Digitisation & Metadata Overview

Thierry Bouche

Cellule MathDoc & institut Fourier, Grenoble

MSRI workshop, April 15th 2005, Berkeley

Outline

Definition

What is metadata?

1. Descriptive material concerning your data.
2. Additional data that was not there in the first place.
3. Anything you're interested in,
   that is not what you consider being 'data'.

   Is OCR (meta)data?

Metadata in a digitisation project

There are three sets of metadata:

Internal Administrative metadata.
Production metadata generated at scan time.
Archiving metadata for the long term.
Rights metadata.

Exposed The subset of all the collected or added metadata that you expose to your users.
This is what allows you to build the pieces of your user interface.
Although 'exposed', some of it may be hidden from the user.

Exported The description of your collections that you're eager to see used elsewhere (mostly because you expect better visibility, *i.e.* links pointing to your resources).

The NUMDAM example

- File names and file systems
- Persistant URLs
- The NUMDAM DTD:
  - Internal metadata
  - Exposed metadata

File names and file systems

- One highly important issue for archiving is
  a consistant naming and organising system for files.
- If everything else is lost or unreadable,
  the file system should be as self explanatory as conceivable.
- Redundance is not a problem.
- It is preferable to dumb numbers whose signification
  depends on external resources that may vanish.

NUMDAM: File names and file systems

The journal article model in NUMDAM is the following:

- Journal identifier (acronym).
- Year of publication.
- Series.
- Volume.
- Issue (identifies the paper volume).
- First page.
- Article order if not alone on that page.

ASENS/ASENS_1959_3_76_3/ASENS_1959_3_76_3_185_0/
These are the directories on CD-ROM
as well as at archive.numdam.org.

NUMDAM: File system for archiving

- ASENS/
  - Volume directories.
  - Global info for the journal.
- ASENS/ASENS_1959_3_76_3/
  - Article directories.
  - Monochrome TIFF files for all the inner pages.
  - Special TIFF files (grey or colour).
  - Cover TIFF pages.
  - XML metadata for the volume.
- ASENS/ASENS_1959_3_76_3/ASENS_1959_3_76_3_185_0/
  - Multipage TIFF file.
  - XML OCR file.
  - PDF and DjVu file.

NUMDAM: File system for posting

- ASENS/
  - Global info for the journal.
- ASENS/ASENS_1959_3_76_3/
  - Article directories.
  - First cover JPEG page.
- ASENS/ASENS_1959_3_76_3/ASENS_1959_3_76_3_185_0/
  - User PDF and DjVu file.
  - HTML abstract.

NUMDAM: Persistant URL

- Each item has a unique identifier.
- Each item has a persistant URL
  which is built as simply as possible from its identifier.
- By prefixing, you get another universal identifier.

http://www.numdam.org/item?id=ASENS_1959_3_76_3_185_0
oai:numdam.org:ASENS_1959_3_76_3_185_0

NUMDAM: Internal metadata

The NUMDAM DTD: Volume record

- Identifier (volume level)
- Journal
  - ISSN
  - Acronym
- Series
- Volume
  - Number
  - Year
- Issue
- Editor (of a specific issue)
- Publisher (rights owner)
- Pages (number)
- Digitisation infos
  - Physical identifier (unused)
  - Creation date
  - Special (anything special)
  - Note (anything worth noticing)

NUMDAM: Internal metadata

The NUMDAM DTD: Special items

- Type (title page, summary, index, ads, special, art, . . .)
- First page (arabic number, many prefixes to deal with printer's idiosyncrasies. . .)
- Last page
- Note
- Monopage files list

NUMDAM: Exposed metadata

The NUMDAM DTD: Articles

- Identifier
- First page
- Last page
- Order (in the volume)
- Order (on the first page)
- Author
  - Last name
  - First name
  - Affiliation
  - Email
- Contributor (translator, redactor, . . .)
- Title (repeatable with a lang attribute)
- Main language
- Relations
  - Is corrected by (and vice versa)
  - Is completed by (and vice versa)
- Note
- Monopage files list
- OCR file

NUMDAM: Exposed metadata

The NUMDAM DTD: Article's bibliographies

- Bibliography
  - bibitem
    - Author
      (First name, last name)
    - Title
    - Year
    - First page

[Continued]

- Optional tags
  - Journal
  - Publisher
  - Location
  - Collection
  - ...

Sharing metadata

There are three circles of metadata:

Complete Everything you know about your collections.
Preliminary catalogue, captured at scan time, or added later. Reserved for insiders of the project.

Published The subset that feeds your website.
Typically under copyright and protected from batch downloads. Serves your users with specific features.

Public The subset that you give away freely.
Should be accurate enough to be useful.
Should be simple enough to be usable at all.

The border line between those circles are political or economical choices. Not so easy to agree upon?

Motivations

- You want that every possible user knows the availability of your collections.
- You can't force them to come to your site and use your interface.
- Working mathematicians with a subscription use MathSciNet or Zentralblatt.
- Historians use hand made lists or Jahrbuch.
- Everybody relies upon rumour, fame or chance.
- They'll fall back googling anyway.

Conjectures

- The distance between you and your possible user might be infinite if you rely on your fame and he relies on his trusted sources of information.
- It is however most probable that the distance amounts to *one* or *two* (kind of Erdős distance).

Solutions

1. Give away your basic catalogue for free.
2. Harvesting and updating must be automated.
3. Make it as detailed as you can afford to.
4. Never forget the “forget” (XML/XSLT) functor!

NUMDAM experiences

- On top of our web server (EDBM), Claude Goutorbe has built an OAI server.
  - It was integrated by tenth of harvesters overnight.
  - Among them, internet search engines like Yahoo, MSN, Excite.
- Under Google's request, we built a "cloaked" interface so that they could index the full OCRed text of articles.
  - Compare the results for "Théorie unitaire du champ physique" in Google and Yahoo!
  - Usage statistics have more than doubled in one month.
  - Searched expressions tend to show that there is not so much noise.

NUMDAM's OAI server

- Sits on top of our web server (EDBM), thus only provides a subset of our published metadata.
- Conforms to OAI-PMH 2.0.
- Provides one *set* per journal, two metadata formats
  - Dublin Core (mandatory)
  - minidml (experimental)

The minidml schema

The 'Article' element in minidml:

- Author **R**
- Title **R**
- Abstract **R**
- Keywords **R**
- Main language
- MSC [2000]
- Citation
- Pages
- Rights
- Reviewid **R**
- Identifier **R**
- Format **R**
- ISSN
- Abbrev
- Jtitle
- Jhome
- Series
- Volume
- Issue
- Date
- Publisher
- Provider

A proposed DML infrastructure

Each digitisation centre should set up:

- A unique identifier for each item.
- An associated persistant URL.
- Structured basic metadata served through OAI.

Then anyone can build an integrated access portal to whatever collection.

Towards an xdml schema

General prinicples.

- Two main objectives:
  1. Provide a basic catalogue so that people can add it to an existing database and set up links.
  2. Make it possible to match an existing reference against your catalogue to decide whether it's already there or not.
- Thus: accuracy, simplicity and versatility.
  1. Don't rely on heuristics to parse your data. Make it clean.
  2. Be explicit on every decision you made (text and math encodings, *e.g.*)
  3. Prefer complex well structured fields over flat structure with misleading element names.
  4. But keep in mind that conversion from and to legacy formats (bibtex. . .) should be possible.

An xdml schema

The basic item types should follow bibtex legacy (bold=mandatory):

- Article (**author**, **title**, **journal**, **year**, volume, number, pages, month, note)
- Book (**author** or **editor**, **title**, **publisher**, **year**, volume, series, address, edition, month, note)
- Proceedings (**title**, **year**, editor, publisher, organization, address, month, note)
- Inbook (**author** or **editor**, **title**, **chapter**, **pages**, publisher, year volume, series, address, edition, month, note)
- Inproceedings (**author**, **title**, **booktitle**, **year**, editor, pages, organization, publisher, address, month, note)
- Thesis (**author**, **title**, **school**, **year**, address, month, note)
- Unpublished (**author**, **title**, **note**, month, year)
- Misc (author, title, howpublished, month, year, note)

A tentative xdml specification

A core set for every item:

- Creator **R** (Given, Middle, Von, Last, Sort, String, URI, attr: author, editor, translator, redactor)
- Title **R** (attr: lang, text-encoding, math-encoding)
- Year
- Main language
- Abstract **R** (attr: lang, text-encoding, math-encoding)
- Keywords **R** (attr: lang, text-encoding, math-encoding)
- MSC **R** (attr: revision)
- Review **R** (attr: service)
- Identifier **R** (attr: scheme)
- Format **R** (Mime-Type, URL, Byte-size)
- Provider (Name, Location, URL, Policy)
- Rights
- Citation (depends on item type)

An xdml schema

Example: Creator, type=author

- **Name string** (Charles de La Vallée Poussin)
- Given name (Charles)
- Middle name ( )
- Von name (de)
- Last name (La Vallée Poussin)
- Sort name (La Vallée Poussin, Charles de)
- URL (serving as kind of identifier)?

An xdml schema

Example: Citation, type=article

- **Citation string**
- Journal
  - **Title**
  - Series
  - Abbreviation
  - Identifier **R** (attr: issn, uri)
  - Home
- Volume
- Issue
- Pages
  - **Collation**
  - First (attr: sysnum,...)
  - Last (attr: sysnum,...)
  - Count

Author (String) + Title + Citation (String) = full printed reference.

An xdml schema

Example: Citation, type=book

- **Citation string**
- Publisher
  - Name
  - Location
  - Home
- Pages
  - **Collation**
  - Count
- Collection
  - **Title**
  - Abbreviation
  - Identifier **R** (attr: issn, uri)
  - Home
- Number

Author (String) + Title + Citation (String) = full printed reference.