GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Using Metadata for the Interlinking of Digitized Mathematics

Thomas Fischer
State and University Library
Göttingen, Germany

Overview

- The Problem: Search vs. Access
- The Situation: Players and Communication
- Suggested Solutions:
  - Metadata Standards
  - Resolving Services
  - Registries

Description of the Problem

Two kind of "searches":

- Trying to find the answer to some question: What are the different differential structures on a 4-manifold?
- Trying to find an article: Where is "Characteristic numbers of 3-manifolds" by Milnor and Thurston available?

- The first kind requires well formulated searches against some appropriate databases.
- The second is in some sense trivial.

The latter question is the one this talk is dealing with:

If you know what you want, how do you get it?

Players: Sources of electronic literature

- Publishers
  - Publishers usually produce electronic versions of mathematical journals
  - Many publishers produce “backfiles”: retrodigitized versions of printed journals
- Preprint servers
  - Preprint servers collect electronic versions of mathematical papers since the early 90’s
  - There are several different of them with no unified structure
- Digitization centres
  - There are several (mostly national) initiatives to produce retro-digitized versions of historical and recent mathematics

Sources of printed mathematics literature

- Local libraries
  - Provide more or less extensive collections of published journals and books
  - Provide additional services if the desired material is not available
- Authors
  - May provide copies of their papers to interested researchers
- Remote libraries
  - Provide articles and books through inter-library loan
  - Some provide remote scanning services, delivering the scanned object to the user's desktop electronically
- Publishers and bookstores
  - Provide options to buy books

Researchers' interest

- Immediate access, if possible
- Most authoritative version
- No or no additional or lowest possible cost
- High quality of presentation

(order of importance may depend on institutional or personal preferences)

Searching for Milnor/Thurston:

- Try Google
- Try Google Scholar
- Try SpringerLink or ScienceDirect
- Zentralblatt Math: Enseign. Math., II. Sér. 23, 249-254 (1977) [ISSN 0013-8584]
- Google: “Enseign. Math”
- http://www.unige.ch/math/EnsMath/EM_en/welcome.html
- http://retro.seals.ch/cntmng?type=pdf&aid=c1:36817 (presented by seals: swiss electronic academic library service, with some problems …)

Problems

Background idea:

World Digital Mathematics Library

But now:

- No unified discovery or access scheme
- No uniform standards of reference
- No uniform quality standards

Not solved or not solvable?

Quality standards for retrodigitization seem to be hard to enforce:

- Different players: publishers, scientific communities, digitization centres
- Money involved in quality of scanning and administration of complex metadata
- Different scopes and long term orientation

Available building blocks

- Review journals (Mathematical Reviews, Zentralblatt MATH)
- Metadata standards (Dublin Core)
- Communication protocol (OAI-PMH)
- Reference standard (OpenURL)
- Willingness to co-operate

Goal: unified access

- Look up the requested paper in Zentralblatt or MathReviews (they might cover preprints and other ‘gray’ literature for that purpose)
- Receive a link to a resolving service
- Obtain the appropriate copy (digital or printed)

Necessary: communication network with sufficient precision

New developments in the world of metadata

- Dublin Core: Abstract Model
- Dublin Core Application Profile: the ePrint Application Profile
- minidml, a DC-based custom format
- Dublin Core Simple, enhanced by best practises

Dublin Core Abstract Model 1

- New: Description sets contain several related descriptions
- Allows distinct descriptions e.g. for Creator, Journal and Article
- With this, authority files become easier to build and manage
- http://dublincore.org/documents/abstract-model/

Dublin Core Abstract Model 2

- A *description set* is a set of one or more *descriptions* about one or more *resources*.
- A *description* is made up of one or more *statements* (about one, and only one, *resource*) and zero or one *resource URI* (a URI reference that identifies the *resource* being described).
- Each *statement* instantiates a *property/value pair* and is made up of a *property URI* (a URI reference that identifies a *property*), zero or one *value URI* (a URI reference that identifies a *value* of the *property*), zero or one *vocabulary encoding scheme URI* (a URI reference that identifies the *class* of the *value*) and zero or more *value representations* of the *value*.

```
<descriptionSet>
  <description
    resourceURI="http://arxiv.org/abs/math/0612096">
    <title>Constructing Smooth Loop Spaces<title>
    <creator descriptionRef="theAuthor"/>
    <date>2006-12-04</date>
    <subject>Differential Geometry; Algebraic
    Topology</subject>
  </description>
  <description descriptionId="theAuthor">
    <firstname>Andrew</firstname>
    <lastname>Stacey</lastname>
    <affiliation>University of Sheffield</affiliation>
  </description>
</descriptionSet>
```

Eprints Application Profile 1

Eprint: a scientific or scholarly research text

(as defined by the Budapest Open Access Initiative)

- a DC Application Profile for describing an eprint
- Each *description set* describes only one eprint (i.e. one ScholarlyWork entity).
- Extends the Dublin Core set by numerous additional fields related to scholarly work
- Uses concept of *related description* for entities that allow a distinctive separate description, e.g. Creator, Funder, Affiliated Institution
- Incorporates parts of the "Functional Requirements for Bibliographic Records" (FRBR), in particular the distinction between *Work, Expression, Manifestation* and *Item.*
- http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile

Why FRBR?

FRBR disentangles some problems with bibliographic references:

- The **Work** is the abstraction of the original product, the ideas.
- An **Expression** is a version of this, e.g. the original one, a translation, the third revision…
- A **Manifestation** is a preprint or a printed and/or digital version in a journal or book.
- An **Item** (copy) is the actual physical object: the article in the journal on the shelf, the specific digital copy on a particular server.

Different properties refer to different levels, think e.g. of a title and a title of the translation, page numbers, publisher, URL…

Eprints Application Profile 2

- Provides as comprehensive a format for describing a scholarly work as desired
- Is well adapted to *electronic sources*, not based on printed matter
- Can be mapped to Dublin Core simple (with some losses)
- Probably the optimal data format available, but not easy to implement

The minidml format

- A DC-based metadata format enriched by the separation of some elements
- Beyond DC simple, minidml provides references:
  - different identifier schemes, e.g. <identifier scheme="oai">
  - <citation>Ann. Inst. Fourier 1, 1-4 (1949)</citation>
  - plus separate subfields for citation
  - <reviewid> giving MR and Zbl numbers
- With 20+ elements plus some schemes the most relevant information on a scholarly article can be captured
- http://minidml.mathdoc.fr/

DC Simple enriched by “Best Practices” 1

“Recommended Practice for Creating Unqualified Dublin Core Records, for Mathematical Literature”
(unwieldy title!) in preparation by

- Thierry Bouche, NUMDAM (Numérisation de documents anciens mathématiques)
- Thomas Fischer, Staats- und Universitätsbibliothek (SUB), Göttingen
- Claude Goutorbe, Cellule MathDoc, Grenoble
- David Ruddy, Project Euclid, Cornell University Library

- Based on experience with and analysis of different OAI metadata schemes used by digitization centres
- Goal: provide rules for DC simple that make the OAI metadata most useful

DC Simple enriched by “Best Practices” 2

Some suggestions:

- Use UTF-8 for special characters outside of mathematical formulas, not the TeX encoding
- Use prefixes to clarify the meaning of data, e.g. “isbn:”, “msc:”, “bibliographicCitation:”
- Use last name first rule for names; use only one version of the name for each author
- Link to reference journals in the <relation> element, using appropriate prefix, e.g. “mr:”, “zbl:”, “jfm:”
- Give full bibliographic citation information in identifier field

Bibliographic Citation and OpenURL

OpenURL: NISO standard Z39-88(2004)

- Basically of the form (without the line break):
  ctx_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&
  <key1>=<value1>&<key2>=<value2>&...&<keyN>=<valueN>
  Essentially a (search) request with standardized fields
- Used for
  - Encoding Bibliographic Citation Information in Dublin Core Metadata (http://dublincore.org/documents/dc-citation-guidelines/)
  - ContextObject in SPAN (COinS): Embedding Citation Metadata in HTML (http://ocoins.info/)
  - SFX: context-sensitive link server from Ex Libris (http://www.exlibrisgroup.com/sfx_openurl.htm)
  - crossRef: an infrastructure for linking citations across publishers (http://www.crossref.org/03libraries/16openurl.html)

Build a registry!

- OAI Data providers:

  Provide data with sufficient information, using a full format like the ePrint application profile or enriched format like minidml and modify the required DC Simple format according to the recommendations

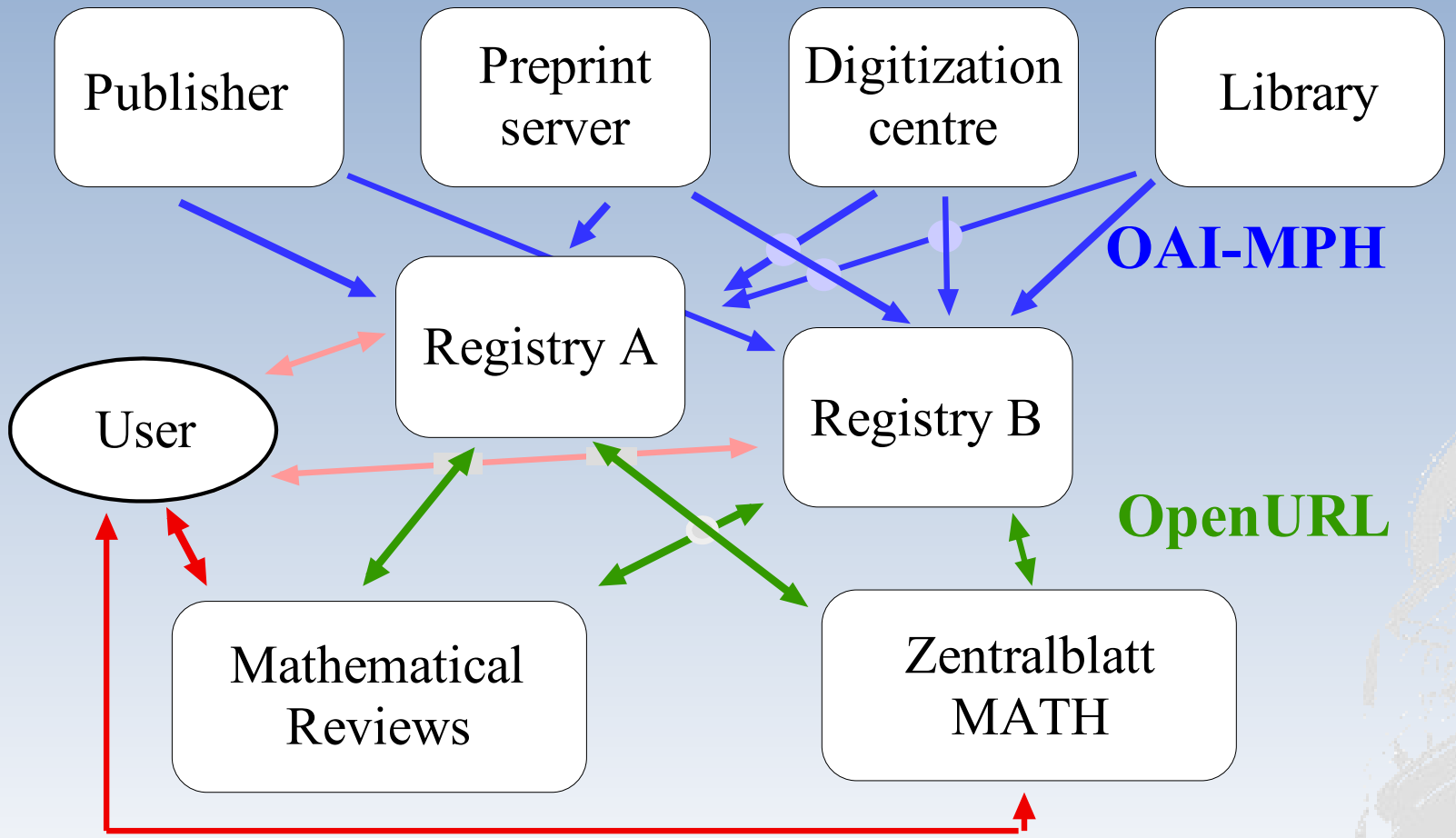
- OAI Service providers:

  Collect data from the available sources and provide an OpenURL service to retrieve the documents

- Review journals:

  Enrich the review data with an appropriate OpenURL, directed to a resolver at one of the registries

A possible communication scheme



Some basic tasks ahead:

- Digitization centres:
  - Get the pages straight and the page numbers right
  - Get full and correct data
- OAI Service providers:
  - Collect data and analyze and organize appropriately, in particular match different versions of the same article
- Review journals:
  - Unify the references to journals

Some references

- Andy Powell, Mikael Nilsson, Ambjörn Naeve, Pete Johnston: Dublin Core Abstract Model (http://dublincore.org/documents/abstract-model)
- Pete Johnston, Andy Powell: Expressing Dublin Core metadata using XML (30.5.2006) (http://dublincore.org/documents/2006/05/29/dc-xml)
- Digital Library Federation and the National Science Digital Library: Best Practices for Shareable Metadata (August 2005) (http://comm.nsdl.org/download.php/653/ShareableMetadataBestPractices.doc)
- JISC, UKOLN, cetis: EPrint Application Profile (September 2006) (http://www.ukoln.ac.uk/repositories/digirep/index/EPrints_Application_Profile)
- ViFa Math /VLib Math: http://www.ViFaMath.de/ (English version in preparation)
- Digitization Registry: http://DigReg.MathGuide.de/ (recommendations for OAI data to appear here)

Thank you for your attention!

Thomas Fischer
State and University Library
Göttingen, Germany
fischer@sub.uni-goettingen.de