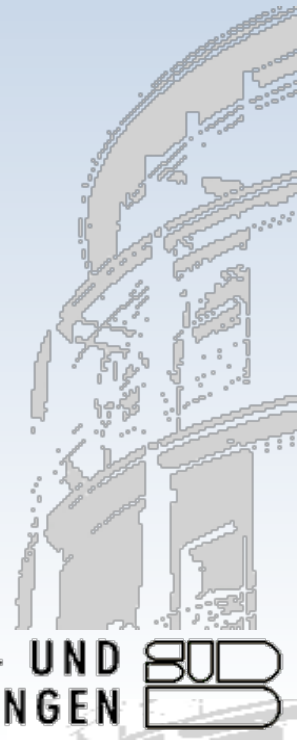
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Integrating OAI metadata: Experiences and challenges

Thomas Fischer
SUB Göttingen
fischer@sub.uni-goettingen.de



NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

Introduction

Last year, the Digitisation Registry (http://DigReg.MathGuide.de/) took the step from collecting Journal Information down to the Article Level:

- standard OAI collecting procedures
- parsed with Perl scripts.

I want to present some experiences from this process as well as some desiderata that sprang from this activity.

The result are rather more questions than answers…

Contents

1. Collecting Mathematics with OAI-PMH: Import and Normalisation
2. Authority Files:
   - Persons
   - Journals
3. Regaining Structure

1. Import Method

- Central Perl Script (oai2digreg.pl)
- Integrates tables:
  - of field conversion (dependent on the source)
  - of MIME Types
  - of Language and Country Codes
  - of Journals
  - of TeX2utf8
  - of TeXSymbols (optional)
- Contains subscript for each OAI data provider, e.g.
  - convertNumdamRecord
  - convertGDZRecord
  - …

Rationale 1: Peculiarities

Not all Dublin Core is simple, e.g.

GDZ (01.10.2007)

- <dc:identifier>ISSN:0037-1912</dc:identifier>

GDZ (25.02.2008)

- <dc:identifier.issn>ISSN:0037-1912</dc:identifier.issn>

Different standards

GDZ (DC Simple)

- <dc:source>Schein, B.M. | Relation algebras and function semigroups. | Semigroup Forum | 1970 | BAND: 1 | Seite: 1</dc:source>

Numdam (minidml)

- <citation>Ann. Inst. Fourier 29, no.1, 49-79 (1979)</citation>

Rationale 2: Normalisation

Tables for unified

- MIME Types
- Language Codes
- Project Codes

(easy)

Tables for normalised

- Characters (TeX dependant, not so easy)
- Journal identifiers (a little complicated)

TeX Characters

For Digitisation projects there seems to be a standard of Unicode text + TeX symbols. This should be made explicit for all digitisation metadata.

(But arXiv: Schr\"odinger, Schr"odinger, Schr{\"o}dinger, Schroedinger, Schrodinger)

TeX symbols of character type could be converted (e.g.: $\nabla \partial \in \forall$) but are probably best left alone

Numdam:

{\bf Z}, {\Bbb Z}, \mathbf{Z}, \mathbb {Z}, \bold Z, Z:

- same meaning
- symbol or coding different

Normalisation

For each Journal a unique identifier, e.g.

- Groupe de travail d'analyse ultramétrique=IdGAU
- Journal de théorie des nombres de Bordeaux=IdJTNB
- Journées équations aux dérivées partielles=IdJEDP

Essentially to get the separated pieces (articles) back together again, for easy search and filtering

(not quite sufficient, see part 3)

But probably also to bring different parts with different names together
(not quite satisfactory, see part 2)

2. Desideratum 1

Authority files for Persons:

- Not so many names: easy
- For different languages: not so easy
- From different nationalities: harder
- And different scripts: very hard?
- Big advantage: Unicode

Example: Virtual International Authority Files (VIAF)

Project started 2003 by Library of Congress, OCLC, DNB

Latest update 2006 at phase 1 of 4

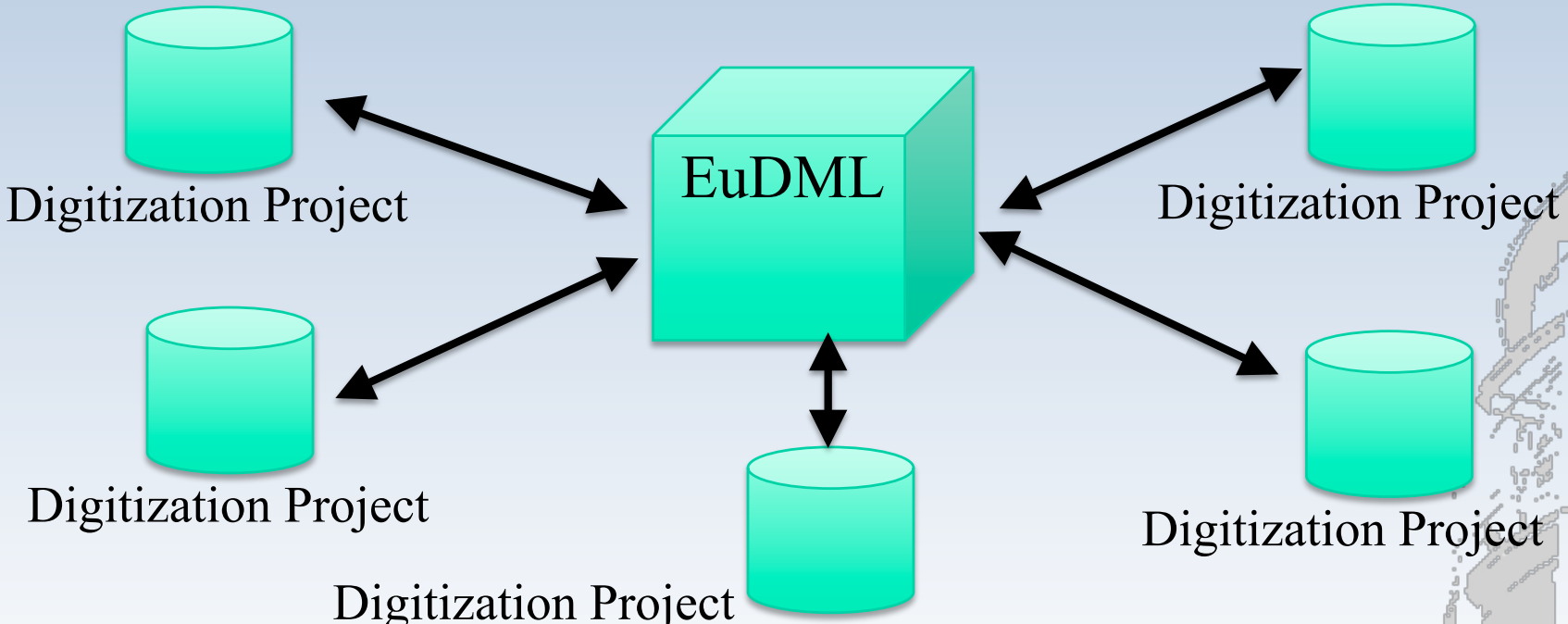
Desideratum 2

Authority files for journals:

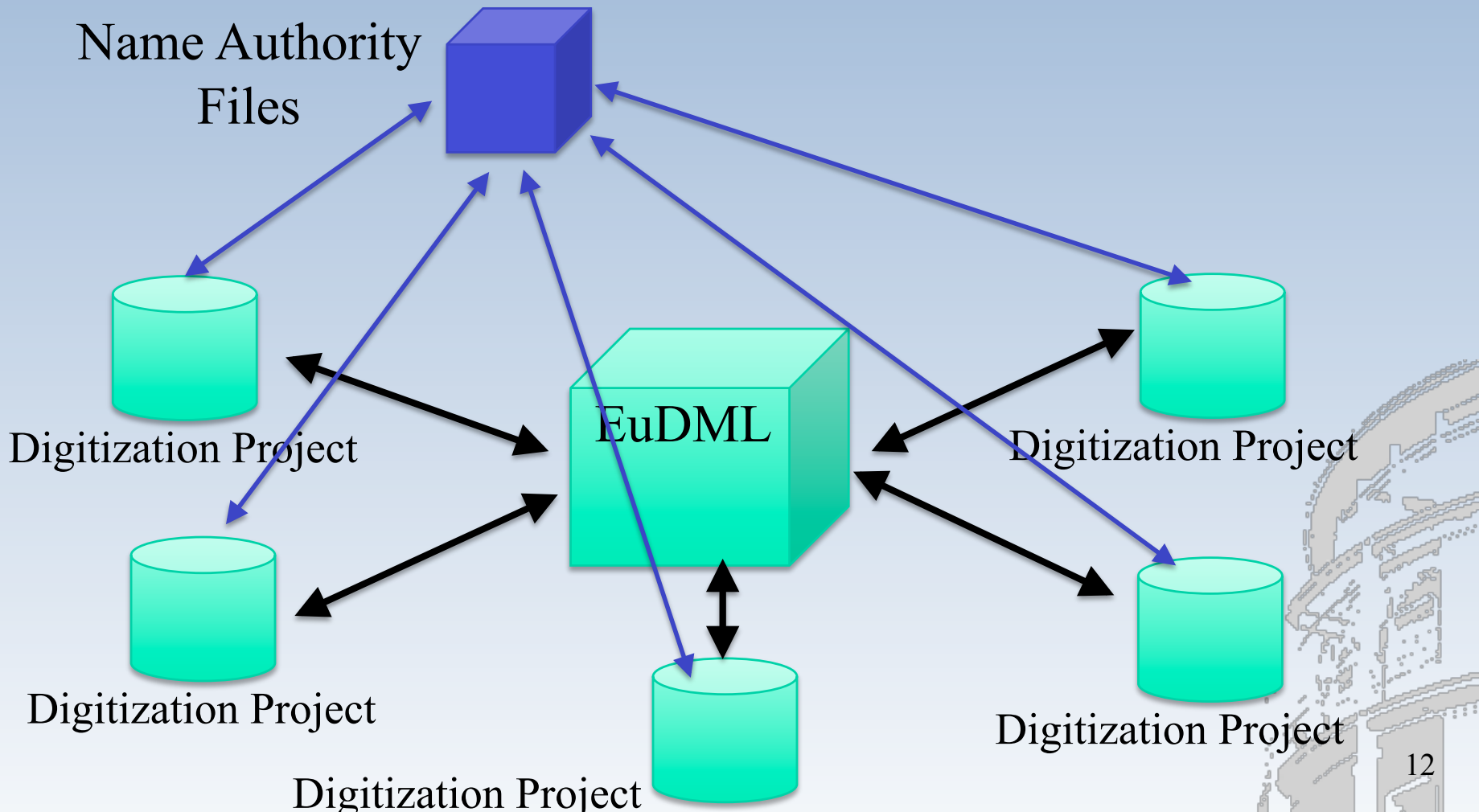
- Unified reference for journals
- Cross-reference for all relevant identifiers:
  - Full name(s?) — Journal für die reine und angewandte Mathematik
  - Abbreviated name(s?) — J. Reine Angew. Math.; J. f. reine u. angew. Math.
  - Nickname(s) — Crelle's Journal
  - ISSN for paper — 0075-4102
  - ISSN for digital — 1435-5345
  - ISSN for retro-digital — 1435-5345 (?)
  - Zbl-identifier — ?
  - MR-identifier — JReiAM (?)
  - NUMDAM-ID — –
  - + …

Example: Société Littéraire de Bruxelles (local copy)

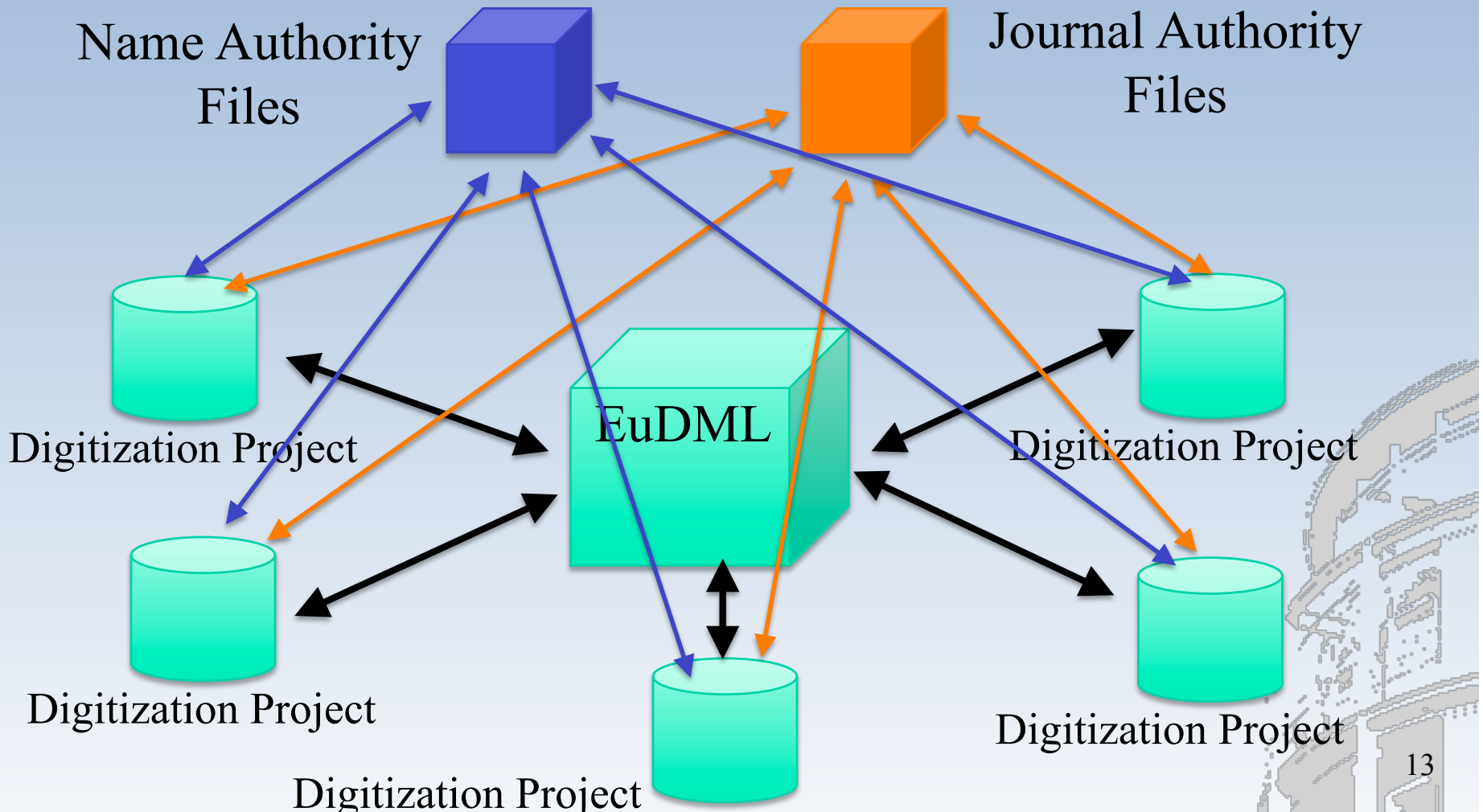
Communication Model?



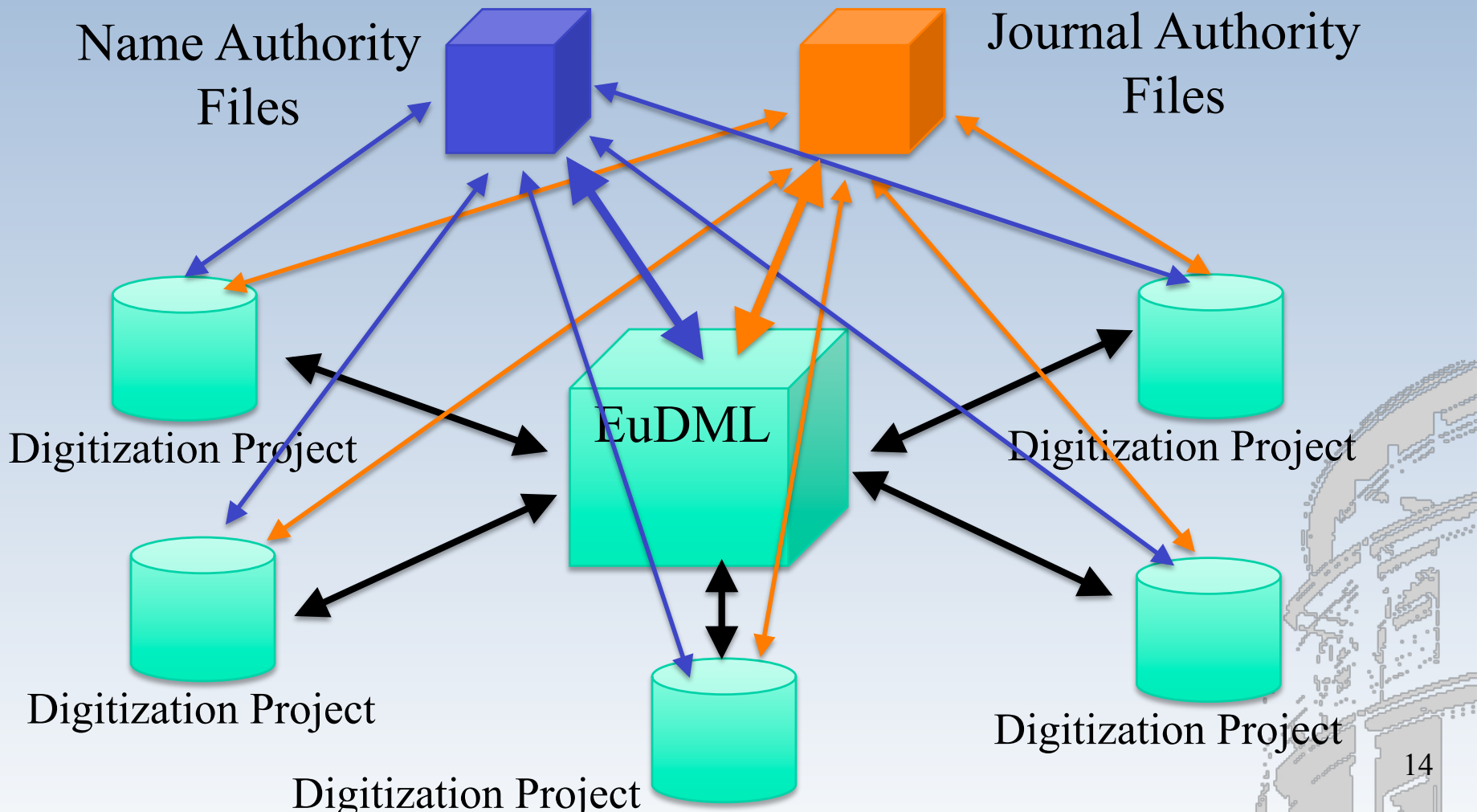
Communication Model?



Communication Model?



Communication Model?



3. Structures

Numdam:

internal identifier: AIF_1979__29_1_107_0

(journal, year, volume, number, page start, extra)

citation: Ann. Inst. Fourier 29, no.1, 107-124 (1979)

allows to reconstruct the journal from data:

Volume, Issue etc.

GDZ:

dc:source: Muth, P. | Ueber alternierende Formen. | Journal für die reine und angewandte Mathematik | 1900 | BAND: 122 | Seite: 89

Additional Datasets?

Basis: Journal information in Journal Authority Files

Is additional information necessary to reconstruct

- Volumes
- Issue

of the journal?

Or can the metadata set be enriched so that everything can be reconstructed from individual data?

Best option:

- metadata with full citation information (volume, issue, pages)
- additional datasets for table of contents

Additional Data?

For full reconstruction of journal issues and volumes, *all* parts are needed.

Is this necessary?

This would include front and back matter and all other parts that are not mathematical articles.

Standard procedure needed to handle this material:

- Type list
- Standardised headings?

Else:

dc:source: | | Mathematische Annalen | 1967 | BAND: 174 | Seite: 311

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Thank you!

Questions?

Contact:

Thomas Fischer

fischer@sub.uni-goettingen.de