Archives & Linked Open Data

are our tools ready to "complete the picture"?

Anila Angjeli
Bibliothèque nationale de France

Some observations from a metadata and standards specialist, a librarian collaborating with archivists to interconnect our data

#SAA12, session 401 - Linking Data Across Libraries, Archives, and Museums
Whether texts, images, websites, etc.

No matter what information you have

Everything is interconnected in a network of interrelated entities

1. You are dealing with real things
2. You only have part of the picture
3. Your objects live in the graph

"Bringing History Alive : Telling stories with Linked Data and open source tools - Australian War Memorial"

by David Peterson, at Drupal Down Under 13-15 January 2012, Melbourne

Collective history building, in action

Pieces of "stories"

Pinned to places

Pieced together

Available to others for reuse

to build new perspectives:

- social,
- historical,
- ethnological,
- …

“Records” from institutions meet those from individuals

http://www.historypin.com
Archival data

• Huge amount of historical records
  – providing documentary evidence of all domains of human activity
    (people, their lives and activities, places, events, intellectual creations, artifacts, …)
  – held in repositories, curated in organized collections
  – preserved in shelves, boxes, folders, digital files

• Represent interrelated pieces of historical evidence
  … as interrelated were the events and people that gave birth to them

• Challenge: reveal those interrelations!
Pro-active approach to access

• Reveal the existence of the records, describe
  • based on archival community standards
    – Content: ISAD(G), ISAAR(CPF), etc.
    – Communication structures: EAD, EAC-CPF, etc.
    – Other "knowledge organization systems": thesauri, etc.

• Make the Finding Aids accessible
  – available on the Internet
  – easily searchable

• Today the mission of access - extended
  – Reveal hidden treasures – inform about the context – highlight implicit connections – "actionable" links to other resources
Keep pace with the changing context

An ever-broadening vision of access to archival records and their interrelation with other resources

• Archival standards (both content and communication ones) are likely to change
• EAC-CPF – released, 2010
• EAD – under revision
• ICA content standards – process underway
a SAA communication standard for describing "people"
A step forward towards archival information as Linked Open Data

- While ISAAR (CPF) focuses primarily on the notion of provenance
- EAC-CPF applies to all "people" documented in the records (contextual information essential to the understanding of records)
- Two focal points:
  - Identity resolution (accurately identify “people”)
  - Make explicit relationships (key dimension of the context)
- Data centric
- Identifiers
- Other name spaces
- Linking mechanisms
For interoperability purposes,
  – title, date, author, etc.
  are duplicated by RDF terms such as
  – Dublin Core, FOAF, etc.
• use of **owl:sameAs**
  for linking to external resources

"EAC-CPF standard is quite different from its relative EAD widely used in archival description, especially because EAC-CPF standard has not the concept of "inheritance" and is more "data-centric" than EAD."

Developed by:
Silvia Mazzini - [Regesta.exe](http://archivi.ibc.regione.emilia-romagna.it/ontology/reference_document/referencedocument.html)
instances expressed in properties and subproperties to encode records and show them as Linked Open Data.
as Linked Open Data - live

HTML5 microdata in chronlist

RDFa owl:sameAs

title="Virtual International Authority File"

Also experimenting with italian RDF via xslt

B. Tingle: http://socialarchive.iath.virginia.edu/NAAC_meeting1_slides.html
Lafayette, Marie Joseph Paul Yves Roch Gilbert du Motier, DLC

(1757 – 1834) Male France French

→ Alternative forms of name

### Related Entries

- Archival Collections (21)
- People (61)
- Corporate Bodies (5)
- Resources (102)

- Linked Data (1)

sameAs

http://viaf.org/viaf/22144104
Prosopography and Linked Data

(from a post by Ethan Gruber (American Numismatic Society) recently on the EAD list)

EG: "... growing interest in ancient prosopography among classicists. Several participants are now working together to build a network of Roman emperors and their relations. ..."

"Having straddled both the museum and library/archival worlds, it makes sense to me to use EAC-CPF for this endeavor."

XForms-based information management system for EAC-CPF records called xEAC (pronounced "zeke") – ongoing, beta version
Simple public interface

Documentation:
http://wiki.numismatics.org/xeac:xeac

Code:
https://github.com/ewg118/xEAC

Features
- Create/edit/validate EAC Records
- Integration with Geonames, Nomisma, and Pleiades for modern and ancient place names
http://wiki.numismatics.org/xeac:placeentry
- Import names and associated authorities from VIAF
http://wiki.numismatics.org/xeac:viaf-import
- Easily maintain links between records and create record stubs
http://wiki.numismatics.org/xeac:relations
- Generic installation instructions
http://wiki.numismatics.org/xeac:generic_installation
... all is about links

complete the "picture" – interrelate - rebuild the history

quote from Ethan Gruber’s post:

"From a numismatic standpoint, I plan to:

- link emperors together with their families,
- link emperors to provincial governors,
- link these people to mints (corporate entities), and
- link mints to anonymous die carvers which are identifiable by their artistic style.

Each of these entities can be

- linked to affiliated coins to start, but we can eventually
- associate sculpture, epigraphy, and literary sources to them."
Archives
not a self-contained world

- Proliferating projects with archival datasets
- Applying Semantic Web and Linked Open Data technologies

examples:
  - Locah, Agrippa, Trove, Australian War memorial, HADOC

Show that:
- A lot can be done to inter-relate the existing archival data with other data from other cultural heritage domains
- By using the existing tools
- By remodeling the existing datasets
Interconnecting French cultural heritage treasures on the Web

- Digital documents (DC)
- Other BnF resources
- External resources
- Web pages for Internet users
- Raw data for machines
- Raw data for machines
- BnF Main catalogue (MARC)
- BnF Archives and Manuscripts catalogue (EAD)
- Modeling
- Matching
- Clustering
- Alignments
- Semantic Web technologies

heterogeneous resources

External resources
Modeling: an FRBR inspired ontology a choice: perspective appropriate to library materials

Pages for "authors" collocate/reorganize heterogeneous resources by type of contribution

Pages for "works" collocate/reorganize heterogeneous resources by "FRBR" level
Key issues

• Expose only the relevant elements to create meaningful clusters of data (FRBR: W-E-M-I) → “re-contextualize”

• EAD Finding Aid perspective – difficult to explore as such
  – Different FA (MS, performing arts, literary archives, etc.) → different EAD implementations
  – Difficult to devise common rules
Key issues

• Id perspective
  – anchors: call numbers of units
  – URIs of the authority records (ARK identifiers) occurring in the FA
  – URI of the FA

• For "authors":
  – trace authority URIs occurring in the FA
  – crawl upwards the hierarchy of the XML tree
  – inference rules to harvest the appropriate information at the higher levels
  – combine information and re-create inheritance
Émile Zola (1840-1902)

Author page
"build the story"

Compiled from different levels of EAD hierarchy
"heritage context rebuilt"

Links to BnF authority files

Unit call number

URL of FA
data.bnf.fr in the RDF file

- `<rdf:Description rdf:about="http://archivesetmanuscrits.bnf.fr/ead.html?id=FRBNFEAD000004233">
  < rdf:type rdf:resource="http://rdvocab.info/uri/schema/FRBRentitiesRDA/Manifestation"/>
  < bnf-onto:cote>Ms -104</ bnf-onto:cote>
  < bnf-onto:cote>Mn 37</ bnf-onto:cote>
  < bnf-onto:cote>Mn 5/194 (1-2)</ bnf-onto:cote>
  < dc:date>1878</ dc:date>
  < dc:date>1890</ dc:date>
  < rdfs:seeAlso rdf:resource="http://archivesetmanuscrits.bnf.fr/ead.html?id=FRBNFEAD000004233"/>
  < dc:title>
  </ dc:title>
</rdf:Description>

- `<rdf:Description rdf:about="http://archivesetmanuscrits.bnf.fr/ead.html?id=FRBNFEAD000004233#frbr:Expression">
  < bnfroles:r110 rdf:resource="http://data.bnf.fr/ark:/12148/cb119297200#foaf:Person"/>
  < marcrel:aut rdf:resource="http://data.bnf.fr/ark:/12148/cb119297200#foaf:Person"/>
  < dc:contributor rdf:resource="http://data.bnf.fr/ark:/12148/cb12276039q#foaf:Person"/>
  < dc:contributor rdf:resource="http://data.bnf.fr/ark:/12148/cb119297200#foaf:Person"/>
</rdf:Description>

Relator code: author of letters

URI of Authority record for Zola

URI of FA

Identification of level in data.bnf.fr

Unit call number
Subject headings as Linked Open data

Alignments with other vocabularies in Linked Open Data

Thesaurus for indexing local archives
Standards, vocabularies, ontologies, …

a conglomeration of tools

• Technological standards
  – Web standards, e.g. RDF, OWL, SKOS, SPARQL

• Content conveying, broadly used vocabularies
  Developed by
  – cross domain professional initiatives, driven by metadata interoperability aspiration
    • e.g. Dublin Core (DC Terms)
  – special interest professional initiatives
    • e.g. Geonames, Bio, RDA Vocabularies
  – movements for creation of social networks on the web (“non-professionals”)
    • e.g. FOAF
What about the specific archival concepts?

- Are the existing tools/vocabularies/ontologies enough
  - to correctly bring the archival datasets in the LOD cloud?
  - to interconnect resources appropriately, and not distort “the story”?

In the experiences with LOD

- Each of the projects has created its own data model (ontology)
  - Local (LOCAH, data.bnf.fr)
  - International cross-domain (EDM- Europeana Data Model)

- Other “bottom-up” experiments
  - Archival ontology
Archival Ontology
a work in progress

An OWL/RDF vocabulary for describing archival collections and the names associated with them.

This vocabulary is a work in progress and is currently being used as a catch-all for creating classes and properties for concepts and relationships that don't exist in other vocabularies.

The vocabulary URI is http://purl.org/archival/vocab/arch#
Contents
- arch
- Classes (8)
- Object Properties (15)
- Annotation Properties (4)
- Datatypes (2)

Annotations (3)
- creator "Aaron Rubinstein" ()
- description "An RDF vocabulary for describing archival collections and the names associated with them" ()
- title "Archival" ()

References
- Classes (8)
- Object Properties (15)
- Annotation Properties (4)
- Datatypes (2)

arch: classes (10)
- Thing
- Agent
- Archive
- Collection
- Family
- Group
- hasProvenance
- heldBy
- holds
- inclusiveEnd
- inclusiveStart
- knows
- Manuscript
- Nothing
- Organization
- Person
- PlainLiteral
- primaryProvenanceOf
- referencedIn
- relatedSubject
- string
- subject
- title
An archival high level ontology, a conceptual model?

Why?

- to address correctly the key concepts of the domain, the interrelationships among objects
- to help in building good systems
- to give to others the key of understanding archival data
- to enable them to use the data in appropriate ways, without distortion
- The revision of EAD is raising such issues
By whom?

Common agreement among the archivists that the best professional forum is

ICA/CBPS is now taking up the challenge of building an archival conceptual model

Two preliminary reports are published

- *Progress report for revising and harmonising ICA descriptive standards*
- *Relationship in archival descriptive systems*

You know your world better than anyone?

Model it

and open your data to the world