Stanford University LIBRARIES & ACADEMIC INFORMATION RESOURCES

Extending the Implementation of PREMIS to Geospatial Resources in the Stanford Digital Repository: An Exploration

By Nancy J. Hoebelheinrich
Metadata Coordinator
Digital Library Systems & Services

DLF Spring Forum, Minneapolis, MN 29
April 2008

To Be Discussed

- What PREMIS data elements are being used currently
- How & why
- PREMIS & Geospatial Resources - a fit?

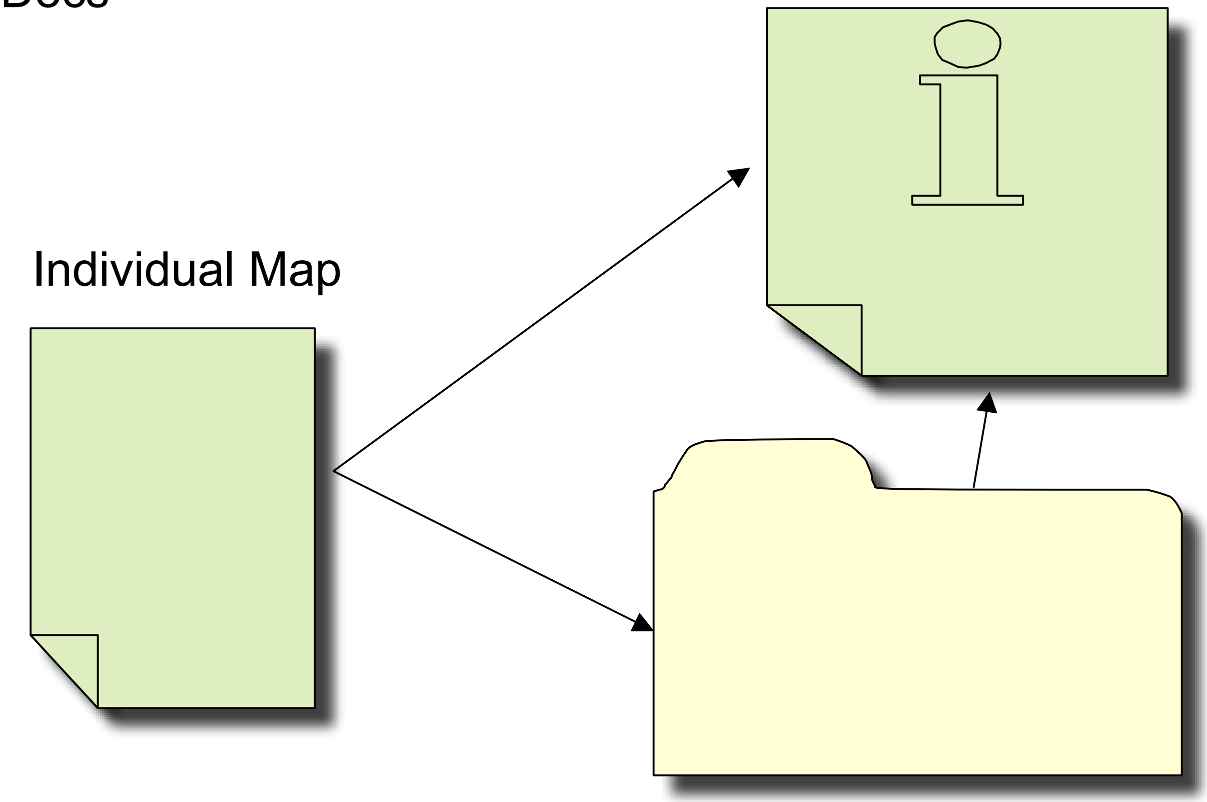
- Investigation (being) done as part of the National Geospatial Digital Archive project, funded by LC as part of NDIIPP.

Scenario 1: David Rumsey Historical Maps Collection

- Comprised of historical maps digitized as Single, Still image TIFFs
- METS Records for
  - Rumsey Deposit Agreement
  - Rumsey "Collection Level" & Auxiliary Files
  - Each Item

METS Documents for Rumsey Collection

Relationships among METS Docs



Individual Map

Deposit Agreement

Collection Level / Auxiliary Files

PREMIS Records contained w/in METS Documents

PREMIS *OBJECT*

PREMIS *RIGHTS*

PREMIS *EVENTS*

- Aspects of digital provenance
- Succinct link to full rights statement
- Important lifecycle events

Use of PREMIS Object Data Elements

- Used in each METS Document referencing files
  - Item, Agreement, “Collection Level” & Auxiliary Files
- Located in the METS <amdSec><techMD> section
- Automatic insertion by Ingest code to retain important provenance info for each file:
  - Original file name from data provider
  - Original checksum
  - Original file size
- Some information redundant, but prefer to retain in case METS sections need to be pulled out separately for action

PREMIS Object Excerpt (v1.1)

Element	Subelement or Attribute	Value
objectIdentifier	objectIdentifierType	filename
	objectIdentifierValue	0372001.tif
preservationLevel		bit preservation
objectCategory		file
objectCharacteristics	compositionLevel	0
fixity	messageDigestAlgorithm	MD5
	messageDigest	0c77e67 bebe3f338 4ec8bf4736648e41
size		315827432
format/ formatDesignation	formatName	TIFF
originalName		0372001.tif

Use of PREMIS Rights data elements

- Rumsey Deposit Agreement METS doc
- Represents the ingested *draft* Agreement with its own METS doc
- Placeholder for:
  - XML or other REL instance of full agreement ***or***
  - Use of METSRights once final agreement template is vetted & agreed upon by University Counsel

Use of PREMIS Rights data elements

- ***How****?*
- <amdSec><rightsMD>
- <mdWrap><xmlData>

- ***Why****?*
- Succinct summary of key information for quick access from METS Document itself
- Locator for more complete expression of terms, conditions;

PREMIS Rights Excerpt (v1.1)

Element	Subelement or Attribute	Value
permissionStatement	xmlID	SDR Access Phase 1
permissionStatementIdentifier	permissionStatementIdentifierType	Repository Permissions
	permissionStatementIdentifierValue	All digital objects falling under SDR Preservation Agreement_BitPreservation, v6.0, David Rumsey Map Collection
grantingAgreement	grantingAgreementIdentification	library_stanford_edu_fcab81ee605011db96c4339be
grantingAgreementInformation	contractAbstract	Version 6.0 of Agreement for Bit Preservation of Rumsey Collection
permissionGranted	act	Public Access
termOfGrant	startDate	2006-11-01
	endDate	2011-11-01
permissionNote /restrictionDefinition	restriction =	="Stanford only ” Stanford community only as defined in agreement .
	restriction =	="SDR_GROUP_xxx" Named group controlled by SUNET group as defined in agreement .
	restriction =	="No access" No access to content content allowed .

Use of PREMIS Event Data Elements

- Event 1:
  - Transform of descriptive MD from MS Access db => XML => MODS
  - Inserted into mets <amdSec><digiprovMD>

- Why this event?
  - In case of questions from outside data provider
  - Retain singular scripts & transform mechanisms
  - Test practicability of recording such events in production environment

PREMIS Event Excerpt (v1.1)

Element	Subelement or Attribute	Value
eventIdentifier	eventIdentifierType	MD_Transformation_Process
	eventIdentifierValue	Rumsey -MODS 3.2 for SDR
eventType		normalization
eventDateTime		2006 -12-01T02:48: 22
eventDetail		Steps of process transforming data provider's descriptive metadata to MODS 3.2 records as required for ingestion into SDR .
eventOutcomeInformation / eventOutcomeDetail /	SDR_Rumsey_Transformation / SDR_RumseyTransformationOutput	The Rumsey Access database, as delivered by Luna Insight, was converted to a single XML document using the MS Access Export function. Both the MS Access database is included as well as the XML file.
		A PERL script was used to break the monolithic XML document representing the MS Access database into many XML documents each representing a single image in the Rumsey collection. The single XML document was broken into separate document s at each occurrence of the "Object" tag. PERL script in text format is included.
		An XSLT was used to make MODS documents for all the Rumsey images. The X SLT file is included.
		SDR conversion code was written to pull geographic coordinates and scale metadata out of SUL MARC records from Unicorn catalog and insert them into the MODS records when available.
		SDR conversion codes was written to insert the comp MODS records into the METS record for each Rumsey digital object.

Scenario 2: Geospatial Files & PREMIS with METS – is it a fit?

- See “An Investigation into Archiving Geospatial data Formats “ prepared for NGDA Project, funded by NDIIPP (http://www.ngda.org/research.php):
  - Shapefiles (vector)
  - Digital Raster Graphics (DRG) raster files (digital representations of USGS topographical maps
  - Digital Ortho Quarter Quads (DOQQs) (images as geoTiffs, tfs or proprietary)
  - Landsat7 satellite images (preliminary)

Scenario 2: Geospatial Files & PREMIS – is it a fit?

- Paper examined approaches of
  - FGDC
  - PREMIS
  - Center for International Earth Science Information Network (CIESIN) ‘s Geospatial Electronic Record (GER) model on basis of:
    - Environment/ computer platform,
    - Semantic underpinnings
    - domain specific terminology,
    - provenance
    - data quality
    - appropriate use

Scenario 2: Geospatial Files & PREMIS with METS – is it a fit?

- Appears ok when:
- Domain specific MD exists, e.g., FGDC for descriptive and technical MD
- Have a number of layers of the resource with MD to be associated, e.g., at representation & file(s) level
- Depending upon the point in resource lifecycle wishing to document

Entering the sticky wicket: PREMIS for geospatial (and other science) data sets?

- Domain specific needs for that are difficult to incorporate:
  - Context
  - Environment including at time of creation
  - “Significant properties”
  - Existence of geospatial format registries

Use of PREMIS Object Data Elements – Scenario3: GIS Dataset

Street network of given metropolitan area

- Dataset 1: official street centerline file used by emergency services to locate street addresses
- Dataset 2: aspects of the road network including topography, angles & geometry of the road network used for a tourist map

Geospatial “Context”

- Placing dataset in Time & Space
- Semantic underpinnings, e.g.,
  - Abstract
  - Description of purpose / research methodology
  - Intended use of data to avoid misinterpretation or misuse
- Where to put?
  - FGDC has place, but does PREMIS, if doesn’t exist in “descriptive” or technical MD?
  - What would be place for this in PREMIS(?)
  - Perhaps <object><relationship><relatedObjectIdentification> for an explanatory website or other source of info?

“Environment” and/or “Significant properties”

- HW info pertinent at time of data creation
- SW info pertinent at time of data creation (?)
- Lineage or “provenance” data e.g., to communicate processing steps used to create scientific data product
- Events, parameters & source data which influenced or impacted the creation of the data set prior to its ingestion into the archive in order to full understand the data that you’re getting

“Environment” & “Significant properties”, continued…

  - Data Quality – describing completeness, logical consistency, attribute accuracy
  - Data Trustworthiness – data creator / provider reliable? = “authentic”
  - Data Provenance – processes & sources for dataset = “understandable & reliable”
  - Understanding of the specific needs of the “designated community”?
- How to do in PREMIS? – v.2 would appear to be better

Use of PREMIS Event Data Elements

- Event :
  - Would prefer the option to describe process of data creation
  - Merge c:\temp\states1;c:\temp \states2; c:\temp\USA
  - (includes process = “merge” and data sources
  - Advantage – can describe events once in repository, unlike FGDC

- Why this event?
  - Important to describe processes during different phases of lifecycle, even prior to ingestion
  - Not to be able to do so – problemmatic for geospatial resources

Issues & Challenges

- Getting domain specific MD would help!
- If not, getting important prez info from data creators & how to determine what is truly necessary for dataset use?
- Establishment of geospatial format registries
- Is this level of documentation still bit preservation?
- Getting buy-in from geospatial domains for use of vocabularies, etc. (see Global Spatial Data Infrastructure: http://www.gsdi.org/Default.asp)

Future directions for NGDA Project

- Further investigation of other geospatial formats including more vector based data such as:
  - layers of the National Atlas,
  - National Map (sections of California)
- Landsat 7 ETM imagery
- Derived data sets from Stanford faculty

Future directions, cont.

- Format Registry investigation - what should be included in a format registry for geospatial
  - Contact with key vendors, e.g. ESRI, SafeSoftware, etc.
- Monitoring what others are doing with e-science data sets, e.g.,
  - NCSU, Johns Hopkins
  - National Australian Archive (NAA)
  - JISC, and DPC in the UK doing with research on scientific data such as vector images, See "Significant Properties"
  - UK DCC SCARP Project (Sharing Curatorial and Re-Use Preservation) – Research on Lineage Data and others

Questions? / comments?

Nancy J. Hoebelheinrich
nhoebel@stanford.edu