SLAC Persistent Archives Testbed Project Metadata Discussions

General Discussions

10/25/2005 Antoine DeTorcy asked about levels of metadata: do we want to structure it hierarchically? 10/27/2005 Wilko Kroeger pointed out that the SRB hierarchy is Collection, Container, File. Jean Deken commented that the archival hierarchy has typically been Repository, Collection, Accession, Container, File. (Once again, we are using the same words, but meaning different things. In the SRB the Collection has a structure: you can have a directory structure within a collection, the collection has levels that are also called collection, until you get to the file level, which is called "file." You get different levels, but the hierarchy is a placement hierarchy, and not necessarily an intellectual hierarchy. It can be an intellectual hierarchy, but it does not have to be. Given this fact, using the SRB hierarchy to structure the metadata may not be workable.

We think that this topic is ripe for further discussion with the SDSC folks: perhaps we need to have our metadata in a separate table that is linked in some way to the items in the SRB?

11/1/05. Jean: Accession-level metadata applies to an accession (intellectually unified group of electronic records ingested at the same time). Accession-level metadata should be associated with the accession, and could be placed in a metadata table (like the current SLAC collections database, SLACARC. To see sample records, look at the results of a search on Richter under Physicist). Item-level metadata applies to an individual item, and should be associated, linked with the item. I have updated the list below to indicate which items should be accession-level, and which items should be item-level.

Attribute-Level Discussions

Injected Metadata

9/15/05: Each object, or each collection has its own metadata. If you have groups that consist of many files and collections . Perhaps we should have our own database tables that link to SRB objects, that give layers. SLAC's metadata database would mediate between the user and the SRB. User won't know the difference, it will be transparent to the user. Layers of metadata for SLAC objects, some (injected) metadata applies to ALL entities. Tricky part is to define the structure of the tables. Could use a template for individual groups of records (ex. SLD, BaBar).

1. **slac.gov.recordgroup** : **Record Group**

Level: Accession-level metadata

Discussion: 434 is the NARA record group number for the US Department of Energy. Other numbers may be appropriate for use for future accessions of records from SLAC.

2. **slac.gov.agency** **: Responsible federal agency**

Level:Accession-level metadata

Discussion: For the SLD records, this is the Department of Energy. In the future, this could be a different funding agency, like NASA (National Aeronautics and Space Administration) or NIH (National Institutes of Health).

3. **slac.gov.referenceby** **: Reference provided by**

Level: Accession-level metadata

Discussion: This metadata attribute is derived from the NARA LCDRG (Life-Cycle Data Requirements Guide). Right now we are using the SLAC Archives & History Office information: once the records have been transferred to NARA, contact information for the cognizant NARA unit will go here.

4. **slac.gov.schedule** **: Applicable records control schedule**

Level: Accession-level metadata

Discussion: This attribute uses the record series description and schedule citation from the authorized government records control schedule. For the SLD records, the applicable schedule is online and the relevant items are linked to each series on the SLC Records Descriptions page.

5. **slac.gov.series** **: Series within the applicable records control schedule to which the records belong.**

Level: Series/Accession-level metadata

Discussion: Series name from the authorized government records control schedule.

6. **slac.gov.description** **: Official series description**

Level: Series/Accession-level metadata

Discussion: Exact wording of series description from the authorized government records control schedule. In future, this will help to pull together same series records from different experiments or laboratories. Since we are using the exact wording in the government schedule, it may be possible to add data for this attribute automatically?

7. **slac.gov.retention** **: Period of time accession should be retained**

Level: Accession-level metadata

Discussion: The retention period is prescribed by the applicable government records control schedule item. Sample values for this attribute:

- Permanent, Offer to Archives 01/2029
- Retain until 10/2015
- Review 01/2009

8. **slac.creator.organization:** **Creating Organization**

Level: Accession-level metadata

Discussion: This is the top-level of the creating organization, so, for the SLD project this attribute value is SLAC.

9. **slac.creator.division** **: Creating Division at SLAC**

Level: Accession-level metadata

Discussion: For the SLD project, the creating division at SLAC is the Research Division, or RD.

10. **slac.creator.group** **: Creating Group at SLAC**

Level: Accession-level metadata

This metadata element could contain the group description from the SPIRES Experiments database: either the narrative description or a link to the description. The downside of having a link is that we do not control the Experiments database and in the past old entries have been deleted. It might be easier to copy over also, because we will only need this metadata element content once. (Actually, the old Experiments db content is on a server at SLAC, it is just not web-accessible at the moment.)

11. **slac.description.type** **: Type of archival description**

Level: Accession-level metadata

Discussion: At SLAC, for records retired to NARA, the description type will always be "Series."

12. **slac.description.by** **: Description author**

Level: Accession-level metadata

Discussion: The name of the person who provided the injected metadata is added here, in the format of Lastname, Firstname. This is a repeatable attribute. Might want to link to slac.description.date?

13. **slac.description.date** **: Date description was completed**

Level: Accession-level metadata
Discussion: Date that metadata was completed or last revised. Repeatable attribute.

14. **slac.description.remarks : Additional information about the accession**

Level: Accession-level metadata
Discussion: Used only if there is some additional information needed.

15. **slac.identifier.copy : Type of copy this is**

Level: Accession-level metadata
Discussion: For the SLD project, all of the copy types are "Preservation". Other types could be: Reference, Duplicate, Original (?)

16. **slac.identifier.contmgt : Content Management System**

Level: Accession-level metadata
Discussion: The name and version of a content management system that may have been used to manage files on the web. Required by NARA (NARA WCG 6.4.7). If no content management system was used, this attribute will be left out of the metadata set.

17. **slac.identifier.websitename : Name of Web Site**

Level: Accession-level metadata
Discussion: Generally found on the home or index page of a web site, generally a header on that page. Might be able to extract automatically? Required by NARA (NARA WCG 6.4.2)

18. **slac.capture.tool : Tool used to capture/crawl website**

Level: Accession-level metadata
Discussion: According to NARA WCG 6.4.5: "include the application used with either a URL to the application's web site or a description of the harvester's capabilities and the log file(s) generated by the harvester that document the harvesting process."

19. **slac.capture.settings : Settings used on capture/crawl tool**

Level: Accession-level metadata
Discussion:Information required by NARA WCG 6.4.5. Format will be determined by tool used.

20. **slac.capture.sitemap : Sitemap of captured/crawled site**

    Level: Accession-level metadata
    Discussion: Include if available (if created by crawl tool), per NARA WCG 6.4.10.

21. **slac.capture.date : Date capture/crawl of website accomplished**

    Level: Accession-level metadata
    Discussion: Information required by NARA WCG 6.4.5.

22. **slac.capture.contact : Person who accomplished capture/crawl of website**

    Level: Accession-level metadata
    Discussion: Information required by NARA WCG 6.4.6. Format is Lastname, Firstname. Email address, telephone number. (Repeatable attribute?)

23. **slac.capture.remarks : Remarks about capture/crawl of website**

    Level: Accession-level metadata
    Discussion: Use this attribute, if necessary, to record additional information about capture/crawl.

24. **slac.pawn.recordset**

    Level: Accession-level metadata
    Discussion: *Record Set* is a PAWN convention that allows the user to establish a link or relationship between more than 1 item or group of items BOTH as they are in transit AND after they have been submitted.

25. **< a name="pawncat" href="MetadataSchem8.html">slac.pawn.category**

    Level" Accession-level metadata
    Discussion: *Category* is the PAWN equivalent of a *Record Series*

Extracted Metadata

1. **slac.gov.access : Access restriction(s)**

Level: Accession-level metadata
Discussion: Access restriction can be prescribed by the government records schedule, or by the creator/creating group at SLAC. Options for this item are: Open, Restricted, or Restricted until xxx.

2. **slac.creator.person** : **Individual responsible for creating the entity.**

Level: Item-level metadata
Discussion: Format as Lastname, Firstname. Should be able to extract from pages. Should be repeatable attribute, since more than one person's name can be associated with an electronic entity.

3. **slac.creator.owner** : **Owner**

Level: Item-level metadata
Discussion: Individual named as owner of the entity, if different from the creator

4. **slac.description.local** : **SLAC-generated narrative description of records**

Level: Accession-level metadata
Discussion: What the records series or web site is called at SLAC, as opposed to what the official government records schedule series description calls it.

5. **slac.description.use** : **Is research use allowed at this time?**

Level:
Discussion: Attribute is either yes (use is allowed) or no (use is not allowed at this time). Jean 10/19/2005: For all of the PAT project SLD records series, this value should be set to "yes". For future projects, this value will probably need to be injected, based on the archivist's appraisal of the records series. Unless a tool could be created to establish this attribute based on the value of the slac.gov.access attribute?

6. **slac.description.webplatform**: **Web Platform**

7. **slac.date.begun** : **Beginning date**

Level: Item-level metadata

8. **slac.date.modified** : **Date last modified**

Level: Item-level metadata

9. **slac.identifier.url** : **Original url for record/resource**

    Level: Item-level metadata

10. **slac.identifier.filename** : **Original filename**

    Level: Item-level metadata

11. **slac.description.format** :

    Level: Item-level metadata

12. **slac.description.filesize** :

    Level: Item-level metadata

13. **slac.identifier.storagelocation** : **Storage location of the copy being described**

    Level: Item-level or Accession-level metadata??

    11/1/05: Question from Jean--will a SLAC-identified accession be stored in the same location on the SRB, or will the location metadata need to be item-level metadata?

14. **slac.identifier.persistent** : **Persistent identifier**

    Level: Item-level metadata

Articles and References

- Ball, Craig. Make Friends With Metadata. *Law Technology News*. January 26, 2006
- Brown, Thomas E. Toward The Appraisal of Web Records archival outlook, July 2006. p. 6
- Cloonan, Michele V. and Shelby Sanett. "The Preservation of Digital Content." *portal: Libraries and the Academy,* Vol. 5, No. 2 (2005) pp.213-237. Baltimore: Johns Hopkins University Press.

- Cobb, Judith, Pearce-Moses, Richard, & Surface, Taylor. ECHO DEPository Project. http://www.ndiipp.uiuc.edu/pdfs/IST2005paper_final.pdf (Viewed on 3 November 2005).
- Crease, Robert P. Experts. Physics World, August 3, 2007. Cites Collins and Evans, who make "a distinction between contributory expertise, possessed by active practitioners of a field, and interactional expertise, whereby someone can speak knowledgeably about a subject without being able to contribute new ideas to it."
- Day, Michael. Collecting and Preserving the World Wide Web: A feasibility study undertaken for the JISC and Wellcome Trust. UKOLN, University of Bath, Version 1.0 - 25 February 2003 (viewed 11/7/2005).
- Deken, Jean and Hasan, Adil. Archiving SLD Records in SRB: The Persistent Archives Test-Bed (PAT) Project at SLAC in 2004 . Stanford University. SLAC-PUB-10857. December 2004
- Digital Preservation Coalition. Report for the DCC/DPC Workshop on Cost Models for preserving digital assets. British Library Conference Centre, Tuesday 26 July 2005.
- Hodge, Gail et al. A Metadata Element Set for Project Documentation. Science & Technology Libraries Volume: 25 Issue: 4 ISSN: 0194-262X Pub Date: 5/23/2005
- IWAW05: Fifth International Web Archiving Workshop. September 22 and 23 2005, Vienna, Austria.
- Masanès, Julien. Towards Continuous Web Archiving D-Lib Magazine December 2002 V. 8 No. 12. (viewed 11/7/2005)
- PADI (Preserving Access to Digital Information). Web Archiving (viewed 11/3/2005). From the page: *This page serves as an introduction to some of the major archiving initiatives that have been established by national libraries around the world to preserve their country's Web heritage. The main models that have emerged for archiving Web content are listed, together with descriptive 'snapshots' of the key projects and collaborative initiatives.*
- Pearce-Moses, Richard and Kaczmarek, Joanne. An Arizona Model for Preservation and Access of Web Documents. (Originally published in DttP: Documents to the People 33:1 (Spring 2005), p. 17-24. http://www.dlapr.lib.az.us/diggovt/azmodel/AzModel.pdf (viewed 11/3/2005)
- Rieger, Oya Y. Preservation in the Age of Large-Scale Digitization: A White Paper. February 2008 Council on Library and Information Resources. Washington DC
- Shirky, Clay. AIHT: Conceptual Issues from Practical Tests. D-Lib Magazine, December 2005. Vol. 11 No. 12. (AIHT = Archive Ingest and Handling Test)
- Tennant, Roy. "Bitter Harvest: Problems & Suggested Solutions for OAI-PMH Data & Service Providers." (California Digital Library)
- Thibodeau, Kenneth. Archival Science and Archival Engineering: building a New Future for the Past . archival outlook, May/June 2006, p. 6.
- web-archive@cru.fr. List-serv on web content archiving, legal deposit of online publications, digital preservation. (viewed 11/7/2005)

Updated: 25 April 2007 J.M. Deken