

The Tap in the Humanities Data Pipeline: Disciplinary Integration of Institutional Data Repositories



Eric Kansa University of California Berkley School of Information
 Amy Barton Purdue University Libraries
 Sorin Matei Purdue University and Discovery Park Fellow

Abstract

Funder mandates, publisher expectations, and open data initiatives are motivating factors for researchers to identify data sharing infrastructures that will provide permanent access to and long-term preservation of their research data. In addition, there is great interest in library science and information science communities to use shared research data in other contexts. Where might institutional data repositories fit in this picture? A use case involves the Purdue University Research Repository (PURR). Research datasets produced through collaborative project spaces within PURR are published with a permanent Digital Object Identifier (DOI) and become available for discovery and citation world-wide. PURR is now developing capacity for other discipline content service providers to access PURR data to include in their content management systems. The use case involves two archaeological content systems, Open Context and Visible Past.

Open Context publishes archeological data contributed by researchers through an editorial review process and aligns data to standards, such as Open Linked Data. Visible Past is an interactive historical atlas enhanced with narratives, pictures, video clips, and audio files. It is also a site where researchers can collaboratively work to produce scholarly products, such as monographs and papers. Visible Past and Open Context are linked via Open Linked Data. Visible Past harvests records from Open Context to include in its content management system.

PURR is preparing to publish archeological datasets from a Rough Cilicia Archeological Survey, 2003 Pottery Sherd Study. Each dataset will include the image(s) of a pottery sherd and a MODS metadata file serialized in XML that contains all the data about the pottery sherd. The MODS file will be accessible for Open Context to harvest, programmatically manipulate, and map to their systems' CIDOC-CRM schemata. The processes to be implemented will enable Open Context to open PURR's downstream humanities data flow to Visible Past.

Methods

Open Context used ArcheoML to serialize object metadata in XML. ArcheoML is no longer maintained. Open Context switched to a conceptual reference model, CIDOC-CRM.

CIDOC-CRM Class Hierarchy

- E27 - Persistent Item
- E20 - Thing
- E22 - Legal Object
- E18 - Physical Thing
- E19 - Physical Object
- E20 - Biological Object
- E21 - Person
- E22 - Man-Made Object
- E84 - Information Carrier
- E24 - Physical Man-Made Thing
- E22 - Man-Made Object
- E84 - Information Carrier
- E25 - Man-Made Feature
- E28 - Collection
- E26 - Physical Feature
- E27 - Site
- E23 - Man-Made Feature

CIDOC-CRM Property Hierarchy

- P12 was intended use of (was made for) E7 Activity E71 Man-Made Thing
- P20 had specific purpose (was purpose of) E7 Activity E5 Event
- P21 had general purpose (was purpose of) E7 Activity E55 Type
- P24 transferred title of (changed ownership through) E8 Acquisition E18 Physical Thing
- P30 transferred custody of (custody transferred through) E10 Transfer of Custody E18 Physical Thing
- P43 has dimension (is dimension of) E20 Thing E54 Dimension
- P44 has condition (condition of) E18 Physical Thing E3 Condition State
- P45 consists of (is incorporated in) E18 Physical Thing E57 Material
- P46 is composed of (forms part of) E18 Physical Thing E28 Physical Thing
- P56 bears feature (is found on) E19 Physical Object E26 Physical Feature
- P49 has former or current keeper (is former or current keeper of) E18 Physical Thing E39 Actor
- P50 has current keeper (is current keeper of) E18 Physical Thing E39 Actor
- P51 has former or current owner (is former or current owner of) E18 Physical Thing E39 Actor
- P52 has current owner (is current owner of) E18 Physical Thing E39 Actor
- P53 has former or current location (is former or current location of) E18 Physical Thing E53 Place

Align PURR's pottery sherd data to CIDOC-CRM.

CIDOC-CRM → PURR Pottery Sherd Data Mapping

A	B	C	D
CIDOC-CRM Class	CIDOC-CRM Property	Pottery Sherd Data Value	Label Class
E55 Type	P2 has_type	Cooking Ware	Label Class
E22 Man-Made_Object	P53 has_current_or_former_location	Antioch / Map Number - 28-C-9-D / Site Name - Collection Area - 14	Location
(E22 Man-Made_Object → P43F has_dimension) E54 Dimension	(P91 has_unit - cm) P90 has_value	0.055	Preserved Height
(E22 Man-Made_Object → P43F has_dimension) E54 Dimension	(P91 has_unit - cm) P90 has_value	0.24	Estimated Diameter
E22 Man-Made_Object	P3 has_note	Rim	Estimated Diameter Note
E22 Man-Made_Object	P3 has_note	fabric: 5YR 8/2 pale yellow core: 2.5 YR 7/6 light red	Exterior/Surface Fabric (Munsell Color Measurement)
E22 Man-Made_Object	P3 has_note	Medium-hard, fine grained. With many small black inclusions, some small red inclusions, and few silver mica inclusions ranging from very small to medium sized.	Fabric Description
E22 Man-Made_Object	P3 has_note	Rim, handle and wall with carination. Ear handle attaches to rim and wall slightly below carination.	Sherd Description
(E22 Man-Made_Object → P4 has_time-span) E52 Time-Span	P82 at_some_time_within	500-699 AD	Time Period
E22 Man-Made_Object	P3 has_note	47 / A / 33	Object Identifier
E15 Identifier_Assignment	P58 has_preferred_identifier	<DOI Identifiers>	DOI
E13 Attribute_Assignment	P140 assigned_attribute	E22	DOI
E13 Attribute_Assignment	P141 assigned (was assigned by)	Nicholas K. Rauh	Data Producer

Methods

Need to create a machine-readable/actionable metadata file for ingestion in Open Context with PURR archeological data.

CIDOC-CRM → MODS Mapping

A	B	C
CIDOC-CRM Class	CIDOC-CRM Property	MODS
E22 Man-Made_Object	P30 has_title	<title display-label="Typology"></title>
E55 Type	P2 has_type	<genre display-label="Class"></genre>
E22 Man-Made_Object	P53 has_current_or_former_location	<location display-label="Site Name"></location>
E22 Man-Made_Object	P53 has_current_or_former_location	<physicalLocation type="Original Map Data"></physicalLocation>
E22 Man-Made_Object	P53 has_current_or_former_location	<physicalLocation type="Map Number"></physicalLocation>
E22 Man-Made_Object	P53 has_current_or_former_location	<physicalLocation type="Site Number"></physicalLocation>
E22 Man-Made_Object	P53 has_current_or_former_location	<physicalLocation type="Collection Area"></physicalLocation>
E22 Man-Made_Object	P91 has_unit - cm) P90 has_value	<physicalDescription display-label="Preserved Height"></physicalDescription>
**	**	<physicalDescription display-label="Preserved Height"></physicalDescription>
**	**	<physicalDescription display-label="Preserved Length"></physicalDescription>
**	**	<physicalDescription display-label="Preserved Length"></physicalDescription>

CIDOC-CRM → MODS Mapping

A	B	C
CIDOC-CRM Class	CIDOC-CRM Property	MODS
**	**	<physicalDescription display-label="Munsell Notes"></physicalDescription>
**	**	<physicalDescription display-label="Fabric Description"></physicalDescription>
**	**	<physicalDescription display-label="Sherd Description"></physicalDescription>
(E22 Man-Made_Object → P4 has_time-span) E52 Time-Span	P82 at_some_time_within	<subject display-label="Time Period"></subject>
E22 Man-Made_Object	P3 has_note	<identifier invalid="yes" display-label="Sherd Number"></identifier>
E15 Identifier_Assignment	P58 has_preferred_identifier	<identifier type="doi" display-label="DOI"></identifier>
E13 Attribute_Assignment	P140 assigned_attribute	<name type="personal"></name>
E13 Attribute_Assignment	P141 assigned (was assigned by)	<name type="personal" display-label="Data Producer"></name>

Result

PURR Archeological Data Serialized in MODS Metadata XML File

```
<?xml version="1.0" encoding="UTF-8"?>
<mods version="3.0" xmlns:xsi="http://www.w3.org/1999/xml" xmlns="http://www.loc.gov/mods/v3" xsi:schemaLocation="http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/mods-3.xsd">
  <recordInfo type="Bibliographic">
    <description type="Text">The Rough Cilicia Archeological Survey Project</description>
    <extension base="http://www.loc.gov/mods/v3" uri="http://www.loc.gov/mods/v3" display-label="Link to Dataset"/>
  </recordInfo>
  <title display-label="Typology">
    <title>West Cilicia Ring Foot Plate (3)</title>
  </title>
  <typeOfResource type="Image">
    <genre display-label="Class">General Common/Quasi-wares</genre>
    <note type="Text">The three dimensional object, the pottery sherd, is portrayed in a still image.</note>
  </typeOfResource>
  <location display-label="Site Name">
    <physicalLocation type="Original Map Data">28-b-21-b-2-27</physicalLocation>
    <physicalLocation type="Map Number">28-B-21-B</physicalLocation>
    <physicalLocation type="Site Number">4</physicalLocation>
    <physicalLocation type="Collection Area">2</physicalLocation>
  </location>
  <physicalDescription display-label="Preserved Height">
    <extent base="text">
      <start>0.028</start>
      <end>0.10</end>
    </extent base="text">
    <physicalDescription display-label="Estimated Diameter">
      <note type="Text">Munsell Measurement</note>
      <physicalDescription display-label="Exterior/Surface Fabric">
        <note type="Text">Fabric Description</note>
        <physicalDescription display-label="Fabric Description">
          <note type="Text">Medium-hard, fine grained, soft, some micaceous (very small) inclusions. Many small red inclusion. Some large red inclusion. Some large white inclusion.</note>
          <physicalDescription display-label="Sherd Description">
            <note type="Text">Rim, handle and wall with carination. Ear handle attaches to rim and wall slightly below carination. Inner surface of foot has small badly worn groove. Floor is lower than end of wall.</note>
          </physicalDescription>
          <identifier type="doi" display-label="DOI">DOI number here</identifier>
          <identifier type="url" display-label="Link to Dataset"></identifier>
          <name type="personal" display-label="Data Producer">
            <namePart>Nicholas K. Rauh</namePart>
            <affiliation>Purdue University</affiliation>
          </name>
        </mods>
      </physicalDescription>
    </physicalDescription>
  </physicalDescription>
</mods>
```

Conclusion



Future Directions

PURR currently makes descriptive Dublin Core metadata available via OAI-PMH. However, we are interested in exposing more than descriptive metadata. We will develop more robust mechanisms to make available serialized metadata for ingestion in service providers' content management systems.

Resources

- Purdue University Research Repository (PURR) <https://purrr.purdue.edu/>
- Open Context <http://opencontext.org/>
- Visible Past <http://visiblepast.net/see/>
- CIDOC-CRM <http://www.cidoc-crm.org/>
- MODS Metadata <http://www.loc.gov/standards/mods/>