Metadata Development for the Persistent Archives Testbed (PAT) Project

Jean Deken

Archives & History Office
Stanford Linear Accelerator Center

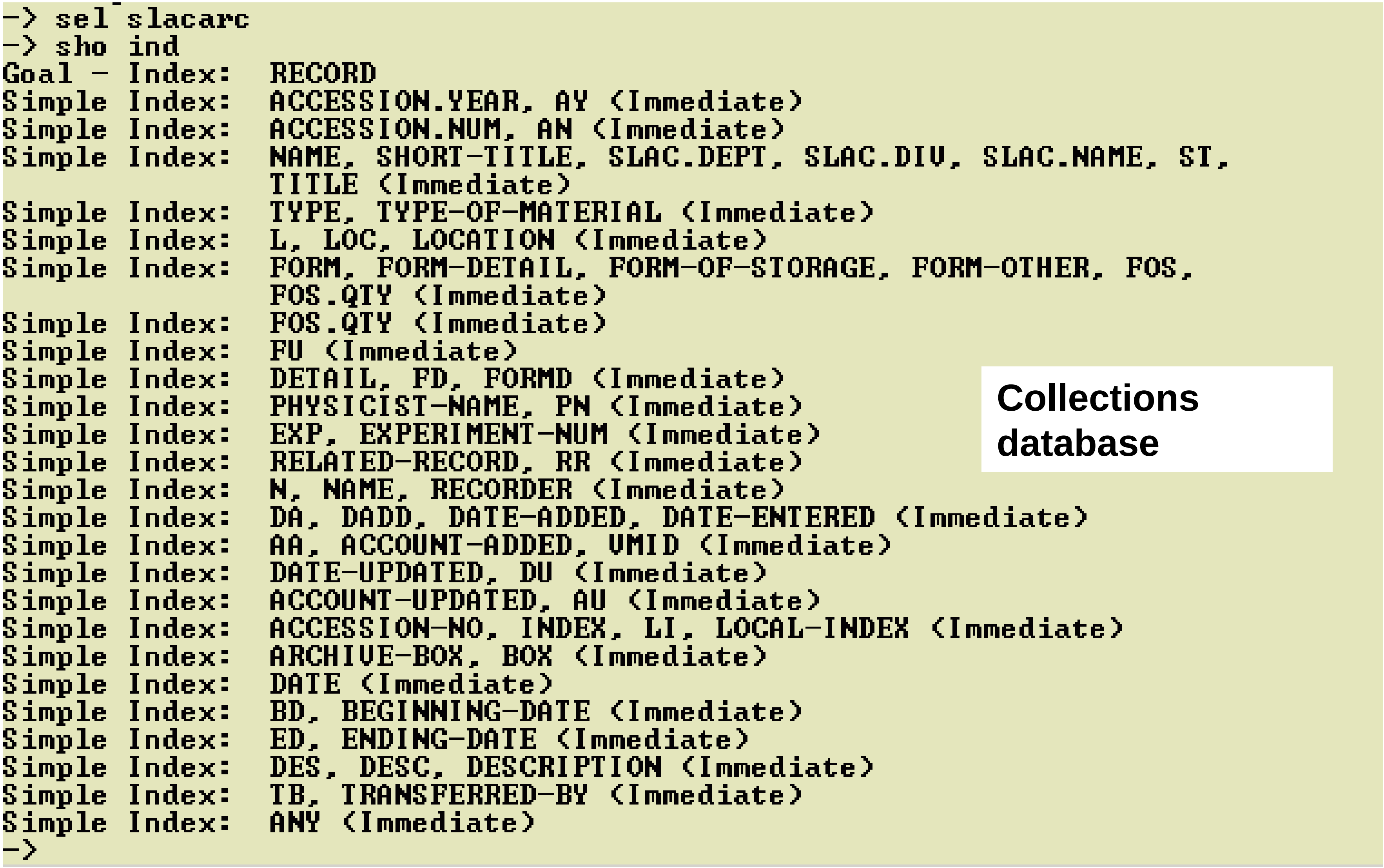
Metadata for SLAC - Overview

- Metadata @ SLAC (before PAT)
- PAT Project Background
- PAT Project Metadata Development / Evolution
- Some conclusions

Metadata @ SLAC Before PAT

- Suite of database indexes
  - **SLACARC – collections**
  - PhotoIndex – photographs and images (some digitized images
  - SLACSpeak – glossary of terms and acronyms
  - SLACNews – index of staff / internal newsletters and periodicals
- Standards
  - MARC / MARC-AMC
  - Locally developed (late 1980's – early 1990's)

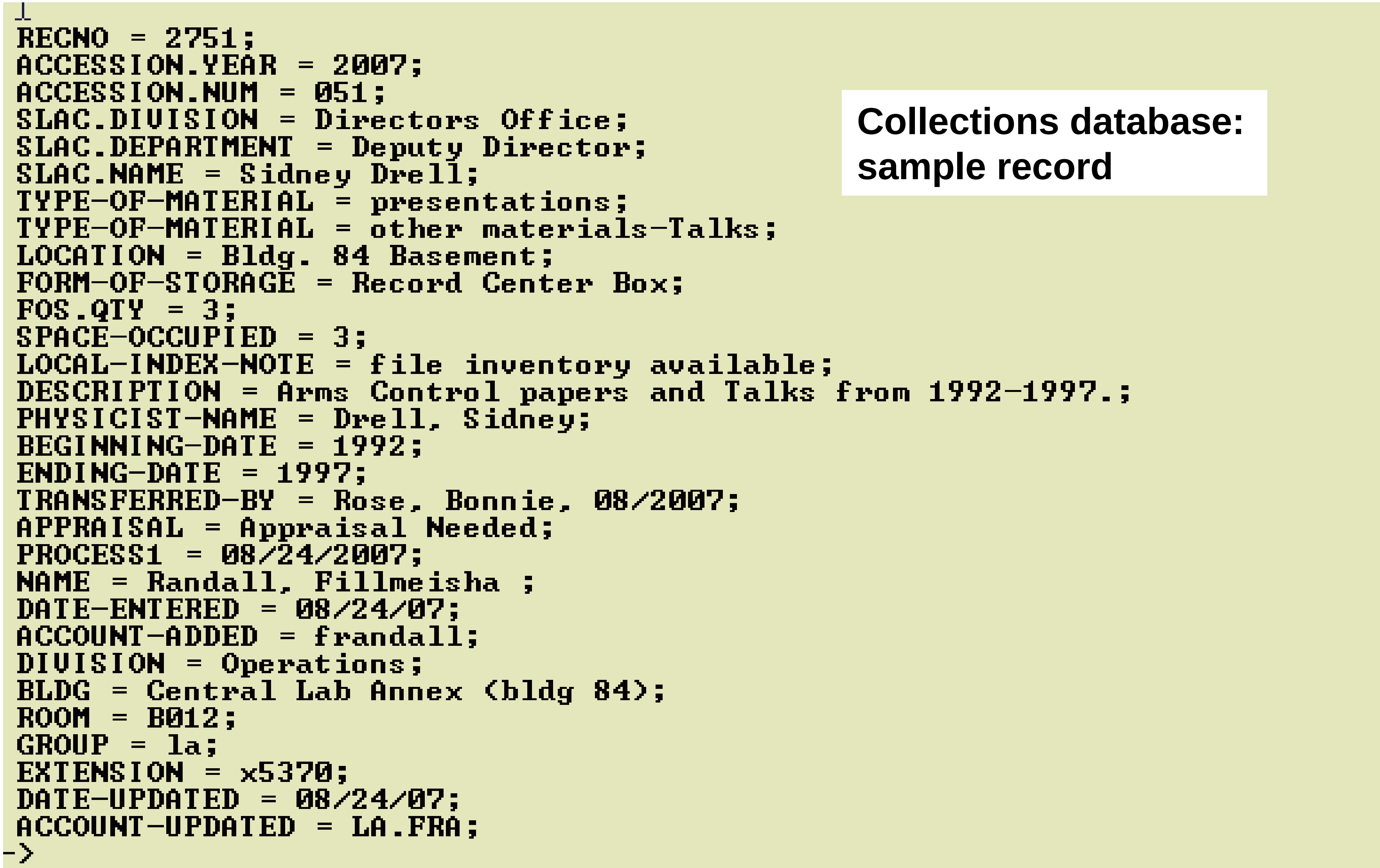
Metadata @ SLAC Before PAT



```
-> sel slacarc
-> sho ind
Goal - Index:   RECORD
Simple Index:   ACCESSION.YEAR, AY (Immediate)
Simple Index:   ACCESSION.NUM, AN (Immediate)
Simple Index:   NAME, SHORT-TITLE, SLAC.DEPT, SLAC.DIV, SLAC.NAME, ST,
                TITLE (Immediate)
Simple Index:   TYPE, TYPE-OF-MATERIAL (Immediate)
Simple Index:   L, LOC, LOCATION (Immediate)
Simple Index:   FORM, FORM-DETAIL, FORM-OF-STORAGE, FORM-OTHER, FOS,
                FOS.QTY (Immediate)
Simple Index:   FOS.QTY (Immediate)
Simple Index:   FU (Immediate)
Simple Index:   DETAIL, FD, FORMD (Immediate)
Simple Index:   PHYSICIST-NAME, PN (Immediate)
Simple Index:   EXP, EXPERIMENT-NUM (Immediate)
Simple Index:   RELATED-RECORD, RR (Immediate)
Simple Index:   N, NAME, RECORDER (Immediate)
Simple Index:   DA, DADD, DATE-ADDED, DATE-ENTERED (Immediate)
Simple Index:   AA, ACCOUNT-ADDED, UMID (Immediate)
Simple Index:   DATE-UPDATED, DU (Immediate)
Simple Index:   ACCOUNT-UPDATED, AU (Immediate)
Simple Index:   ACCESSION-NO, INDEX, LI, LOCAL-INDEX (Immediate)
Simple Index:   ARCHIVE-BOX, BOX (Immediate)
Simple Index:   DATE (Immediate)
Simple Index:   BD, BEGINNING-DATE (Immediate)
Simple Index:   ED, ENDING-DATE (Immediate)
Simple Index:   DES, DESC, DESCRIPTION (Immediate)
Simple Index:   TB, TRANSFERRED-BY (Immediate)
Simple Index:   ANY (Immediate)
->
```

Collections database

Metadata @ SLAC Before PAT



```
RECNO = 2751;
ACCESSION.YEAR = 2007;
ACCESSION.NUM = 051;
SLAC.DIVISION = Directors Office;
SLAC.DEPARTMENT = Deputy Director;
SLAC.NAME = Sidney Drell;
TYPE-OF-MATERIAL = presentations;
TYPE-OF-MATERIAL = other materials-Talks;
LOCATION = Bldg. 84 Basement;
FORM-OF-STORAGE = Record Center Box;
FOS.QTY = 3;
SPACE-OCCUPIED = 3;
LOCAL-INDEX-NOTE = file inventory available;
DESCRIPTION = Arms Control papers and Talks from 1992-1997.;
PHYSICIST-NAME = Drell, Sidney;
BEGINNING-DATE = 1992;
ENDING-DATE = 1997;
TRANSFERRED-BY = Rose, Bonnie, 08/2007;
APPRAISAL = Appraisal Needed;
PROCESS1 = 08/24/2007;
NAME = Randall, Fillmeisha ;
DATE-ENTERED = 08/24/07;
ACCOUNT-ADDED = frandall;
DIVISION = Operations;
BLDG = Central Lab Annex (bldg 84);
ROOM = B012;
GROUP = la;
EXTENSION = x5370;
DATE-UPDATED = 08/24/07;
ACCOUNT-UPDATED = LA.FRA;
->
```

Collections database: sample record

PAT Project – Background

- Persistent Archives Testbed (PAT)
  - Goal:

  conduct case studies that test the ability to implement the SDSC's Storage Resource Broker (SRB) data grid (http://www.npaci.edu/DICE/SRB) technology using a variety of archival collections.
  - Participants:
    - States of CA, KY, MI, MN, OH
    - Federal gov't: NHPRC, NARA, SLAC, Korea
    - Universities: GaTech, UCLA, UI-UC, U of FLA
    - Others …

PAT Project – Background

- Persistent Archives Testbed (PAT)
  - SLAC test collection – SLAC Large Detector (SLD) Collaboration
  - 1983-1988
  - Early and prolific user of world-wide web
  - No further need to keep data confidential
  - Many types of electronic documents
  - Meet US Department of Energy (DOE) / NARA criteria for retention

PAT Project – Background

- Persistent Archives Testbed (PAT)
  - Initial electronic records appraisal – manual “crawl” of web
  - Preliminary list of records series
  - Interviewed collaboration’s key staff
    - Data Czar
    - Web manager
    - Spokesperson
  - Automated Web crawls

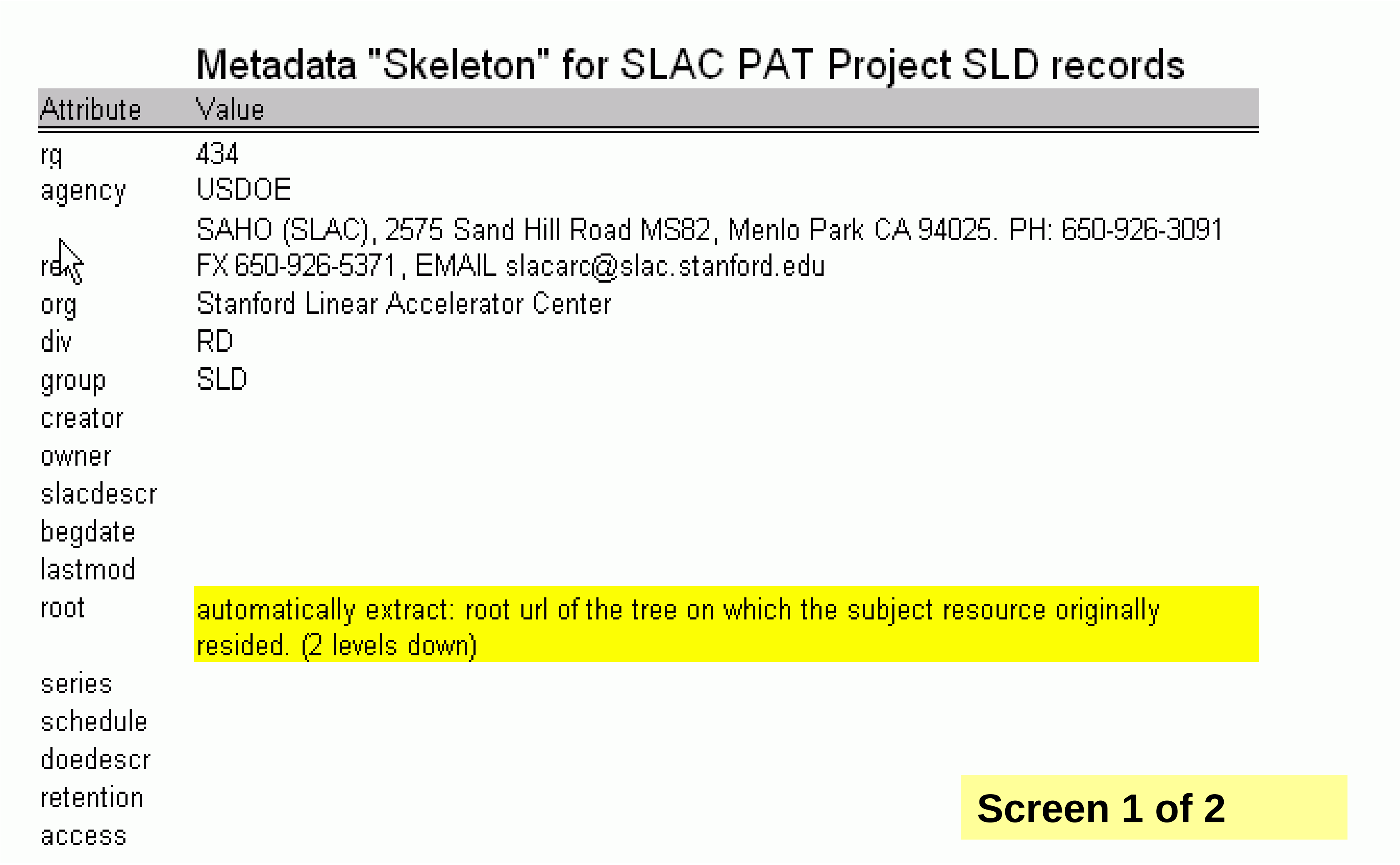
PAT Project Metadata Development

- Began with the data elements that we currently use for our archives collections database, SLACARC
- Looked at
  - **Dublin Core**
  - **METS**
  - **NARA metadata scheme (LCDRG)**
- Methodology: Concatenated exploration

(= make it up as you go along)

(Paul Conway-U of Michigan)

PAT Project Metadata Development

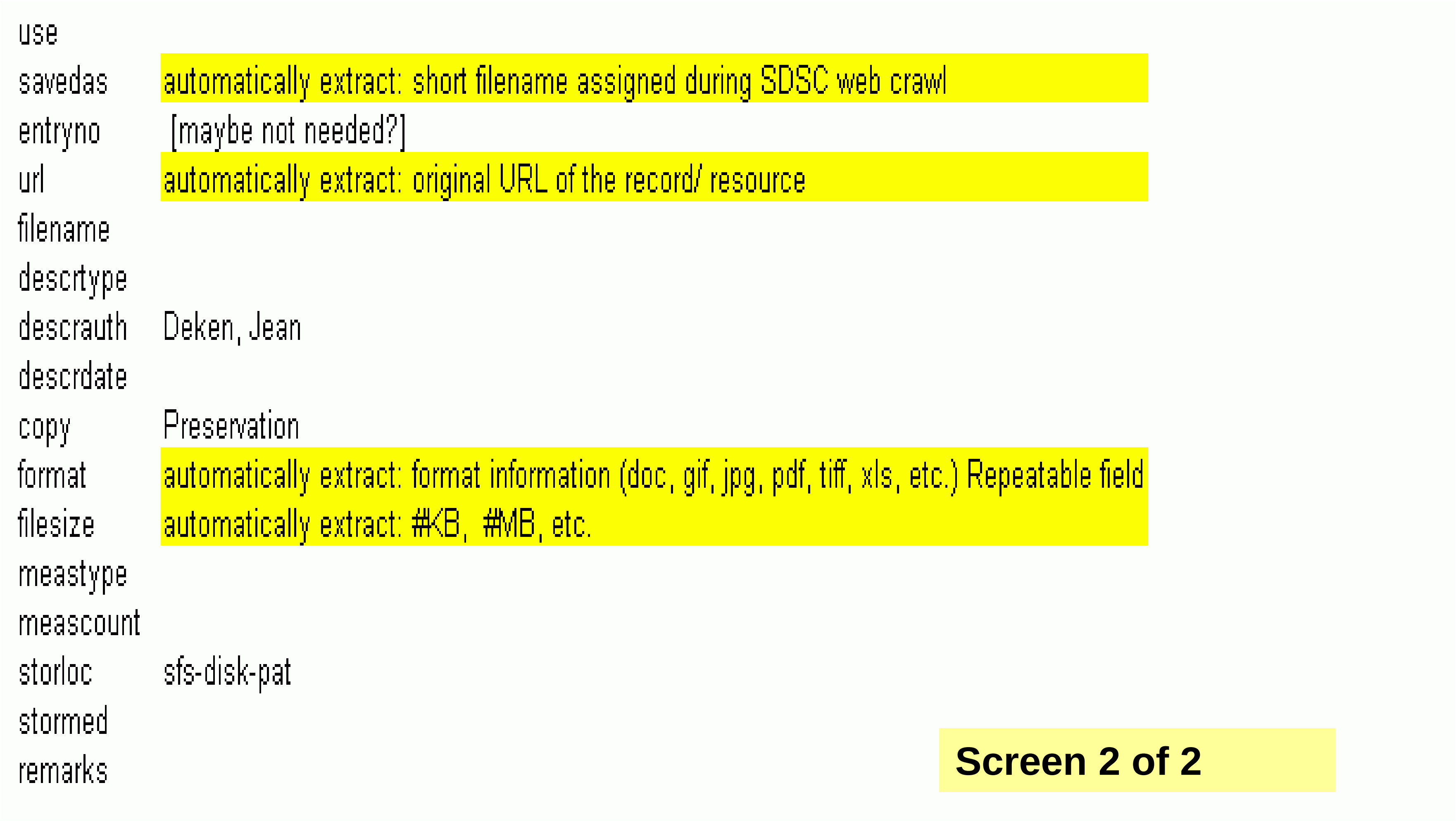


Metadata "Skeleton" for SLAC PAT Project SLD records

Attribute	Value
rg	434
agency	USDOE
re	SAHO (SLAC), 2575 Sand Hill Road MS82, Menlo Park CA 94025. PH: 650-926-3091 FX 650-926-5371, EMAIL slacarc@slac.stanford.edu
org	Stanford Linear Accelerator Center
div	RD
group	SLD
creator	
owner	
slacdescr	
begdate	
lastmod	
root	automatically extract: root url of the tree on which the subject resource originally resided. (2 levels down)
series	
schedule	
doedescr	
retention	
access	

Screen 1 of 2

PAT Project Metadata Development



use	
savedas	automatically extract: short filename assigned during SDSC web crawl
entryno	[maybe not needed?]
url	automatically extract: original URL of the record/ resource
filename	
descrtype	
descrauth	Deken, Jean
descrdate	
copy	Preservation
format	automatically extract: format information (doc, gif, jpg, pdf, tiff, xls, etc.) Repeatable field
filesize	automatically extract: #KB, #MB, etc.
meastype	
meascount	
storloc	sfs-disk-pat
stormed	
remarks	

Screen 2 of 2

PAT Project Metadata Development

- Applied metadata “skeleton” to some records
- Revised / iterated elements
- Developed / refined definitions
- Searched literature
  - Bibliography on project web site
  - Hodge, Gail et al. A Metadata Element Set for Project Documentation. Science & Technology Libraries Volume: 25 Issue: 4

PAT Project Metadata Development

- Categorized elements

Injected/ injectable:

- added to digital object
- based on outside information / outside needs

Extracted / extractable:

- information inherent in the digital object
- able to be obtained from it automatically (in theory)

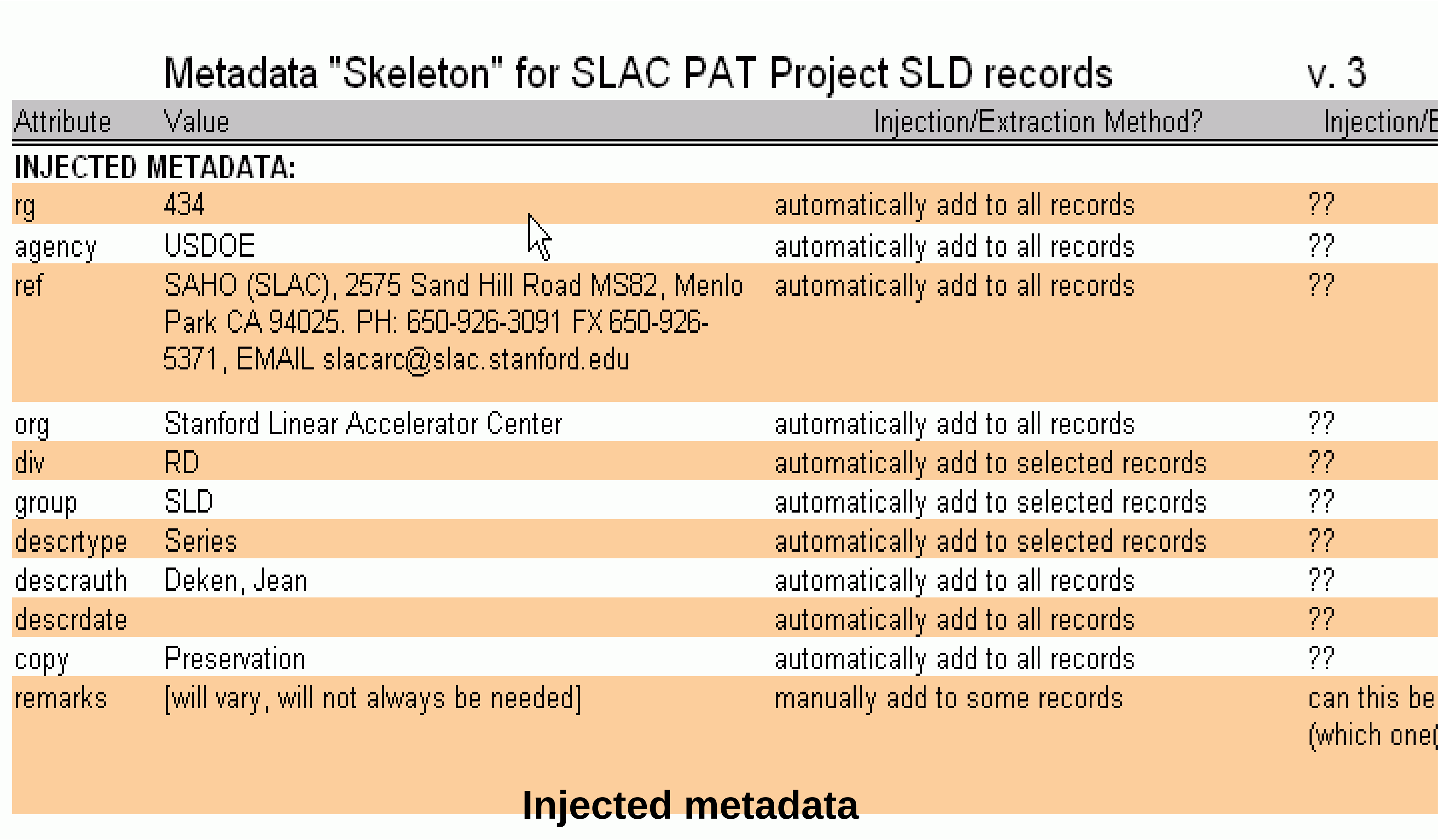
PAT Project Metadata Development

- Classified elements
  - **slac.gov**
    - Recordgroup, agency, referenceby, schedule, series, description, retention
  - **slac.creator**
    - Organization, division, group, person, owner
  - **slac.description**
    - Type, by, date, remarks, local, use, webplatform, webserver, format, filesize

PAT Project Metadata Development

- Classified elements
  - **Slac.identifier**
    - Copy, contmgt, websitename, url, filename, storagelocation, persistent
  - **Slac.capture**
    - Tool, settings, sitemap, date, contact, remarks
  - **Slac.pawn**
    - UMD – UMIACS test software “PAWN”
    - Recordset, category
  - **Slac.date**
    - Begun, modified

PAT Project Metadata Development



Metadata "Skeleton" for SLAC PAT Project SLD records v. 3

Attribute	Value	Injection/Extraction Method?	Injection/E
INJECTED METADATA:			
rg	434	automatically add to all records	??
agency	USDOE	automatically add to all records	??
ref	SAHO (SLAC), 2575 Sand Hill Road MS82, Menlo Park CA 94025. PH: 650-926-3091 FX 650-926-5371, EMAIL slacarc@slac.stanford.edu	automatically add to all records	??
org	Stanford Linear Accelerator Center	automatically add to all records	??
div	RD	automatically add to selected records	??
group	SLD	automatically add to selected records	??
descrtype	Series	automatically add to selected records	??
descrauth	Deken, Jean	automatically add to all records	??
descrdate		automatically add to all records	??
copy	Preservation	automatically add to all records	??
remarks	[will vary, will not always be needed]	manually add to some records	can this be (which one(

Injected metadata

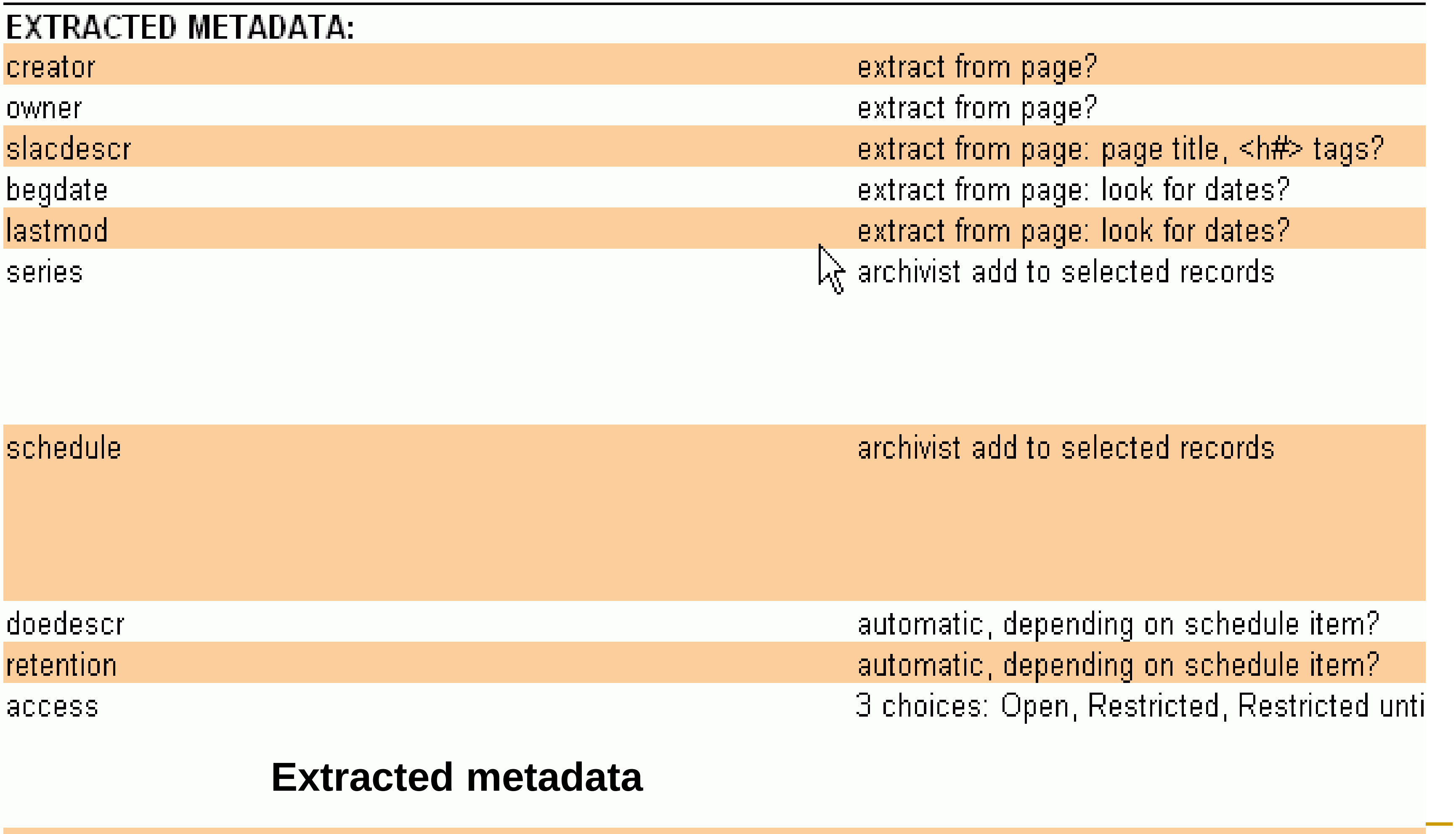
PAT Project Metadata Development

- Elements injected at the folder level (part 1):
  - slac.gov.recordgroup: 434
  - slac.gov.agency: USDOE
  - slac.gov.referenceby: SAHO (SLAC),2575 Sand Hill Road MS82, Menlo Park CA 94025.PH:650-926-3091 FX:650-926-5371 EMAIL:slacarc@slac.stanford.edu
  - slac.gov.schedule: N1-434-96-9,Item1.A.1
  - slac.gov.retention: Permanent
  - slac.creator.organization: Stanford Linear Accelerator Center

PAT Project Metadata Development

- Elements injected at the folder level (part 2):
  - slac.creator.division: RD
  - slac.creator.group: SLD
  - slac.description.type: Series
  - slac.description.by: Jean Deken
  - slac.description.date: [current date: yyyy.mo.day]
  - slac.identifier.copy: Preservation
  - slac.identifier.websitename: Introduction to the SLD Collaboration
  - slac.capture.tool: SDSC crawl tool written by C. Cowart

PAT Project Metadata Development

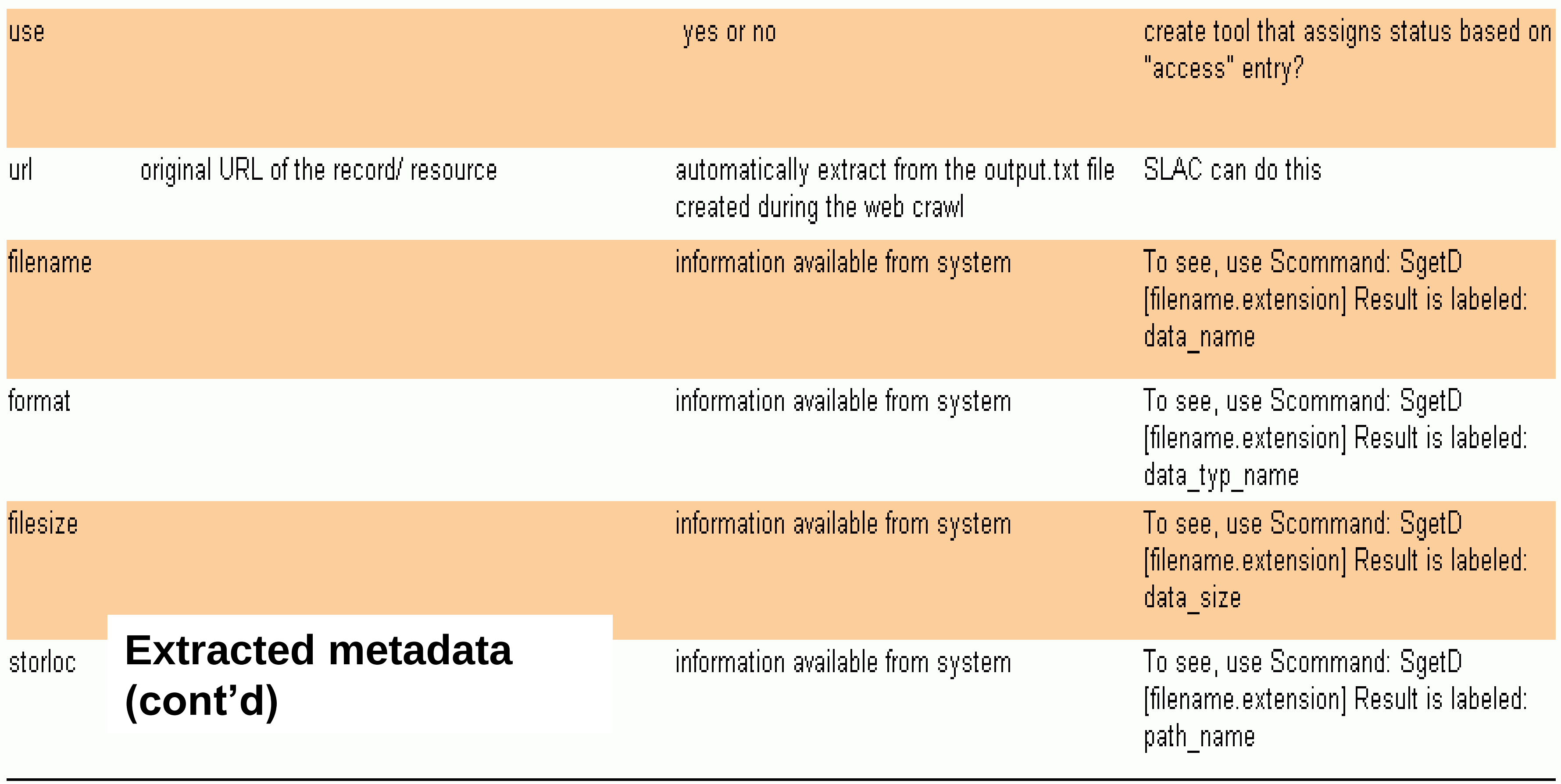


EXTRACTED METADATA:

creator	extract from page?
owner	extract from page?
slacdescr	extract from page: page title, <h#> tags?
begdate	extract from page: look for dates?
lastmod	extract from page: look for dates?
series	archivist add to selected records
schedule	archivist add to selected records
doedescr	automatic, depending on schedule item?
retention	automatic, depending on schedule item?
access	3 choices: Open, Restricted, Restricted unti

Extracted metadata

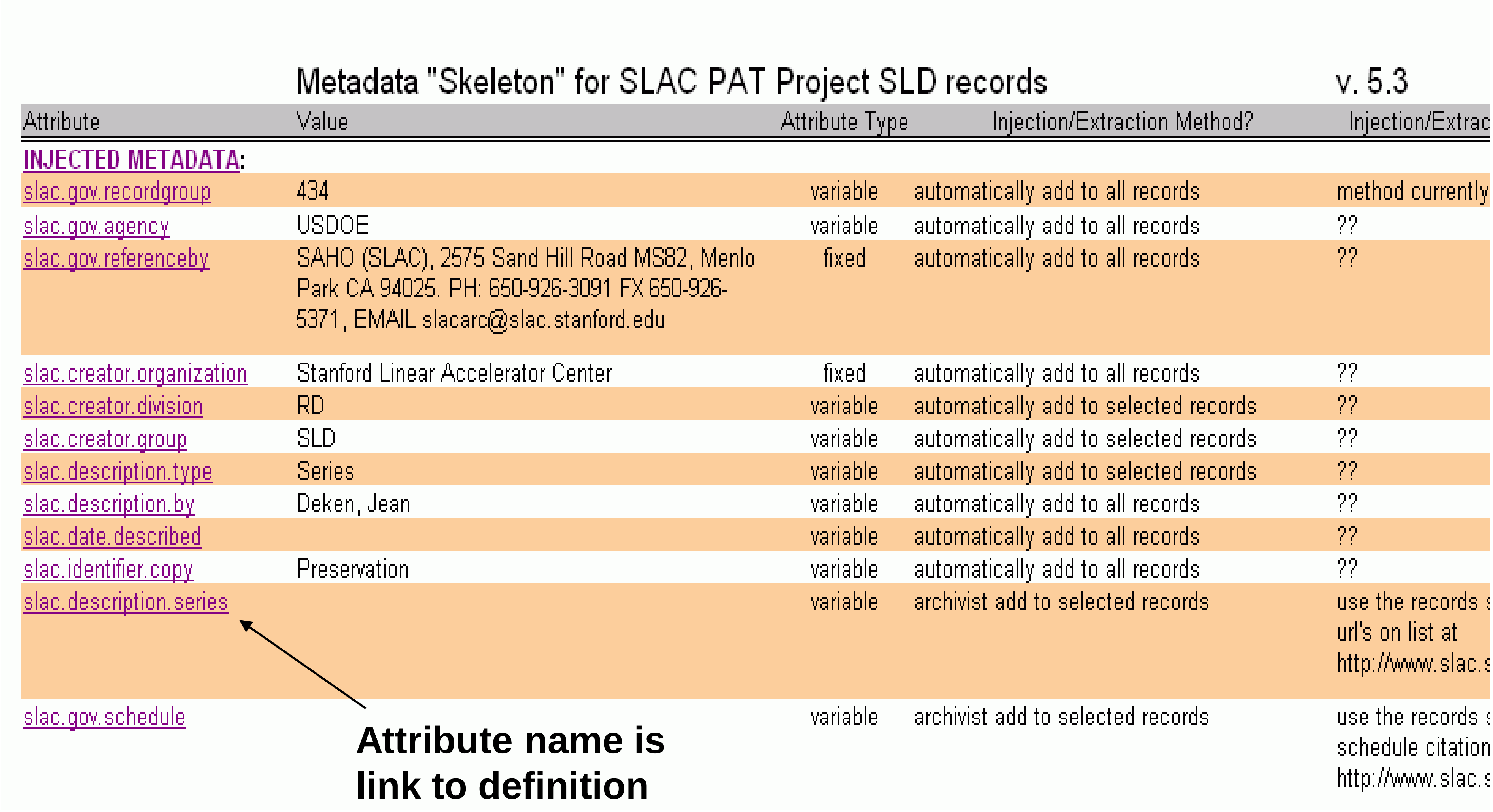
PAT Project Metadata Development



use		yes or no	create tool that assigns status based on "access" entry?
url	original URL of the record/ resource	automatically extract from the output.txt file created during the web crawl	SLAC can do this
filename		information available from system	To see, use Scommand: SgetD [filename.extension] Result is labeled: data_name
format		information available from system	To see, use Scommand: SgetD [filename.extension] Result is labeled: data_typ_name
filesize		information available from system	To see, use Scommand: SgetD [filename.extension] Result is labeled: data_size
storloc		information available from system	To see, use Scommand: SgetD [filename.extension] Result is labeled: path_name

Extracted metadata (cont’d)

PAT Project Metadata Development

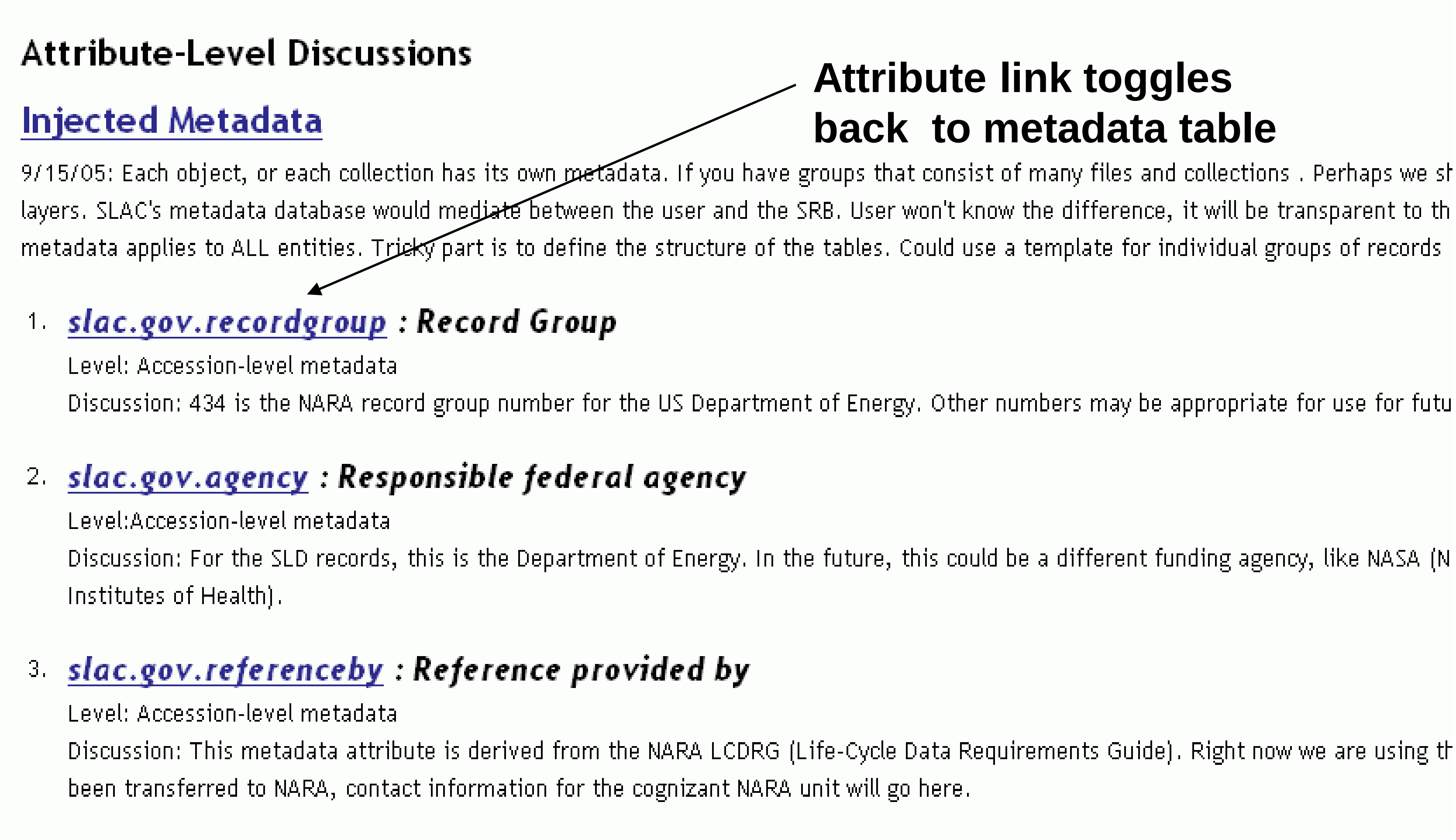


Metadata "Skeleton" for SLAC PAT Project SLD records v. 5.3

Attribute	Value	Attribute Type	Injection/Extraction Method?	Injection/Extrac
INJECTED METADATA:				
slac.gov.recordgroup	434	variable	automatically add to all records	method currently
slac.gov.agency	USDOE	variable	automatically add to all records	??
slac.gov.referenceby	SAHO (SLAC), 2575 Sand Hill Road MS82, Menlo Park CA 94025. PH: 650-926-3091 FX 650-926-5371, EMAIL slacarc@slac.stanford.edu	fixed	automatically add to all records	??
slac.creator.organization	Stanford Linear Accelerator Center	fixed	automatically add to all records	??
slac.creator.division	RD	variable	automatically add to selected records	??
slac.creator.group	SLD	variable	automatically add to selected records	??
slac.description.type	Series	variable	automatically add to selected records	??
slac.description.by	Deken, Jean	variable	automatically add to all records	??
slac.date.described		variable	automatically add to all records	??
slac.identifier.copy	Preservation	variable	automatically add to all records	??
slac.description.series		variable	archivist add to selected records	use the records s url's on list at http://www.slac.s
slac.gov.schedule		variable	archivist add to selected records	use the records s schedule citation http://www.slac.s

Attribute name is link to definition

PAT Project Metadata Development



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 sh
layers. SLAC's metadata database would mediate between the user and the SRB. User won't know the difference, it will be transparent to th
metadata applies to ALL entities. Tricky part is to define the structure of the tables. Could use a template for individual groups of records

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 futu

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 (N
   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 th
   been transferred to NARA, contact information for the cognizant NARA unit will go here.

Attribute link toggles back to metadata table

Some Conclusions …

- Start from where you are
- Accept that metadata is evolving…
- Follow standards that make sense for you
  - Your repository
  - Your resources
  - Your needs
- Be systematic
- Document, document, document !!

Contact Information …

SLAC PAT / TPAP project website:

http://www.slac.stanford.edu/history/projects.shtml

Jean Deken

jmdeken@slac.stanford.edu