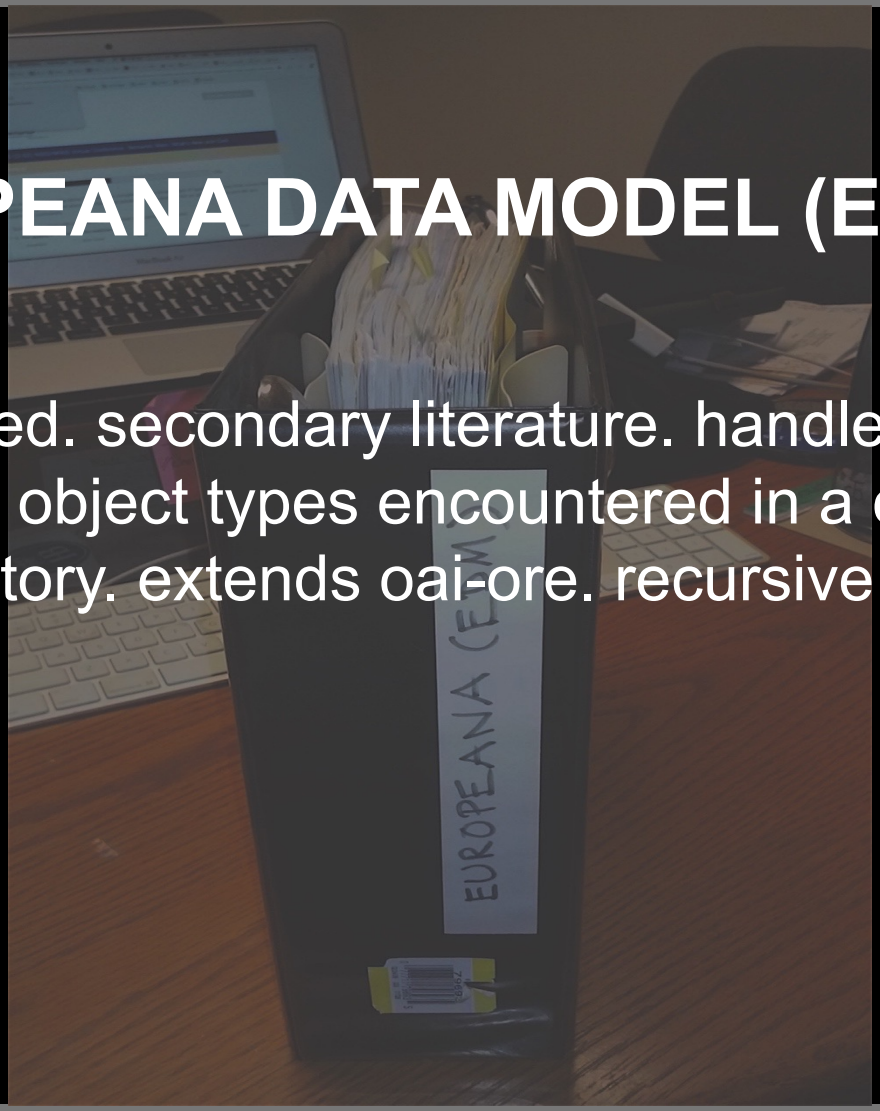
Within the OAIS model, three types of information package are identified: the Submission Information Package (SIP), which is sent from the information producer to the archive; the Archive Information Package (AIP), which is the information package actually stored by the archive; and the Dissemination Information Package (DIP), *which is the information package transferred from the archive in response to a request by a consumer*.

Brian Lavoie, "Meeting the challenges of digital preservation: The OAIS reference model", OCLC Newsletter, No. 243:26-30 (January/February 2000). (my emphasis)

Processing (cont'd)

SIPs are created as linked data (Turtle -> RDF/XML). AIPs are RDF triples in an RDF triplestore (database). DIPs are produced as structured XML (could be JSON as well) in response to SPARQL queries, or the semantic web query language for RDF triplestores. Our DIPs are therefore precisely "information package[s] transferred from the archive in response to a request by a consumer". They are lightweight, easy to transport, robust, and actionable, using standard tools for the purpose (e.g., cURL).

How do we do this?



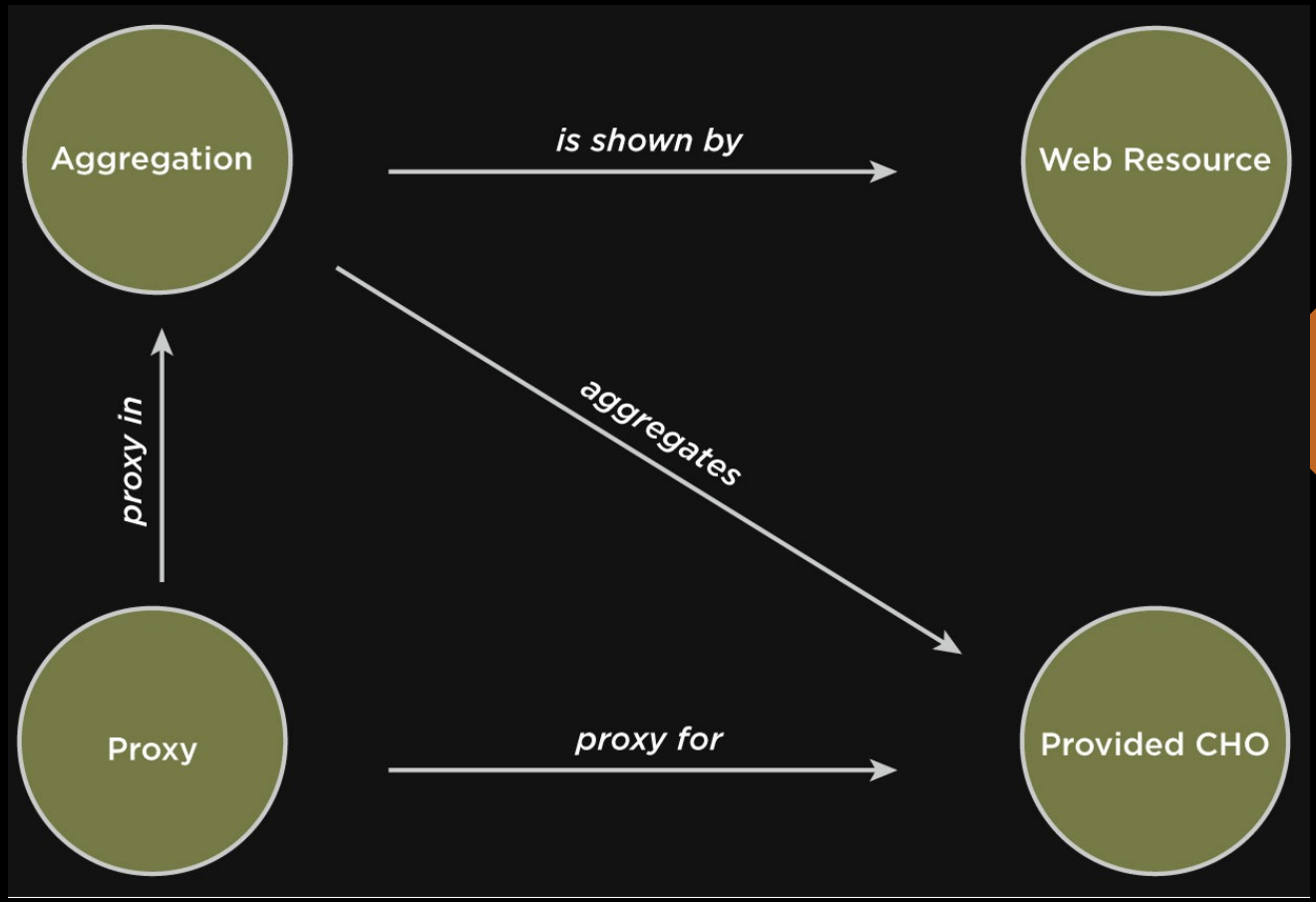
EUROPEANA DATA MODEL (EDM)

well-documented. secondary literature. handles the variety of collections and object types encountered in a cultural heritage repository. extends oai-ore. recursive.

The challenge

Pick a complex intellectual object in the digital repository to model--a serial title--and see whether one can apply *all required elements* specified by EDM. If one can do this, one should be able to model less complex objects. See also whether one can reuse existing data elements to avoid using any not already defined by others.

Modelling the issue



ProvidedCHO (highlights)

dc:title and/or dc:description are required.
dc:title “University of Chicago Record";

Link to the plain-text OCR for the issue.
dc:description <.../mvol-[NNNN]-[MMMM]-[PPPP].txt>;

A part is also a providedCHO (consider a page in an art book
used as a teaching resource in its own right, for example).
dcterms:hasPart <[NOID]/[URI for providedCHO]/00000001>;
dcterms:hasPart <[NOID]/[URI for providedCHO]/00000002>;

WebResource (highlights)

dc:format "application/pdf";

premis:objectIdentifierType "ARK”;
premis:messageDigestAlgorithm "SHA-256";
premis:messageDigest "4f6237c25a51382c3f6c489 …";
premis:messageDigestOriginator "/sbin/sha256";
premis:size 31011220;
premis:formatName "application/pdf";
premis:eventType "creation";
premis:eventDateTime "[ISO 8601]"^^xsd:dateTime;

Aggregation (highlights)

edm:aggregatedCHO [URI for the providedCHO]

a website

edm:isShownAt <http://pi.lib.uchicago.edu/[persistent link]>;

a PDF file

edm:isShownBy <.../mvol-[NNNN]-[MMMM]-[PPPP].pdf>;

a thumbnail

edm:object <.../00000001.jpg>;

Proxy

For the provided MARC record

<x0971s4d8g8wb/Maps/Chi1890/G4104-C6P33-1897-B536/G4104-C6P33-1897-B536.mrc>

dc:format "application/marc";

ore:proxyFor <x0971s4d8g8wb/Maps/Chi1890/G4104-C6P33-1897-B536>;

ore:proxyIn <x0971s4d8g8wb/aggregation/Maps/Chi1890/G4104-C6P33-1897-B536>;

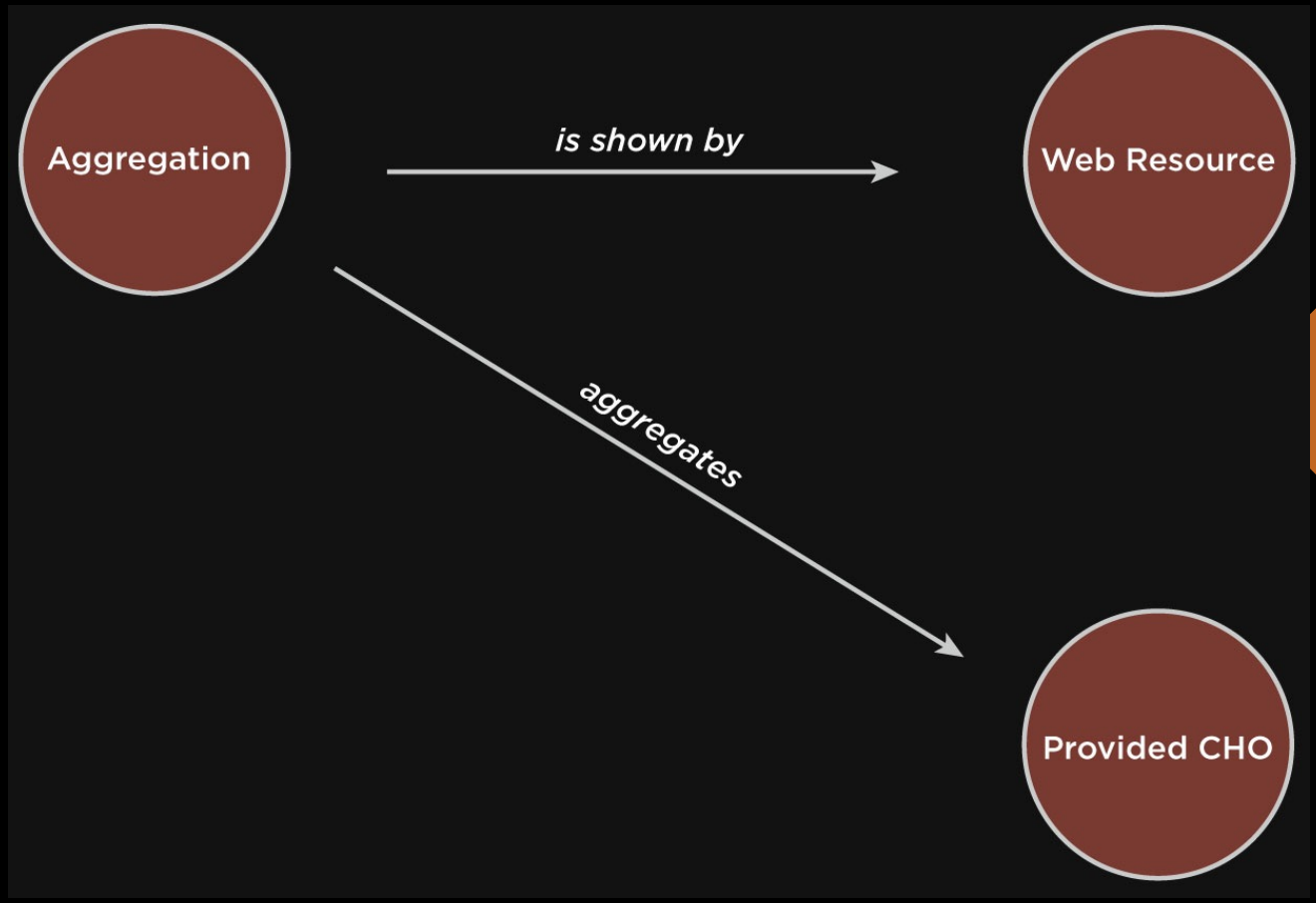
a ore:Proxy.

Recapitulation

ore:Aggregation	Required in EDM
edm:ProvidedCHO	Required in EDM
edm:WebResource	Required in EDM
ore:Proxy	Optional in EDM

Europeana also models Agent, Place, TimeSpan and Concept "to allow these entities to be modelled as separate entities from the CHO with their own properties if the data can support such treatment."

Modelling the Page Object



ProvidedCHO for first page object (highlight)

dc:description <.../[URI for OCR].xml>

For a page object, the dc:description is a file of OCR for the page which is structured as XML. Words are accompanied by coordinates, which allows software which supports this functionality to draw a bounding box around a search term showing where on the page image it is located.

Structured OCR example

<line l="109" t="494" r="240" b="503" spacing="37 5 60 5 24">Edward McCormick Blair</line>

t = top

b = bottom

l = left

r = right

l + spacing = r

ProvidedCHO for second page object (highlights)

dc:description <.../[URI for OCR].xml>

dc:title "Page 1";

edm:isNextInSequence <[URI for preceding page object]>;

WebResource for a digital masterfile (highlights)

dc:format "image/tiff";

mix:imageWidth 2208;
mix:imageHeight 2688;

premis:eventDateTime "[ISO 8601]"^^xsd:dateTime;

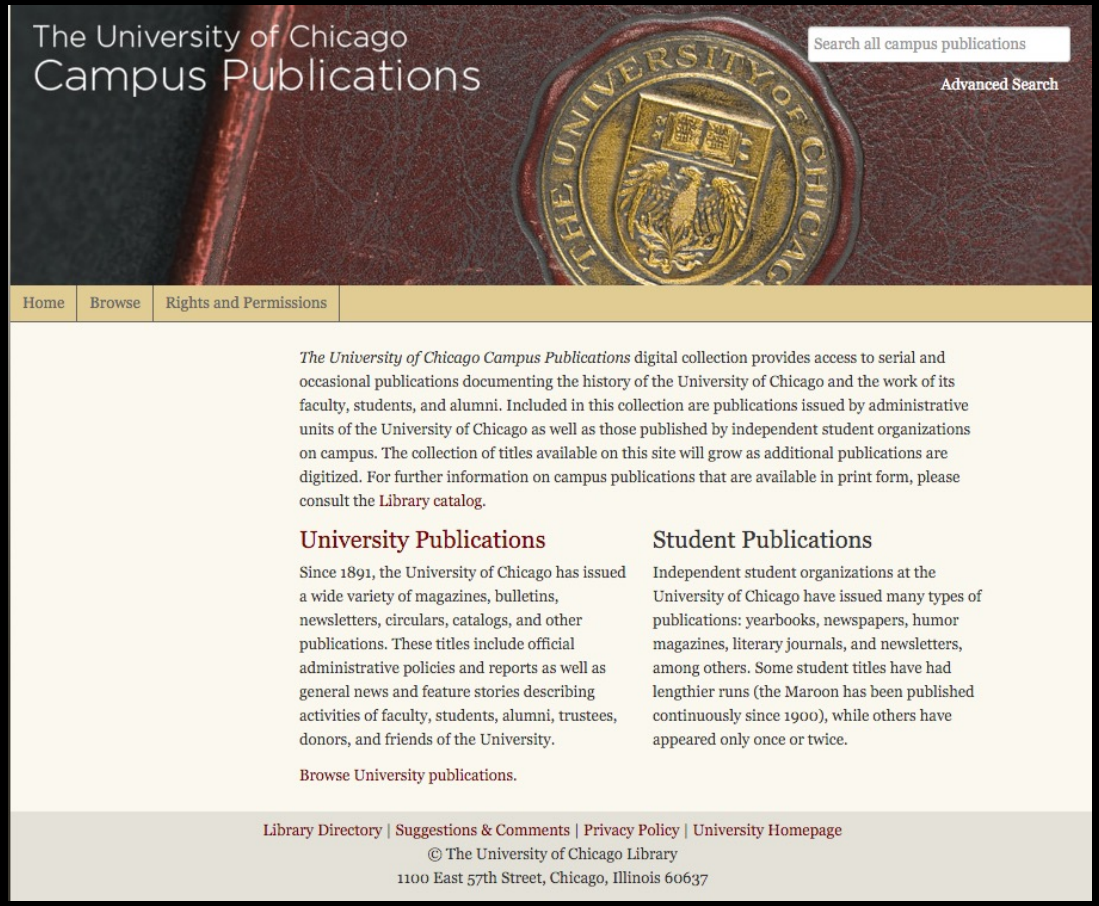
Aggregation (highlights)

edm:aggregatedCHO [URI for the providedCHO]

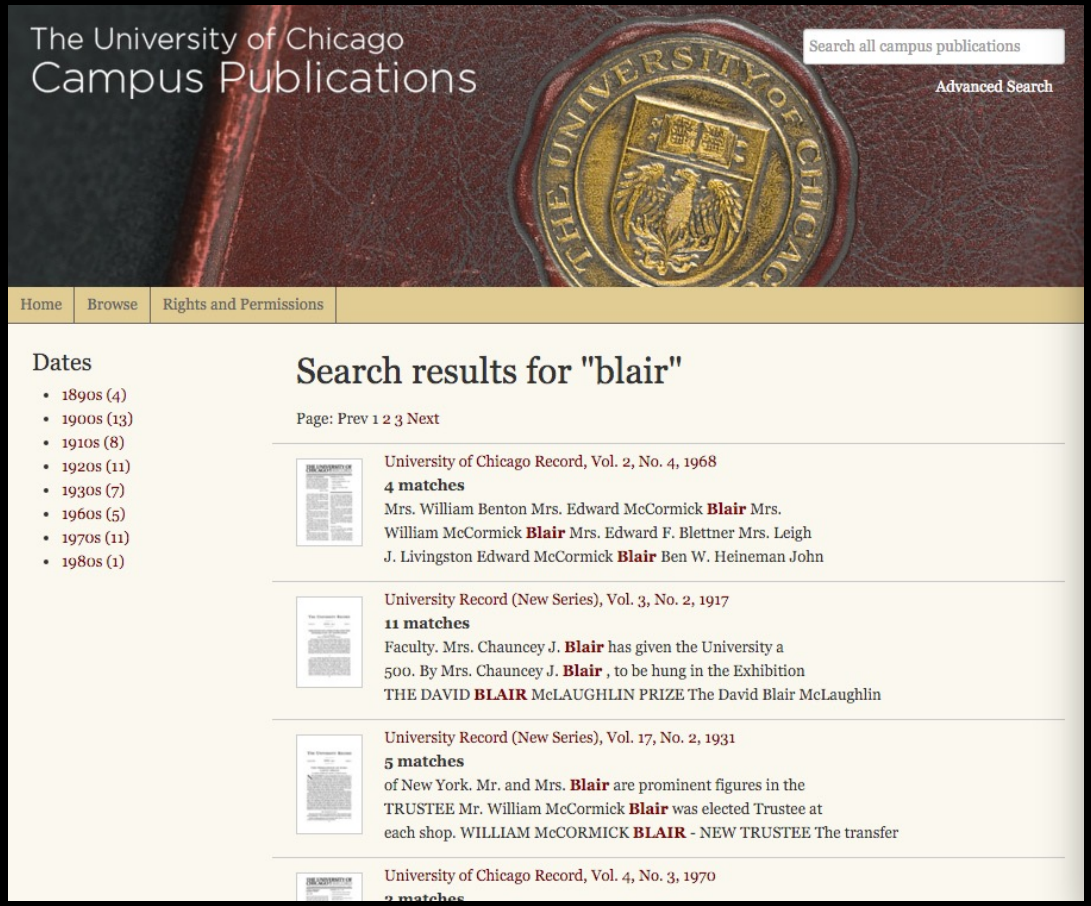
The page object is shown by the digital masterfile
edm:isShownBy <.../mvol-0007-0013-0001_0001.tif>;

The derivative access copy of the tiff image.
edm:object <.../mvol-0007-0013-0001_0001.jpg>;

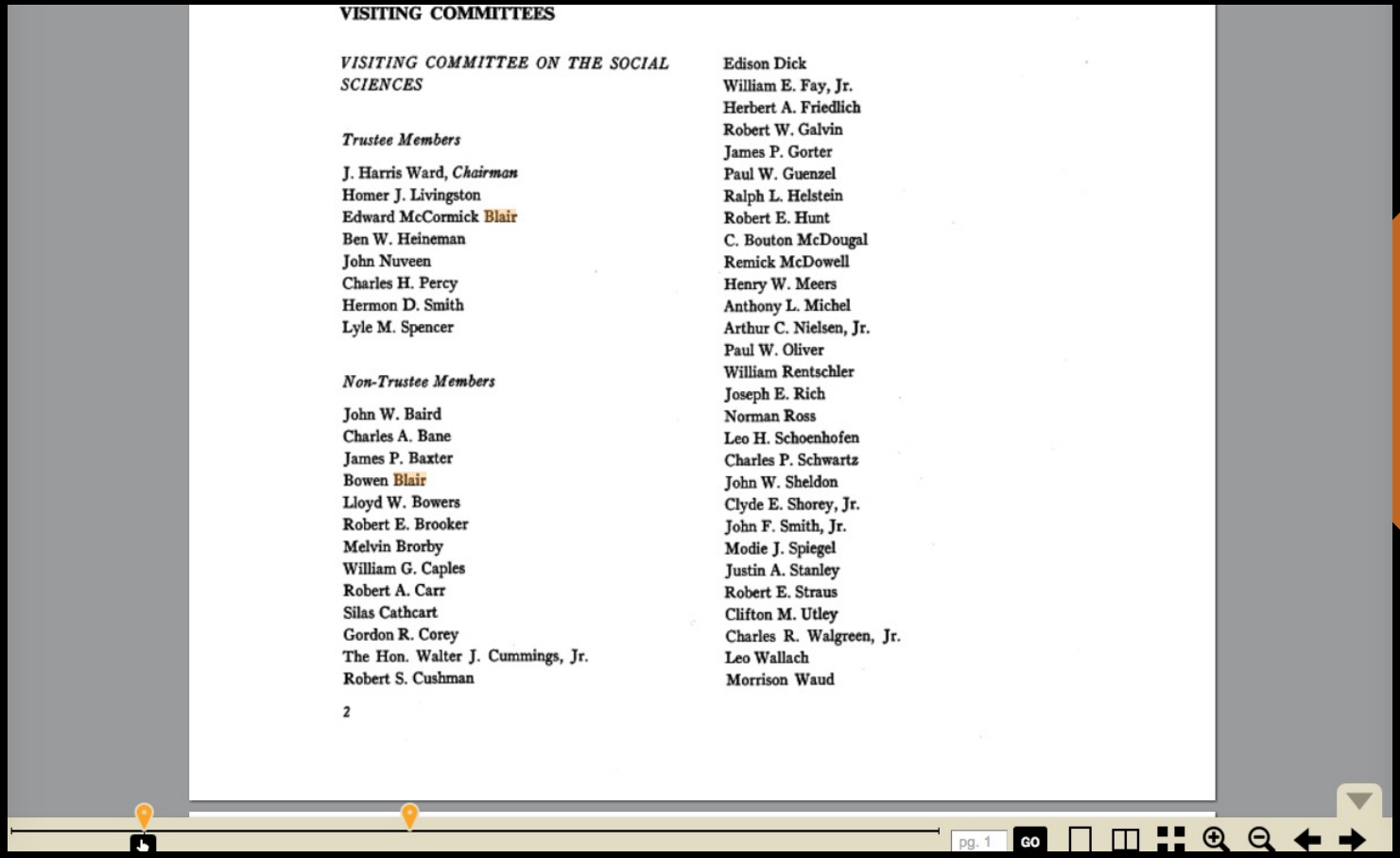
How have we used this?



Search for *blair*



Blair is highlighted on the page



VISITING COMMITTEES

VISITING COMMITTEE ON THE SOCIAL SCIENCES

Trustee Members

J. Harris Ward, *Chairman*
Homer J. Livingston
Edward McCormick Blair
Ben W. Heineman
John Nuveen
Charles H. Percy
Hermon D. Smith
Lyle M. Spencer

Non-Trustee Members

John W. Baird
Charles A. Bane
James P. Baxter
Bowen Blair
Lloyd W. Bowers
Robert E. Brooker
Melvin Brorby
William G. Caples
Robert A. Carr
Silas Cathcart
Gordon R. Corey
The Hon. Walter J. Cummings, Jr.
Robert S. Cushman

Edison Dick
William E. Fay, Jr.
Herbert A. Friedlich
Robert W. Galvin
James P. Gorter
Paul W. Guenzel
Ralph L. Helstein
Robert E. Hunt
C. Bouton McDougal
Remick McDowell
Henry W. Meers
Anthony L. Michel
Arthur C. Nielsen, Jr.
Paul W. Oliver
William Rentschler
Joseph E. Rich
Norman Ross
Leo H. Schoenhofen
Charles P. Schwartz
John W. Sheldon
Clyde E. Shorey, Jr.
John F. Smith, Jr.
Modie J. Spiegel
Justin A. Stanley
Robert E. Straus
Clifton M. Utley
Charles R. Walgreen, Jr.
Leo Wallach
Morrison Waud

2

Note the bounding box around the name

How does this work?

We generate DIPs from the RDF triplestore by means of SPARQL queries.

A SPARQL query (fragment)

```
select ?tiff ?width ?height

from <http://lib.uchicago.edu/campub>
where { ?tiff dc:format "image/tiff" .
        ?tiff mix:imageWidth ?width .
        ?tiff mix:imageHeight ?height .
        ?tiff a edm:WebResource }
```

Fragment of a DIP (XML)

```
<result>
  <binding name="tiff">
   <uri>http://ark.lib.uchicago.edu/ark:/61001/[path to tiff image]</uri>
  </binding>
  <binding name="width">
   <literal datatype="http://www.w3.org/2001/XMLSchema#integer">4384</literal>
  </binding>
  <binding name="height">
   <literal datatype="http://www.w3.org/2001/XMLSchema#integer">5376</literal>
 </binding>
</result>
```

Bounding box

In order to create the outlines of the bounding box correctly from the information in the file of OCR, we need to know the dimensions of the original TIFF image, since the coordinates are specified with reference to it, not the derivative image. *All we need to extract from the repository are the technical metadata for height and width, not the TIFF image itself*.

DIP DIP DIP DIP DIP (fragment)

Dip dip dip dip dip dip dip dip
Mum mum mum mum mum mum
Get a job
Sha na na na - sha na na na na

Another dissemination use case

Suppose I want all scores added to the Chopin Early Editions collection since the last time I made this request.

Another SPARQL query (fragment)

```
select ?score ?masterfile
from <http://lib.uchicago.edu/chopin>
where {
    ?aggregation4score edm:aggregatedCHO ?score .
    ?score dcterms:hasPart ?page .
    ?aggregation4page edm:aggregatedCHO ?page .
    ?aggregation4page edm:isShownBy ?masterfile .
    ?masterfile dc:format "image/tiff" .
    ?masterfile premis:eventDateTime ?date .
    filter (?date >= "2014-02-04T00:00:00"^^xs:dateTime) .
    ?masterfile a edm:WebResource
}
```

DIPs redux

“μήτε πλεονάζει μήτε ἐλλείπη”

Aristotle, *Ethica Nicomachea,* II. 5. 1106a 31-32

“se deve buscar lo preciso, y huir de lo superfluo”

Juan Antonio de Arrieta Arandia y Morentín, 1688

Processing redux

Archivists want to be able to leverage the accessions database to help them automate the production of the inventory portion of a finding aid. Once they add the descriptive elements and finish archival processing, *we can use the resulting EAD markup to generate linked data according to the Europeana data model*. How do we know we can do this?

The literature shows us how

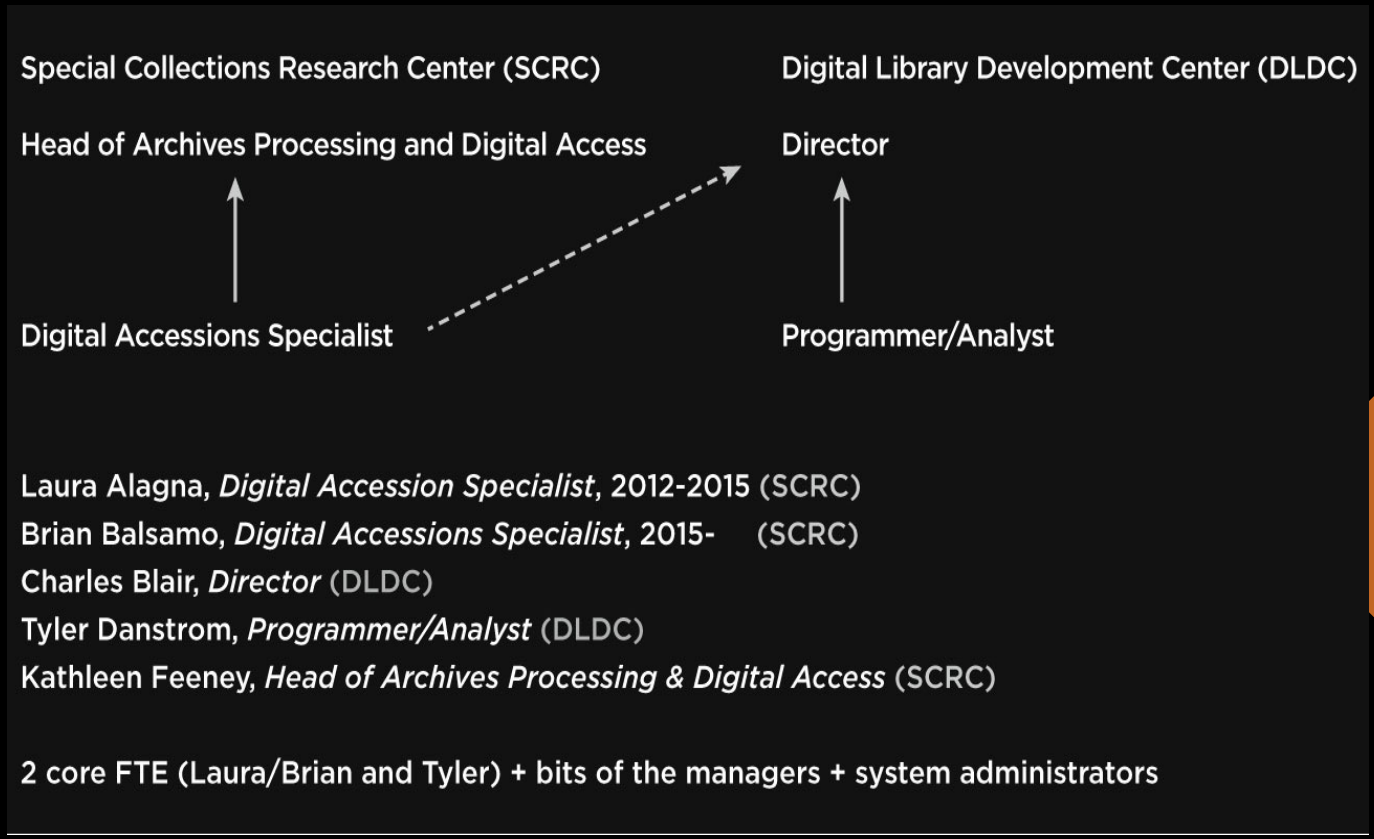
Casarosa, Vittore; Meghini, Carlo; Gardasevic, Stanislava. (2013). "Improving Online Access to Archival Data". Digital Libraries & Archives, pp. 153-162.

Gardasevic, Stanislava. (2011). "Opening Archives to the General Public, a data modelling approach". Master thesis. International Master in Digital Library Learning.

Hennicke, Steffen; Olensky, Marlies; de Boer, Victor; Isaac, Antoine; Wielemaker, Jan. (2011). "Conversion of EAD into EDM Linked Data". In: Proceedings of the 1st International Workshop on Semantic Digital Archives.
<http://www-e.uni-magdeburg.de/predoiu/sda2011/sda2011_06.pdf>.

Concluding thoughts

Credits



Laura Alagna, *Digital Accession Specialist*, 2012-2015 (SCRC)
Brian Balsamo, *Digital Accessions Specialist*, 2015- (SCRC)
Charles Blair, *Director* (DLDC)
Tyler Danstrom, *Programmer/Analyst* (DLDC)
Kathleen Feeney, *Head of Archives Processing & Digital Access* (SCRC)

2 core FTE (Laura/Brian and Tyler) + bits of the managers + system administrators

Vector graphics by Kathy Zadrozny. *Get a Job* – The Silhouettes – 1957.
Presentation by chas@uchicago.edu