PLAZI
taking care of freedom

Digital Object and Text Encoding Standards

Terry Catapano

2014-03-11

Digital Objects

- Aggregation of digital files and metadata
  - Images
    - TIFF
    - JPEG
    - Thumbnail
  - Metadata
    - Object
    - Technical
    - Rights

Standards

- METS (Metadata Object Transmission Standard)
  - http://loc.gov/mets
  - XML Schema
  - Namespaces
  - Metadata
    - Descriptive
    - Administrative
      - Technical
      - Provenance
      - Rights
    - Structural
      - Hierarchies of div’s + xlink

METS “External” Schema

- http://www.loc.gov/standards/mets/mets-extenders.html
- Descriptive: Dublin Core, MODS, EAD, etc…
- Administrative: PREMIS
  - http://www.loc.gov/standards/premis/
  - Object
  - Agent
  - Rights
  - Event

BagIt

- Lightweight packaging of digital content
- Similar to DarwinCore Archive
- Good Tool Support:
  - Libraries in Python, Ruby, Perl, Java, PHP

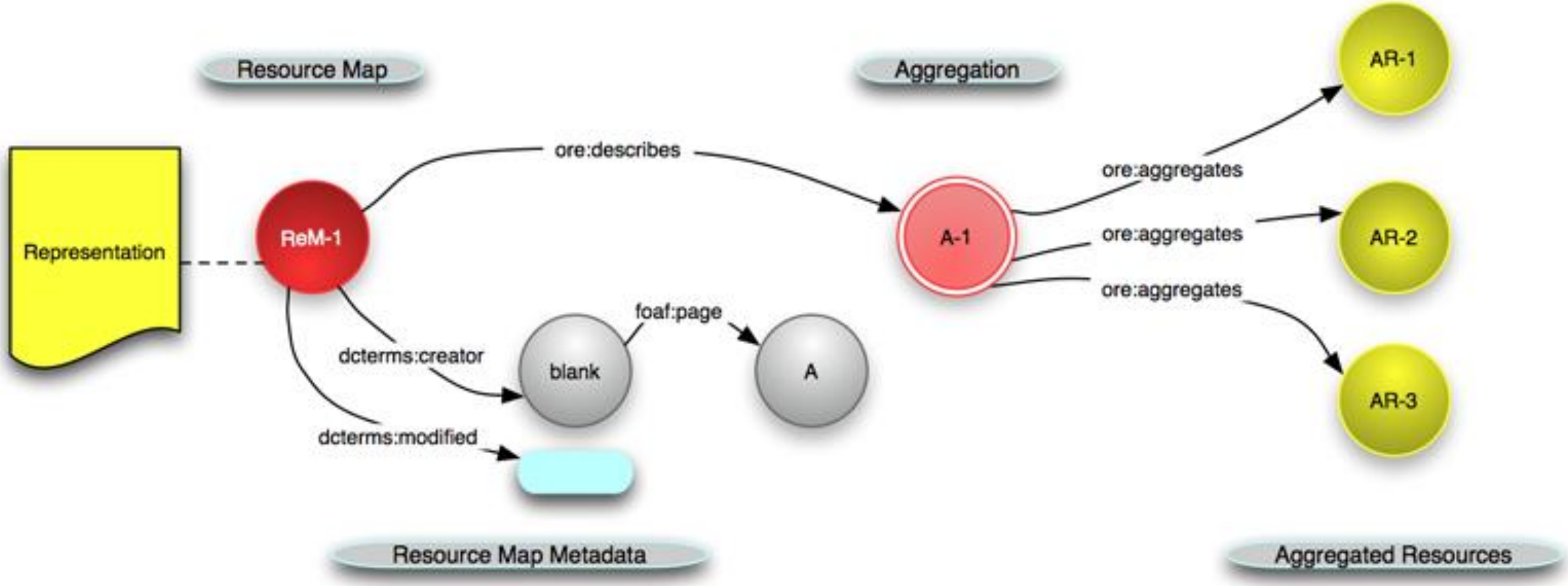
BagIt example

```
acnathist_edsfb_0800200C9A66_bag/
   manifest-md5.txt
                  49afbd86a1ca9f34b677a3f09655eae9 data/acnathist_edsfb_0800200C9A66/images/q172.png
                  408ad21d50cef31da4df6d9ed81b01a7 data/acnathist_edsfb_0800200C9A66/images/q172.txt
   bagit.txt
                  BagIt-version: 0.96
                  Tag-File-Character-Encoding: UTF-8
   bag-info.txt
                  Source-organization: ACME Natural History Museum
                  Organization-address: Box 350, Boston, MA 02134
                  OCR files of the field notebook of the entomologist Endora D. Stitz...
                  Bag-date: 2008-04-15
                  External-identifier: http://digital.acnathist.org/obj/8ADC1E30-A90C-11E3-A5E2-0800200C9A
   /data
                  acnathist_edsfb_0800200C9A66/images/q172.png
                  acnathist_edsfb_0800200C9A66/images/q172.txt
```

OAI-ORE

- Open Archives Initiative Object Reuse and Exchange
- RDF-based
- Aggregations of Web Objects
  - Aggregation
  - Resource Map
  - Aggregated Resources

OAI-ORE Basic Example



Tools and Systems

- File Characterization
  - JHOVE
  - Hashdeep
  - FIDO/PRONOM
  - FITS
- Archivematica: Digital Content Processing and Preservation
- Fedora/Islandora; Hydra

Text Encoding

- XML
- Text Encoding Initiative
  - http://tei-c.org
  - Guidelines for Text Encoding and Interchange
  - Customizable; TaxonX
- Journal Article Tag Set (JATS)
  - NLM/NCBI DTD
  - NISO Standard
  - Archiving, Publishing, Authoring Tag Sets
  - PubMed Central
- TaxPub
  - JATS: Generic Markup
  - Taxonomic Extensions
  - "Domain Schemas" (DWC) for fine grained semantics

TaxonX collection_event

```
<tax:collection_event>
        <tax:xmldata>
                <dwc:DecimalLongitude>49.7</dwc:DecimalLongitude>
                <dwc:DecimalLatitude>-12.97</dwc:DecimalLatitude>
                <dwc:Country>Madagascar</dwc:Country>
                <dwc:StateProvince>Foret
d'Ampondrabe</dwc:StateProvince>
                <dwc:Locality>Daraina</dwc:Locality>
                <dwc:YearCollected>2003</dwc:YearCollected>
                <dwc:MonthCollected>12</dwc:MonthCollected>
                <dwc:DayCollected>10</dwc:DayCollected>
                <dwc:Collector>B.L.Fisher</dwc:Collector>
    </tax:xmldata>Foret d'Ampondrabe , 26.3 km 10° NNE
Daraina</tax:collection_event> ;
<tax:collection_event>
```

TaxPub

- Taxonomic Treatments
- Treatment Sections
  - Nomenclature
  - Other (Description, Diagnosis, Etymology, etc…)
- Phrase Level
  - Taxon Names
  - Treatment Citations
  - *Material Citation*

TaxPub: material-citation

```
<tp:material-citation><named-content content-
type="dwc:typeStatus">Holotype</named-content>.
    Occurrence: catalogNumber: <named-content content-
type="dwc:catalogNumber"
        >SAM-DIP-A007146</named-content>; recordedBy:
<named-content content-type="dwc:recordedBy"
        >H. Brown</named-content>; sex: <named-content
content-type="dwc:sex">1
    male</named-content>; lifeStage: <named-content
content-type="dwc:lifeStage"
        >Adult</named-content>; otherCatalogNumbers:
<named-content
        content-type="dwc:otherCatalogNumbers">AAM-
000450</named-content></tp:material-citation>
```

Encoded Archival Context

- XML Schema
- Archival Authority Control
    - Corporate, Personal, Family Entities
- Strong features for relations to related materials and entities
- Social Networks and Archival Contexts Project
- http://socialarchive.iath.virginia.edu
- Data drawn from VIAF, APENet, Archivegrid
- http://socialarchive.iath.virginia.edu/xtf/view?docId=dickey-donald-cr.xml