Describing resources I: MARC

CERN-UNESCO School on Digital Libraries
Rabat, Nov 22-26, 2010

Annette Holtkamp
CERN

Metadata

- data about data
- structured, descriptive information about a resource
- key to resource discovery
- useful for records management, archiving
- metadata element:
  - field for storing specific information (like title)
- metadata value:
  - content of one metadata element
  - may be taken from predefined vocabulary

Metadata types

- descriptive
  - identification and retrieval
    - title, author, abstract…
- structural
  - presentation
    - chapters of a book,…
- administrative
  - management and preservation
    - version, technical info, access control

Metadata schema

- defined set of metadata elements
- serving a specific purpose
  - e.g. specific discipline, type of resource
- specify name and meaning of its elements
- optional rules
  - content, representation, element values, syntax…
- metadata standards
  - MARC, Dublin Core…

MARC

MAchine Readable Cataloguing

- international standard for representing and communicating bibliographic records
- developed in the 60s
- catalogue card oriented
- high degree of complexity
  - all purpose
- basis of most library catalogs, huge user base

http://www.loc.gov/marc

MARC21

- evolution of MARC
- combination of US and Canadian MARC formats
- internationalization
- Unicode
  - standard for encoding and representing text in multilingual environments
  - > 100k characters
  - 93 scripts

Formats

- bibliographic
  - books, periodicals, computer files, maps, music, visual materials, mixed materials
- authority
  - authorized forms of names and subjects
- classification
  - classification numbers or index terms
- holdings
  - single-part, multi-part and serial items
  - copy-specific information
- community information
  - non-bibliographic ressources of a community
    - scientists, institutions, conferences

Bibliographic record - structure

- Leader
  - basic information about the item
    - e.g. type of material
  - information for the processing of the record
    - record length, status, character coding scheme…
  - fixed field, first 24 character positions of each bibl record
- directory
  - Computer-generated index to location of control and data fields
  - 12 characters at position 24
- control fields 00x
  - 001 – control number / system nr
  - 003 – control number identifier, MARC code of organization 003 SzGeCERN
  - 005 – date and time of latest transaction, version identifier
  - 008 – general information on material
    - e.g. 1-character alphabetic code at pos 23 specifying form of material (b: microfiche)
- data fields

Data fields - structure

- three-character numeric tags
- often repeatable
- up to 2 indicators
  - interpret or supplement the data found in the field
  - lowercase alphabetic or numeric character
- numerous subfields
  - lowercase alphabetic or numeric character
  - independently defined for each field
  - sometimes repeatable

Data fields - classes

0xx – control, number and code fields

1xx – main entry fields

2xx – title/publication fields

3xx – physical descriptions

4xx – series fields

5xx – note fields

6xx – subject fields

7xx – added entry fields

8xx – series, holdings, location…

9xx – reserved for local implementation

complete list at http://www.loc.gov/marc/bibliographic/

01x-04x – Number and code fields

010 – Library of Congress control number

020 – ISBN

- $a – ISBN
- $u – medium (non-standard)
- 020__ $$a9783540632931$$uprint version, paperback

022 – ISSN

024 – other standard identifiers (e.g. DOI)

041 – language code

- e.g. eng for English

05x-08x – classification and call nr fields

050 – Library of Congress call number

080 – UDC Universal Decimal Classification number

080__ $$a514.763

082 – DDC Dewey Decimal Classification number

084 – other classification number

088 – report series number

088__ $$aCERN-PH-TH-2010-240

1xx – Main entry

100 – Personal name

- $a - personal name
- $e – relator term
- $u – affiliation
- $i – author id (undefined subfield, used by Inspire)

100__ $$aClerbaux, Barbara$$eed.$$iINSPIRE-00314890$$uBrussels U.

110 – Corporate name

- $a – corporate name
- $b – subordinate unit
- $g – acronym

100__ $$aCentre des Recherches Nucleaires$$gCERN

2xx – title information

245 – Title

- $a – Title
- $b – subtitle
- 245__ $$aRemoving The Haystack$$bThe CMS Trigger and Data Acquisition Systems

246 – varying form of title

242 – translated title

250 – edition statement

- $a – edition

260 – publication, imprint

- $a – place of publication
- $b – name of publisher
- $c – date of publication
- 260__ $$aLondon$$bImperial College Press$$c2010

3xx – Physical description

300 – Physical description

$a – pagination, duration in minutes…

$b – other physical characteristics

300__ $$aStreaming video ; 2 DVD video$$b720x576 4/3, 25

4xx – Series information

490 – series

$a – series

$v – volume information

490__ $$aLecture Notes in Mathematics$$v1358

5xx – note fields

500 – general note

502 – dissertation note

506 – restrictions on access

- indicator 1
  - 0 – no restriction
  - 1 – restrictions apply
- $a – terms governing access
- $d – authorized users
- 5061_ $$aRestricted$$dais-users [CERN]

520 – summary

- $a – summary (abstract)

540 – terms governing use and reproduction

- $a – terms governing access, e.g. CC license
- $b – body imposing these terms, e.g. publisher
- $u – URI

542 – copyright information

- $d – copyright holder
- $f – copyright statement
- $g – copyright date
- $u – URI

6xx – subject fields

650 – topical terms

- indicator 1: level of subject
  - 1 – primary
  - 2 – secondary
- indicator 2: thesaurus
  - 0 – Library of Congress subject heading
  - 7 – Source specified in subfield $2
- $a – topical term or geographic name
- $2 – source

65017 $$2arXiv$$aParticle Physics - Theory

653 – index term

- $a – uncontrolled term (e.g. author keywords)
- $9 – source (e.g. author)

6531_ $$9CERN$$acomputer networks

69x – local subject access fields

690C_ $$aBOOK

7xx – added entry fields

700 – additional authors

710 – additional corporate names

76x-78x – linking entries

specify different relationships to a related item

773 – host item entry

- vertical relationship (book chapters, journal articles)
- $p – title (journal name)
- $v – volume
- $n – issue
- $y – year
- $c – pagination, article id
- $u – url
- $a – DOI
- $e – relationship code
- $w – record control nr of parent record

773__ $$a10.1088/1748-0221/5/09/P09003$$cP09003$$pJ. Instrum.$$v5$$y2010

787 – nonspecific relationship entry

- example: linking slides with proceedings contribution
- $w – record control nr of related record
- $i – relationship information (slides, conference paper…)

787__$$w1234567$$islides

85x – holdings, location

852 – location

- $a – location
- $b – sublocation or collection
- $c – shelving location

856 – electronic location and access

- indicator 1: access method
  - 4: http
- $q – electronic format type (html, pdf, jpeg…)
- $u – URI
- $y – link text

8564_ $$uhttp://arxiv.org/pdf/1011.1200.pdf$$yPreprint

9xx – local fields

999 – references

- $o reference number
- $m Miscellaneous
- $h authors
- $a DOI
- $u Uniform Resource Identifier
- $r report number
- $s journal reference

999C5$$o1$$hR.W. Robinett and J.L. Rosner$$sPhys. Rev. D 25 (1982) 3036$$a10.1103/PhysRevD.25.3036

Control subfields

Fields within a record may be linked via subfield 8 or 6:

$8 - Field link and sequence number

- $8 [linking number].[sequence number]\[field link type]
- linking number
  - occurs in subfield $8 in all fields that are to be linked
- sequence number
  - indicates the relative order for display of the linked fields
- field link type
  - code indicating the reason for the link

$6 – links fields that are different script representations of each other

Records are linked to authority records via subfield 0:

$0 - Authority record control nr or standard nr

Bibliographic record: web display

Information | References | Discussion | Fulltext

Article

Report number	arXiv:0801.1651 ; CERN-PH-TH-2008-004 ; FTPI-MINN-2008-01 ; UMN-TH-2008-2630
Title	**Sparticle Discovery Potentials in the CMSSM and GUT-less Supersymmetry-Breaking Scenarios**
Author(s)	Ellis, Jonathan Richard (CERN) ; Olive, Keith A (Univ. Minnesota, Minneapolis, MN, USA) ; Sandick, Pearl (Univ. Minnesota, Minneapolis, MN, USA)
Imprint	11 Jan 2008. - 20 p.
In:	*J. High Energy Phys.* 08 (2008) 013
Subject category	hep-ph
Abstract	We consider the potentials of the LHC and a linear e^+e^- collider (LC) for discovering supersymmetric particles in variants of the MSSM with soft supersymmetry-breaking mass parameters constrained to be universal at the GUT scale (CMSSM) or at some lower scale M_{in} (GUT-less models), as may occur in

Bibliographic record: MARC

001__ 1080272
003__ SzGeCERN
005__ 20081003111503.0
0248_ $$aoai:cds.cern.ch:1080272$$pcerncds:CERN
035__ $$9arXiv$$aoai:arXiv.org:0801.1651
035__ $$9SPIRES$$a7620977
037__ $$aarXiv:0801.1651
041__ $$aeng
088__ $$aCERN-PH-TH-2008-004
088__ $$aFTPI-MINN-2008-01
100__ $$aEllis, Jonathan Richard$$uCERN
245__ $$aSparticle Discovery Potentials in the CMSSM and GUT-less Supersymmetry-Breaking Scenarios
269__ $$c11 Jan 2008
300__ $$a20 p
520__ $$aWe consider the potentials of the LHC and a linear e^+e^- collider (LC) for discovering supersymmetric…
595__ $$aOA
65017 $$2arXiv$$ahep-ph
690C_ $$aARTICLE
690C_ $$aCERN
700__ $$aOlive, Keith A$$uUniv. Minnesota, Minneapolis, MN, USA
773__ $$c013$$pJ. High Energy Phys.$$v08$$y2008
8564_ $$uhttp://arxiv.org/pdf/0801.1651.pdf$$yFulltext
8564_ $$uhttp://cdsweb.cern.ch/record/1080272/files/jhep082008013.pdf$$ySISSA/IOP OA article

Conference record

Information | References | Discussion | Fulltext

Conference

Conference title	**24th International Symposium on Lepton Photon Interactions at High Energies**
Related conference title(s)	Lepton Photon 09
Date(s), location	17 - 22 Aug 2009, DESY, Hamburg, Germany
Conference contact	email: lp09@desy.de
Imprint	2009
Subject category	Particle Physics

external link:

Conference home page

Contributions to this conference in CDS

Particle Physics in the LHC Era and beyond

by Altarelli, Guido

Top Quark Pair Production Cross section at LHC with ATLAS

by Doxiadis, AD

MARC XML

- XML schema based on MARC21
- developed by Library of Congress
- XML: Extensible Markup Language
  - set of rules for encoding arbitrary data structures
  - separates content (metadata) from presentation

MARC XML: elements

- <collection>
  - file of several records
- <record>
  - delineates records within a collection
- <leader>
  - MARC leader data string
- <control field>
  - MARC control field data string
- <data field>
- <subfield>

MARC XML: datafield

- MARC field tags and indicators are expressed as attributes of a datafield element

  <datafield tag=”100” ind1=”1” ind2=” ”>

- Each subfield a separate element
  - subfield code as attribute

  <subfield code=”a”>…</subfield>

Example: book editor

```
<datafield tag="100" ind1=" " ind2=" ">
  <subfield code="a">Clerbaux, Barbara</subfield>
  <subfield code=“e">ed.</subfield>
  <subfield code=“i”>INSPIRE-00314890</subfield>
  <subfield code="u">Brussels U.</subfield>
</datafield>
```

MARC XML

- aim: easy sharing of bibl info
- easy access at subfield level
- lossless conversion from MARC21
- manipulated and transformed via XSL stylesheets
  - Extensible Stylesheet Language
- “bus” for conversion between different standards