ISO 16363 & OAI-PMH

PURDUE UNIVERSITY RESEARCH REPOSITORY

By Neal Harmeyer, Amy Hatfield, and Brandon Beatty

PURDUE UNIVERSITY LIBRARIES

PRESERVATION

BY NEAL HARMEYER

WHY PRESERVE?

- Scholarly research necessitates ability to refer back, or build upon, previous work—without preservation, this becomes impossible over time.
- Items accessible today are not guaranteed to be accessible tomorrow.
- Obsolescence, technology failures, disasters, etc. can damage or destroy—effective preservation mitigates that eventuality.

PURR DIGITAL PRESERVATION POLICY

- The PURR Digital Preservation Policy is a guiding document for the management of content within the repository.
- The Policy states that a focused attention to preservation is an “essential component of PURR services as it enables long-term access, and as such it requires attention throughout the data management process.”
- Development of long-term preservation strategies, strategic plans, and actions are taken from this foundational document.
    - The Libraries is committed to preserving and maintaining all PURR content for at least a period of ten years after it is published within the repository.

PURDUE UNIVERSITY LIBRARIES

FROM POLICY TO TRUSTWORTHINESS

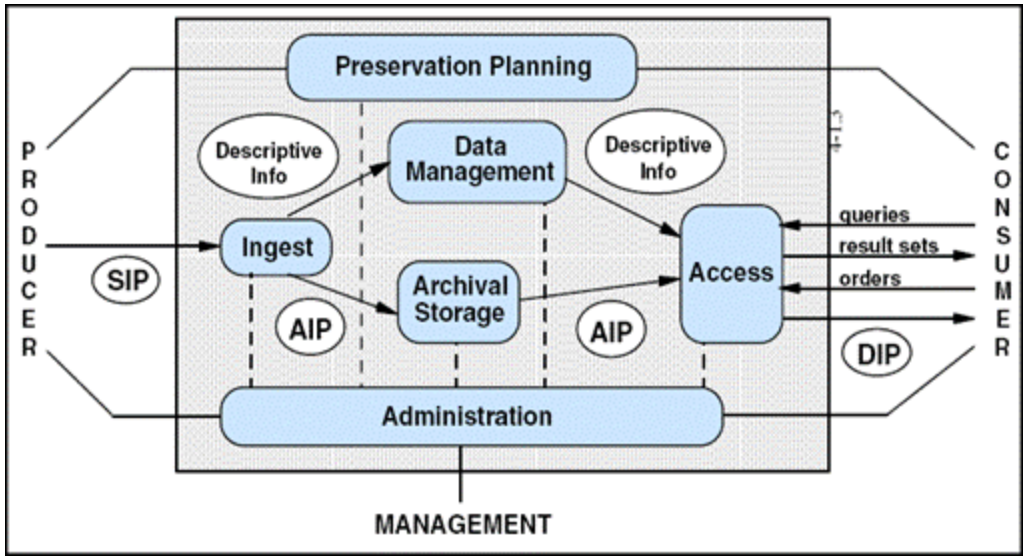
- Via the mandate of the PURR Digital Preservation Policy, a robust preservation system must be implemented.
- Stringent preservation planning should come from an internationally recognized standard.
- ISO 16363, Audit and Certification of Trustworthy Digital Repositories, provides metrics designed to establish a functional and reliable digital preservation environment.

DOCUMENTATION – PLAN AND STRATEGIES

- Preservation Strategic Plan
  - Lays out overall objectives
  - Lists imperative preservation activities
- Preservation Strategies
  - Determines specific strategies for preservation of digital objects
  - Lists preservation actions necessary for long-term preservation and access

DOCUMENTATION - OAIS MODEL

- The Open Archival Information System (OAIS) Reference Model is a standard in digital preservation.
- Various preservation planning aspects – ingest, data management, archival storage, and access – are modeled.
- The goal is to create a trustworthy system from producer to consumer.



PURDUE UNIVERSITY LIBRARIES

DOCUMENTATION – FIXITY AND FORMATS

- Digital objects must undergo fixity checks on a regular basis.
  - At ingest, a cryptographic hash is created for each object.
  - On a set schedule, the current hash is compared to preservation hash to check fixity.
- File formats must be determined and validated to ensure long-term preservation techniques are appropriately applied.
  - Files are checked against a format registry database at submission.
  - Preservation strategies and actions are determined by file format.
  - Formats are normalized to archival standards.
- As archival best practices change, preservation actions will change.

DOCUMENTATION – INFORMATION PACKAGES

- An information package is a group of digital objects within a preservation system.
- There are three types of information packages.
  - Submission Information Package (SIP)
    - Delivered by producer and initiated when user creates a project
    - Includes: digital object(s), descriptive information (metadata)
  - Archival Information Package (AIP)
    - Created from SIP
    - Contains digital object(s) and Preservation Descriptive Information (more metadata)
  - Dissemination Information Package (DIP)
    - Derived from AIP
    - Access piece for consumer upon request

PURDUE UNIVERSITY LIBRARIES

METADATA

Amy Hatfield, MLS PURR Metadata

Networks and Matrix Computations (nmcomp)

General project by David F Gleich

private project

Updates 1
Info
Team 2
Files 5
Publications 1
To-do
Notes

Manager options:
Edit project
Invite people to join
Learn about projects

Publications » dataset "Graph of Flickr Photo-Sharing Social Network Crawled in May 2006"

Versions » Version 1.0 (draft)

✓ Content | Description | ✓ Authors | x Gallery | x Tags | ✓ License | x Notes

Edit Description

Save changes

Fill in title, abstract and description sections below.

Title: REQUIRED

Graph of Flickr Photo-Sharing Social Network Crawled in May 2006

Synopsis: REQUIRED

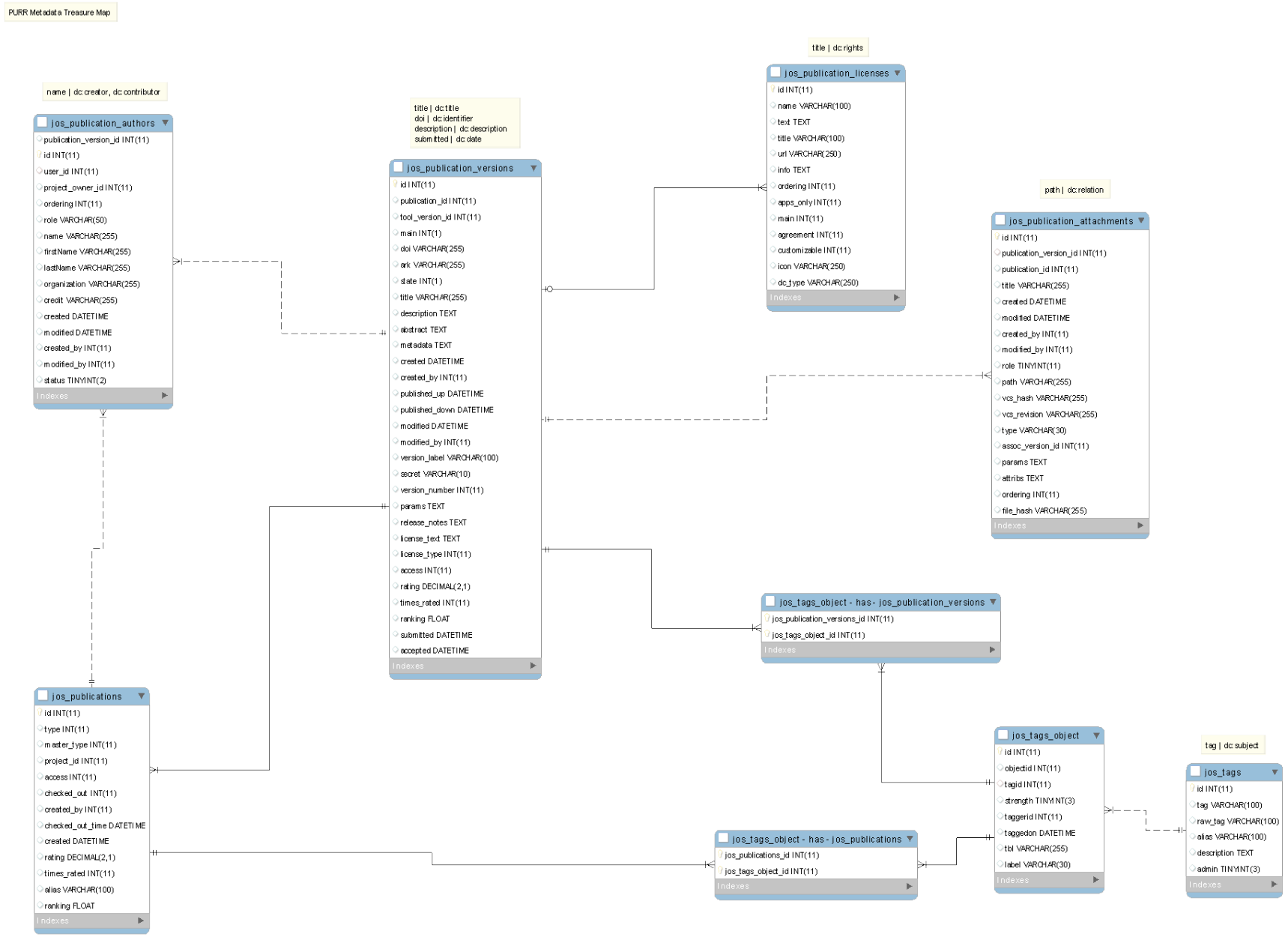
Graph of the Flickr photo-sharing social network crawled in May 2006 returning a graph with 820,878 nodes and 9,837,214 edges. Dataset is distributed as a TSMAT file with README file with code to read file in Python and MATLAB.

Abstract / Description: REQUIRED

Flickr is a popular online-community for sharing photos, with millions of users. This graph is representative of its social network, in which the node set V represents users, and the edge set E is such that (u, v) is in E if and only if a user u has added user v as his/her contact. We start with a crawl extracted from Flickr in May 2006. This crawl began with a single user and continued until the total personalized PageRank on the set of uncrawled nodes was less than 0.0001. The result of the crawl was a graph with 820,878 nodes and 9,837,214 edges.

Preview:

Flickr is a popular online-community for sharing photos, with millions of users. This graph is representative of its social network, in which the node set V represents users, and the edge set E is such that (u, v) is in E if and only if a user u has added user v as his/her contact. We start with a crawl extracted from Flickr in May 2006. This crawl began with a single user and continued until the total personalized PageRank on the set of uncrawled nodes was less than 0.0001. The result of the crawl was a graph with 820,878 nodes and 9,837,214 edges.



PURDUE UNIVERSITY LIBRARIES

Database population

DUBLIN CORE

Dublin Core Metadata Used By PURR

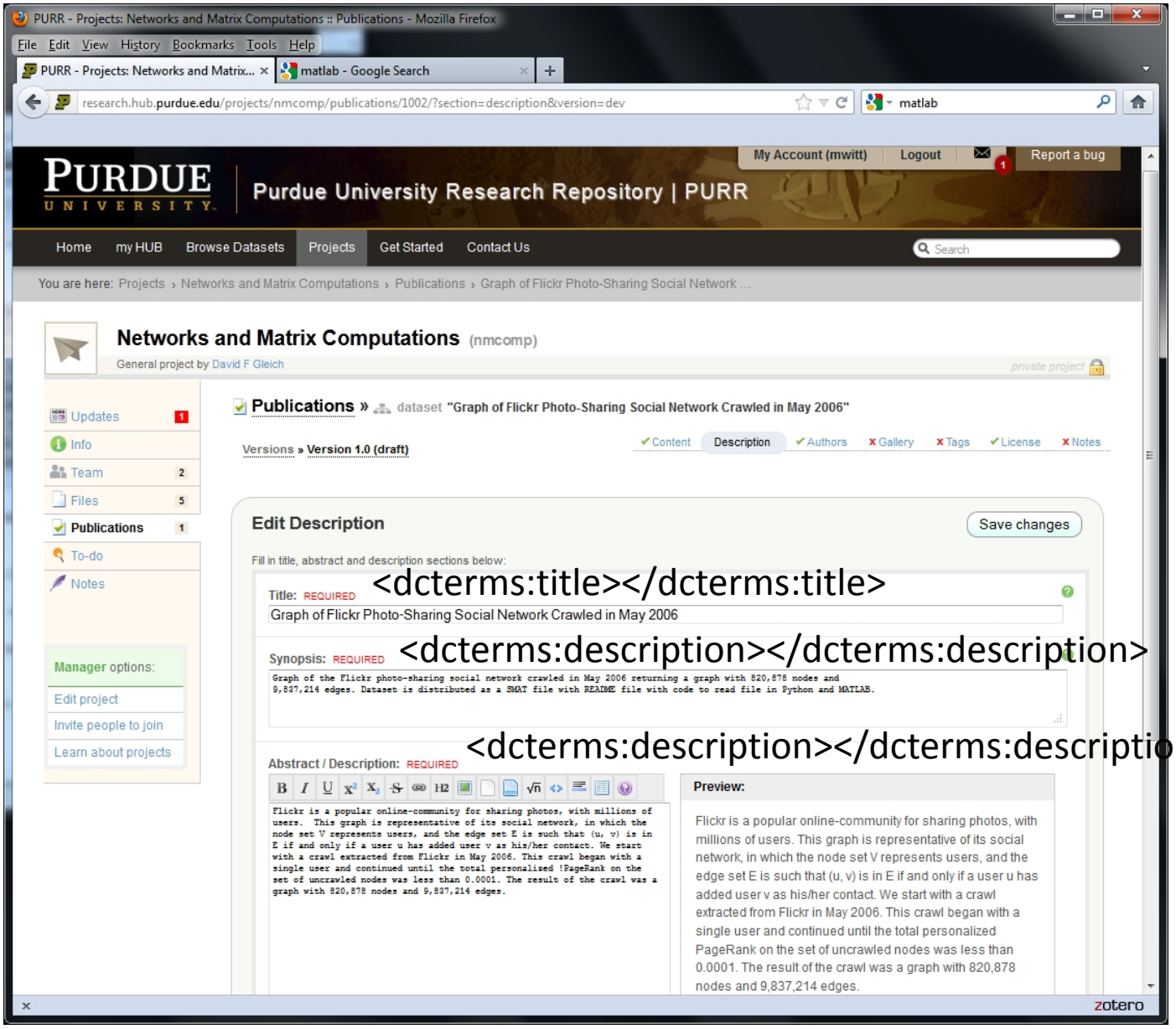
PURR	dcterms	Repeatable	Required	User-Contributed	System-Generated	Description
Author jos_publications	Creator (text string)		Yes	Yes		An entity primarily responsible for making the resource. **(This field is for listing the PRIMARY author or project lead.)**
Author jos_publications	Contributor (text string)	Yes		Yes		An entity responsible for making contributions to the resource. **(this field is for listing additional authors/project members)**
Created jos_publications_versions	Date (ISO:8601)		Yes		Yes - timestamp	A point or period of time associated with an event in the lifecycle of the resource.
abstract + synopsis + notes jos_publications_versions	Description (free text)	Yes		Yes		An account of the resource. **(Each PURR field will have a DC description field equating to 3 description fields.)**
Format jos_publications_master_types	Format (MIME type)		Yes		Yes -- always 'Bagit'	The file format, physical medium, or dimensions of the resource.
Identifier	Identifier		Yes		Yes - DOI minted at SIP publication	An unambiguous reference to the resource within a given context.

QUALIFIED DUBLIN CORE SCHEMA

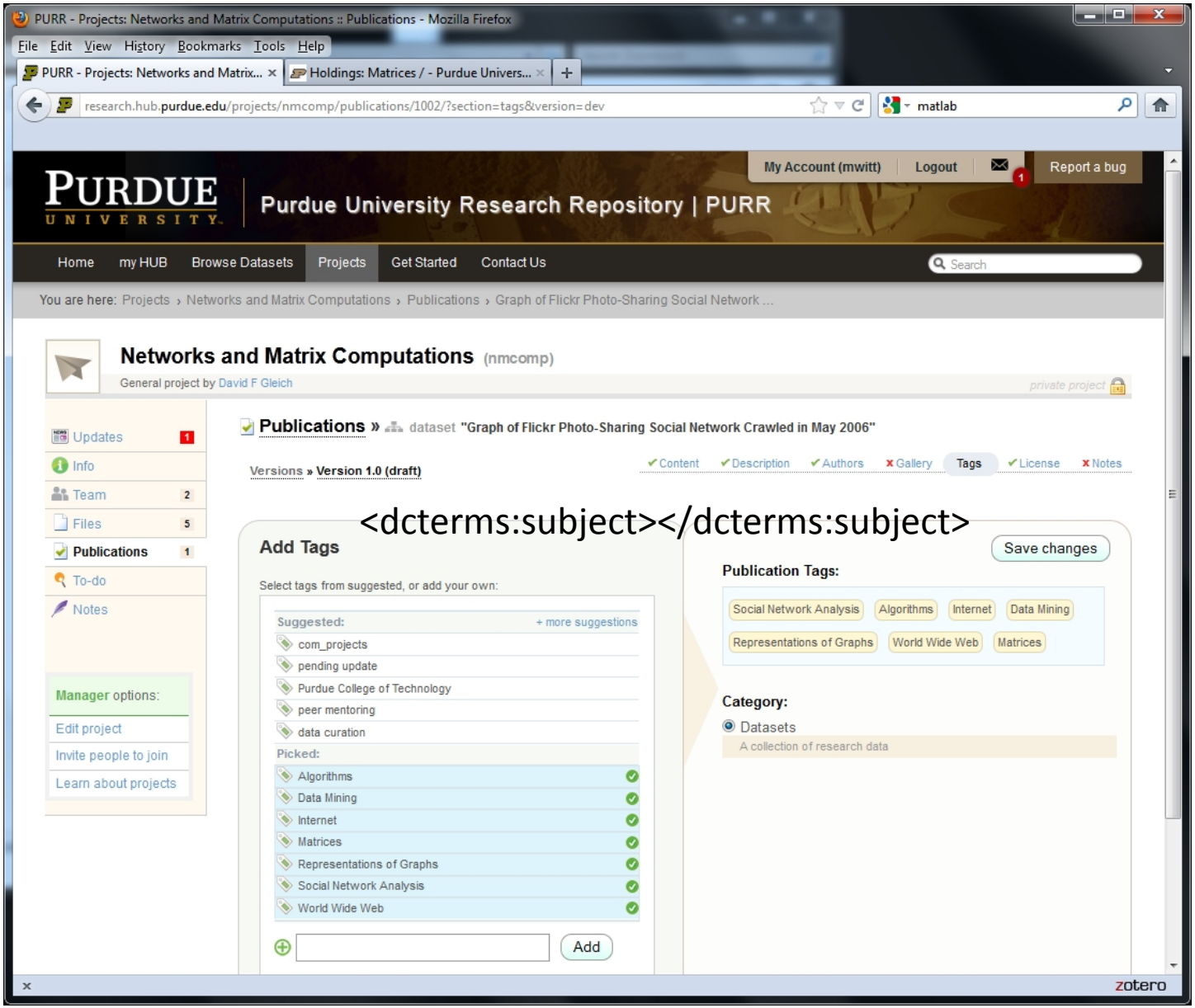
<dcterms:creator>Principle Author - Required</dcterms:creator>
<dcterms:contributor>Other Authors - Optional/Repeatable</dcterms:contributor>
<dcterms:date>Submission Timestamp (ISO 8601) - Required</dcterms:date>
<dcterms:description>Abstract - Currently Required/Repeatable</dcterms:description>
<dcterms:description>Synopsis - Currently Required/Repeatable</dcterms:description>
<dcterms:description>Notes - Currently Required/Repeatable</dcterms:description>
<dcterms:format>BagIt - Hard coded - Required</dcterms:format>
<dcterms:identifier>DOI - Required</dcterms:identifier>
<dcterms:publisher>Purdue University Research Repository - Hard coded - Required</dcterms:publisher>
<dcterms:rights>Information about rights held in and over the resource</dcterms:rights>
<dcterms:subject>Tags - Required/Repeatable</dcterms:subject>
<dcterms:title>Required</dcterms:title>
<dcterms:type>Dataset - Hard coded - Required</dcterms:type>

PURDUE UNIVERSITY LIBRARIES

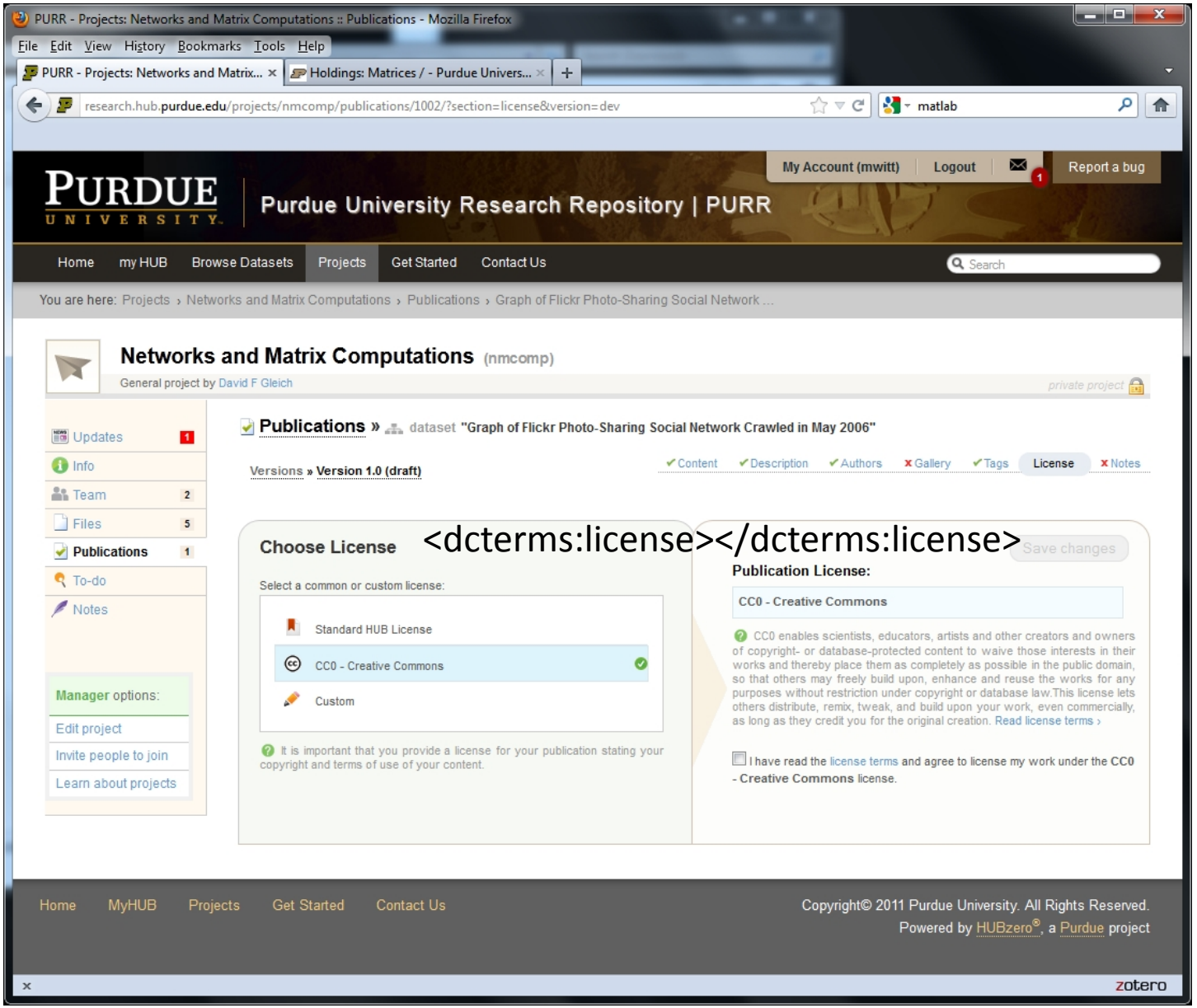
IMPLEMENTATION



PURDUE UNIVERSITY LIBRARIES



PURDUE UNIVERSITY LIBRARIES



PURDUE UNIVERSITY LIBRARIES

Metadata Encoding and Transmission Standard (METS) Wrapper

```
<mets:mets…
 <mets:dmdSec ID="DC">
  <mets:mdWrap MDTYPE="DC">
   <mets:xmlData>
                          Dublin Core Terms
   </mets:xmlData>
  </mets:mdWrap>
 </mets:dmdSec>
 <mets:amdSec>
 <mets:techMD ID="object1">
  <mets:mdWrap MDTYPE="PREMIS:OBJECT">
   <mets:xmlData>

   </mets:xmlData>
  </mets:mdWrap>
 </mets:techMD>
 <mets:digiprovMD ID="event1" >
  <mets:mdWrap MDTYPE="PREMIS:EVENT">
   <mets:xmlData>

   </mets:xmlData>
  </mets:mdWrap>
 </mets:digiprovMD>
 </mets:amdSec>
</mets:mets>
```

```
<dcterms:dcterms…
 <dcterms:creator>Principle Author - Required</dcterms:creator>
 <dcterms:contributor>Other Authors - Optional/Repeatable</dcterms:contributor>
 <dcterms:date>Submission Timestamp - Required</dcterms:date>
 <dcterms:desctiption>Abstract - Optional/Repeatable</dcterms:desctiption>
 <dcterms:description>Synopsis - Optional/Repeatable</dcterms:description>
 <dcterms:description>Notes - Optional/Repeatable</dcterms:description>
 <dcterms:format>Bagit - Hard coded</dcterms:format>
 <dcterms:identifier>DOI - Required</dcterms:identifier>
 <dcterms:publisher>Purdue University Research Repository - Hard

coded</dcterms:publisher>
 <dcterms:rights>Information about rights held in and over the resource</dcterms:rights>
 <dcterms:subject>Tags - Optional/Repeatable</dcterms:subject>
 <dcterms:title>Required</dcterms:title>
 <dcterms:type>Dataset - Hard coded</dcterms:type>
 </dcterms:dcterms>
</mets:xmlData>
```

PURDUE UNIVERSITY LIBRARIES

PREMIS Preservation Metadata

```
<mets:mets…
 <mets:dmdSec ID="DC">
  <mets:mdWrap MDTYPE="DC">
   <mets:xmlData>

   </mets:xmlData>
  </mets:mdWrap>
 </mets:dmdSec>
 <mets:amdSec>
 <mets:techMD ID="object1">
  <mets:mdWrap MDTYPE="PREMIS:OBJECT">
   <mets:xmlData>
        Administrative Metadata
   </mets:xmlData>
  </mets:mdWrap>
 </mets:techMD>
 <mets:digiprovMD ID="event1" >
  <mets:mdWrap MDTYPE="PREMIS:EVENT">
   <mets:xmlData>

   </mets:xmlData>
  </mets:mdWrap>
 </mets:digiprovMD>
 </mets:amdSec>
</mets:mets>
```

```
<premis:object xsi:type="premis:file" xsi:schemaLocation="info:lc/xmlns/premis-v2
                        http://www.loc.gov/standards/premis/v2/premis-v2-0.xsd">
 <premis:objectIdentifier>
  <premis:objectIdentifierType>CHECKSUM - Required</premis:objectIdentifierType>
   <premis:objectIdentifierValue>Generated checksum</premis:objectIdentifierValue>
  </premis:objectIdentifier>
  <premis:preservationLevel>
   <premis:preservationLevelValue>full</premis:preservationLevelValue>
   <premis:preservationLevelDateAssigned>00000000
                                        </premis:preservationLevelDateAssigned>
  </premis:preservationLevel>
  <premis:objectCharacteristics>
   <premis:compositionLevel>0</premis:compositionLevel>
    <premis:fixity>
     <premis:messageDigestAlgorithm>Name of CHECKSUM

algorithm</premis:messageDigestAlgorithm>
      <premis:messageDigest>Generated checksum</premis:messageDigest>
      <premis:messageDigestOriginator>PURR</premis:messageDigestOriginator>
     </premis:fixity>
     <premis:size>000000</premis:size>
      <premis:format>
       <premis:formatDesignation>
        <premis:formatName>File format</premis:formatName>
        <premis:formatVersion>If the format is versioned, formatVersion should be
                              recorded. It can be either a numeric or
chronological
designation.</premis:formatVersion>
       </premis:formatDesignation>
        <premis:formatRegistry>
         <premis:formatRegistryName>DROID or Unix
              Tools?</premis:formatRegistryName>
         <premis:formatRegistryKey>(e.g., fmt/10)</premis:formatRegistryKey>
         <premis:formatRegistryRole>specification</premis:formatRegistryRole>
        </premis:formatRegistry>
      </premis:format>
```

PURDUE UNIVERSITY LIBRARIES

```
<mets:mets…
 <mets:dmdSec ID="DC">
  <mets:mdWrap MDTYPE="DC">
   <mets:xmlData>

   </mets:xmlData>
  </mets:mdWrap>
 </mets:dmdSec>
 <mets:amdSec>
 <mets:techMD ID="object1">
  <mets:mdWrap MDTYPE="PREMIS:OBJECT">
   <mets:xmlData>
```

Technical Metadata

```
   </mets:xmlData>
  </mets:mdWrap>
 </mets:techMD>
 <mets:digiprovMD ID="event1" >
  <mets:mdWrap MDTYPE="PREMIS:EVENT">
   <mets:xmlData>

   </mets:xmlData>
  </mets:mdWrap>
 </mets:digiprovMD>
 </mets:amdSec>
</mets:mets>
```

```
<premis:creatingApplication>
<premis:creatingApplicationName>Software used to create the file. Repeatable for multiple software used.</premis:creatingApplicationName>
<premis:creatingApplicationVersion>Software version</premis:creatingApplicationVersion>
<premis:dateCreatedByApplication>00000000</premis:dateCreatedByApplication>
</premis:creatingApplication>
```

PURDUE UNIVERSITY LIBRARIES

```
<mets:mets…
 <mets:dmdSec ID="DC">
  <mets:mdWrap MDTYPE="DC">
   <mets:xmlData>

   </mets:xmlData>
  </mets:mdWrap>
 </mets:dmdSec>
 <mets:amdSec>
 <mets:techMD ID="object1">
  <mets:mdWrap MDTYPE="PREMIS:OBJECT">
   <mets:xmlData>
```

Technical Metadata

```
   </mets:xmlData>
  </mets:mdWrap>
 </mets:techMD>
 <mets:digiprovMD ID="event1" >
  <mets:mdWrap MDTYPE="PREMIS:EVENT">
   <mets:xmlData>

   </mets:xmlData>
  </mets:mdWrap>
 </mets:digiprovMD>
 </mets:amdSec>
</mets:mets>
```

```
<premis:hardware>
<premis:hwName>Name of hardware</premis:hwName>
<premis:hwType>Processor</premis:hwType>
<premis:hwOtherInformation>(e.g., 60 mhz
minimum)</premis:hwOtherInformation>
</premis:hardware>
<premis:hardware>
<premis:hwName>(e.g., 64 MB RAM)</premis:hwName>
<premis:hwType>Memory</premis:hwType>
<premis:hwOtherInformation>(e.g., 32 MB minimum)</premis:hwOtherInformation>
</premis:hardware>
<premis:environmentExtension>
<hardwareInformation/>
<softwareInformation/>
</premis:environmentExtension>
```

PURDUE UNIVERSITY LIBRARIES

```
<mets:mets…
 <mets:dmdSec ID="DC">
  <mets:mdWrap MDTYPE="DC">
   <mets:xmlData>

   </mets:xmlData>
  </mets:mdWrap>
 </mets:dmdSec>
 <mets:amdSec>
 <mets:techMD ID="object1">
  <mets:mdWrap MDTYPE="PREMIS:OBJECT">
   <mets:xmlData>
```

Provenance Metadata

```
   </mets:xmlData>
  </mets:mdWrap>
 </mets:techMD>
 <mets:digiprovMD ID="event1" >
  <mets:mdWrap MDTYPE="PREMIS:EVENT">
   <mets:xmlData>

   </mets:xmlData>
  </mets:mdWrap>
 </mets:digiprovMD>
 </mets:amdSec>
</mets:mets>
```

```
<premis:eventType>validation</premis:eventType>
<premis:eventDateTime>2006-06-06T00:00:00.001</premis:eventDateTime>
<premis:eventDetail>jhove1_1e - validation software</premis:eventDetail>
 <premis:eventOutcomeInformation>
  <premis:eventOutcome>successful</premis:eventOutcome>
 <premis:eventOutcomeDetail>
   <premis:eventOutcomeDetailNote>Well-formed and valid</premis:eventOutcomeDetailNote>
   <premis:eventOutcomeDetailExtension>
    <logfileInfo>
     <in/>
     <out/>
    </logfileInfo>
   </premis:eventOutcomeDetailExtension>
  </premis:eventOutcomeDetail>
 </premis:eventOutcomeInformation>
```

PURDUE UNIVERSITY LIBRARIES

```
<mets:mets…
 <mets:dmdSec ID="DC">
  <mets:mdWrap MDTYPE="DC">
   <mets:xmlData>

   </mets:xmlData>
  </mets:mdWrap>
 </mets:dmdSec>
 <mets:amdSec>
 <mets:techMD ID="object1">
  <mets:mdWrap MDTYPE="PREMIS:OBJECT">
   <mets:xmlData>

   </mets:xmlData>
  </mets:mdWrap>
 </mets:techMD>
 <mets:digiprovMD ID="event1" >
  <mets:mdWrap MDTYPE="PREMIS:EVENT">
   <mets:xmlData>
```

Provenance Metadata

```
   </mets:xmlData>
  </mets:mdWrap>
 </mets:digiprovMD>
 </mets:amdSec>
</mets:mets>
```

```
<premis:eventType>ingestion</premis:eventType>
<premis:eventDateTime>2006-06-
06T00:00:00.002</premis:eventDateTime>
<premis:eventDetail>Ingest tool/software (e.g.,
ingester1_0.exe)</premis:eventDetail>
<premis:eventOutcomeInformation>
<premis:eventOutcome>successful</premis:eventOutcome>
</premis:eventOutcomeInformation>
```

```
<premis:eventType>migration</premis:eventType>
<premis:eventDateTime>2006-07-06T00:00:00.006</premis:eventDateTime>
<premis:eventDetail>Name of software used to migrate version (e.g., Adobe Acrobat v. 9)
</premis:eventDetail>
<premis:eventOutcomeInformation>
<premis:eventOutcome>successful</premis:eventOutcome>
</premis:eventOutcomeInformation>
```

PURDUE UNIVERSITY LIBRARIES

Archival Information Package (AIP)

```
<mets:mets xmlns:lc="http://www.loc.gov/mets/profiles/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:mets="http://www.loc.gov/METS/" xmlns:premis="info:lc/xmlns/premis-v2"
xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/mets.xsd">
    <mets:dmdSec ID="DC">
        <mets:mdWrap MDTYPE="DC">
            <mets:xmlData>
                <dcterms:dcterms xmlns:dcterms="http://purl.org/dc/terms/"
xsi:schemaLocation="http://purl.org/dc/terms/ http://dublincore.org/schemas/xmls/qdc/dcterms.xsd">
                    <dcterms:creator>Principle Author - Required</dcterms:creator>
                    <dcterms:contributor>Other Authors - Optional/Repeatable</dcterms:contributor>
                    <dcterms:date>Submission Timestamp - Required</dcterms:date>
                    <dcterms:desctiption>Abstract - Optional/Repeatable</dcterms:desctiption>
                    <dcterms:description>Synopsis - Optional/Repeatable</dcterms:description>
                    <dcterms:description>Notes - Optional/Repeatable</dcterms:description>
                    <dcterms:format>Bagit - Hard coded</dcterms:format>
                    <dcterms:identifier>DOI - Required</dcterms:identifier>
                    <dcterms:publisher>Purdue University Research Repository - Hard coded</dcterms:publisher>
                    <dcterms:rights>Information about rights held in and over the resource</dcterms:rights>
                    <dcterms:subject>Tags - Optional/Repeatable</dcterms:subject>
                    <dcterms:title>Required</dcterms:title>
                    <dcterms:type>Dataset - Hard coded</dcterms:type>
                </dcterms:dcterms>
            </mets:xmlData>
        </mets:mdWrap>
    </mets:dmdSec>
    <mets:amdSec>
        |
```

Dissemination Information Package (DIP)

Puuuurrrrrrrrrrr....

PURDUE UNIVERSITY LIBRARIES

Dissemination Information Package (DIP)

Searchable – within PURR

Discoverable – through other systems…

PURDUE UNIVERSITY LIBRARIES

PURDUE
UNIVERSITY
LIBRARIES

OAI-PMH

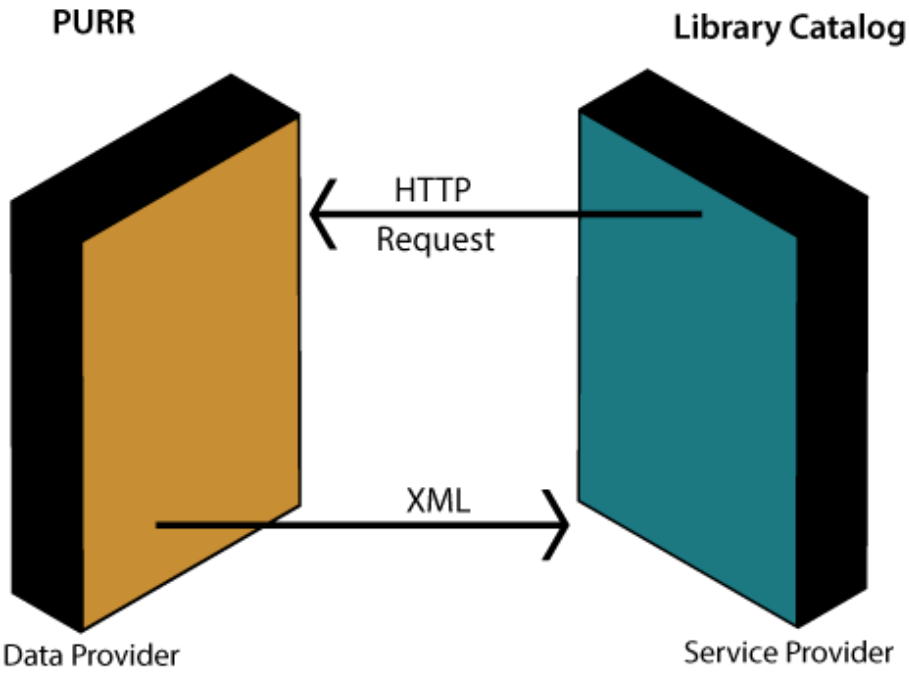
OPEN ARCHIVES INITIATIVE PROTOCOL FOR METADATA HARVESTING

OAI-PMH

OPEN ARCHIVES INITIATIVE PROTOCOL FOR METADATA HARVESTING

Application-independent framework based on metadata harvesting. There are two classes of participants in the OAI-PMH framework:

- **Data Providers** administer systems that support the OAI-PMH as a means of exposing metadata; and
- **Service Providers** use metadata harvested via the OAI-PMH as a basis for building value-added services.



PURDUE UNIVERSITY LIBRARIES

OAI-PMH XML OUTPUT

HUBNAME.ORG/?OPTION=COM_OAIPMH&VERB=LISTRECORDS&METADATAPREFIX=OAI_DC


```
▼<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2012-09-20T11:07:55-04:00</responseDate>
  <request verb="ListRecords" metadataPrefix="oai_dc">http://research.purdue.edu/?option=com_oaipmh</request>
 ▼<ListRecords>
  ▼<record>
   ▼<header>
     <identifier>http://dx.doi.org/10.5072/FK299999</identifier>
     <datestamp>2012-08-30T10:38:47-04:00</datestamp>
     <setSpec>Dataset</setSpec>
    </header>
   ▼<metadata>
    ▼<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
     http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
      <dc:title>oaipmh.xml</dc:title>
      <dc:creator>Brandon M Beatty</dc:creator>
      <dc:subject>test</dc:subject>
      <dc:subject>Advanced Exercises</dc:subject>
      <dc:date>2012-08-30 10:38:47</dc:date>
      <dc:identifier>http://dx.doi.org/10.5072/FK299999</dc:identifier>
      <dc:description>The is the Abstract and Description.</dc:description>
      <dc:type>Dataset</dc:type>
      <dc:publisher>Purdue University Research Repository</dc:publisher>
      <dc:rights>CC0 - Creative Commons</dc:rights>
      <dc:contributor>Brandon M Beatty</dc:contributor>
      <dc:contributor>Courtney Matthews</dc:contributor>
      <dc:relation>oaipmh.xml</dc:relation>
     </oai_dc:dc>
    </metadata>
   </record>
  ▼<record>
   ▼<header>
     <identifier>http://dx.doi.org/10.5072/FK299999</identifier>
     <datestamp>2012-08-10T13:16:51-04:00</datestamp>
     <setSpec>Dataset</setSpec>
    </header>
   ▼<metadata>
    ▼<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
     http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
      <dc:title>BagItPHP 19</dc:title>
      <dc:creator>Mark Leighton Fisher</dc:creator>
      <dc:subject>test</dc:subject>
      <dc:date>2012-08-10 13:16:51</dc:date>
      <dc:identifier>http://dx.doi.org/10.5072/FK299999</dc:identifier>
      <dc:description>BagItPHP 19</dc:description>
      <dc:type>Dataset</dc:type>
      <dc:publisher>Purdue University Research Repository</dc:publisher>
      <dc:rights>CC0 - Creative Commons</dc:rights>
      <dc:contributor>Mark Leighton Fisher</dc:contributor>
      <dc:relation>new-headshot.jpg</dc:relation>
      <dc:relation>purdue-headshot.jpg</dc:relation>
```


PURDUE UNIVERSITY LIBRARIES

THANK YOU

Neal Harmeyer – Digital Archivist – harmeyna@purdue.edu
Amy Hatfield – Metadata Specialist – hatfiea@purdue.edu
Brandon Beatty – PURR Software Developer – bbeatty@purdue.edu

PURDUE UNIVERSITY LIBRARIES