

北海道大学数学教室における e-print サーバの構築

行木孝夫*

北海道大学大学院理学研究科数学専攻

概要

北海道大学数学教室における e-print サーバの構築について述べ、利用するソフトウェアとプロトコルを紹介する。メタデータを OAI-PMH プロトコルに載せて共有することで情報を有効に活用できることを実例によって示す。これは 21 世紀 COE プロジェクト「非線型における特異性の数学」にて「数学の海」として実施されるものである。

1 はじめに

近年のネットワーク環境の変化は著しく、研究に関する主な情報交換はメーリングリストと web への移行を完了しつつある。しかしながら、利便性を向上させたかに見える ML と web とを中心とした情報交換には新たな問題点が指摘されている。あらかじめ情報の所在を知らなければアクセスできないという問題である。

以前は数学教室の掲示板に貼られていた情報が電子的に流通する結果、その情報を研究グループを核とするコミュニティの外から得るにはサーチエンジンの提供する雑多な情報を濾過しなければならない。

この問題をある程度まで解決するには、何らかの形で情報の所在を示すメタ情報を共有する機構が必要である。実装手段は多く考えられるが、今のところはデファクトスタンダードになりつつあるプロトコル OAI-PMH[3] によって実装することが可能である。実際、arXiv をはじめとしてサポートしているサーバも多い。

北大数学教室では e-print サーバ間の連携に OAI-PMH を採用するサーバプログラムを利用することで、問題の解決を図りながら e-print サーバの構築を行った。arXiv 等とほぼ同等の機能を実現できるものである。以下では OAI-PMH を実装するソフトウェアの選定からメタデータの収集、サーバの構築までを簡単に解説する。

2 概要

OAI-PMH に対応したメタデータサーバをリポジトリ、データプロバイダとよぶ。分散したデータプロバイダから OAI-PMH によってメタデータを収集し、何らかの機能を提供するサーバをサービスプロバイダとよぶ。ユーザはサービスプロバイダの提供する機能を利用する (図 1)。

*nami@math.sci.hokuai.ac.jp

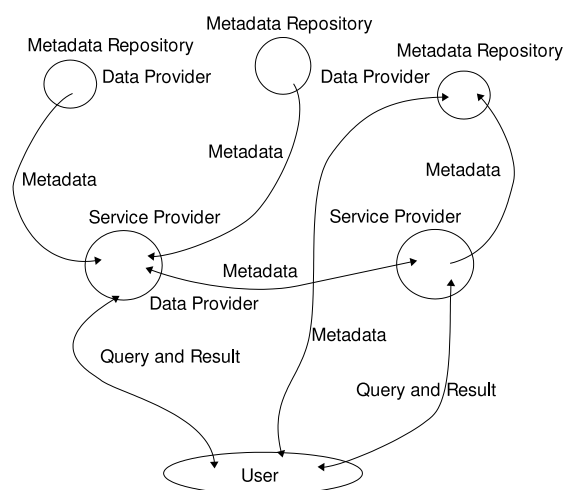


図 1: 概念図

3 データプロバイダの構築

3.1 ソフトウェアの選定

E-print サーバの実現に際しては、適切な形式でファイルを置き、目次としての HTML を書くだけのサーバであれば簡単である。しかし、投稿者管理、版管理、OAI-PMH の実装などを考えると相当の規模になることを考慮し、既存のオープンソースソフトウェアを利用することにした。候補は表 1 の二件であり、それぞれに特徴がある。

ソフトウェア	所在	開発元	特徴
GNU EPrints	www.eprints.org	Southampton University	Perl ベース。apache+mod_perl+MySQL
DSpace	www.dspace.org	MIT と HP の共同開発	Java ベース。JSP+PostgreSQL

表 1: ソフトウェアの候補

どちらも必要な機能を満たしている。機能比較は [7]などを参照のこと。両者を比較すれば DSpace は大規模なサーバに向いている。「大規模」とは、複数の電子ジャーナルを同時に構築するという場合、あるいは複数の学部にもたがるリポジトリを構築する場合である。

数学教室の規模を勘案し、Perl ベースの GNU EPrints を採用した。海外での GNU EPrints の事例については [6]などを参照。

3.2 コンテンツと分類

- 北大数学教室プレプリントシリーズ
- 北大数学教室講究録

- 研究集会の講演論文、予稿、論文集
- ソフトウェア
- ビデオアーカイブ (主に公開講座、一部の講義)
- Hokkaido Mathematical Journal(予定)

プレプリント、講究録は北大数学教室発行のものに限っている。教室メンバー、COE 協力研究者であれば任意に投稿できる。

研究集会の講演論文、予稿、論文集は主催者から申請があれば収録する。研究集会の担当者が講演者から集約し、まとめて登録する体制を整えた。本来は講演者に入力してもらうべきではあるが、一度きりの登録には入力が煩雑であろう。

ソフトウェア、ビデオアーカイブとは、数学関連ソフトウェア、講演記録を意味する。前者は無視できない存在であり、後者は近年の研究交流においては不可欠になるであろう。

別にセミナーや研究集会情報の設定も可能である。

3.3 盗作、改変問題

問題は、電子化され、タイトルと著者を改変された情報から正しいプレプリントをレフェリーが見つげ出すのは困難であるという点に尽きる。OAI-PMH によって教室間のプレプリント情報を共有し、少なくとも abstract までを利用できることになれば、状況は大きく前進する。なぜなら、盗作者が abstract を正しく変更するという自体は考えにくいからである。

4 サービスプロバイダの構築

メタデータを取得するデータプロバイダとして、arXiv, EUCLID, MathPreprints, JuNii を選択した。再利用を考慮し、ミラーサーバを設置する。

4.1 ミラーサーバの設置

ミラーサーバには [3] の関連ツールリストから oaia を選択した。これは ListRecords の出力を record ごとに 1 レコードとして RDB へ格納し、インターフェースとして CGI を用意している。以下にミラー状況を示す。

明らかに arXiv の量は頭抜けており、EUCLID がこれに続く。両者が数学関連の最大のデータプロバイダであることは間違いない。EUCLID は電子ジャーナルを供給し、arXiv はプレプリントであるから、重複も比較的少ないと考えられる。

4.2 検索

収集したメタデータによる検索サービスをフリーソフトウェアの NAMAZU により提供している。Google API を用いた結果と合わせて表示することで、効率的な検索結果となるよう配慮した。

Identifier	URI	Harvest Method	Records
arXiv	www.arxiv.org	ListRecords	323549
euclid	projecteuclid.org	ListRecords	29541
MathPreprints	www.mathpreprints.com	ListRecords	1223
HUMATH	eprints.math.sci.hokudai.ac.jp	ListRecords	890
HMJ	hmj.math.sci.hokudai.ac.jp	ListRecords	1100
NUMDAM	www.numdam.org	ListRecords	8931

表 2: ミラーリング中のデータプロバイダとメタデータ件数

4.3 Mathematics Subject Classification の共相関解析

ここでは収集したメタデータから各分野の相関関係を示すことにした。使用したメタデータは Project EUCLID の提供する oai_dc であり、そこから日本の数学教室発行のジャーナルを抽出した。後者には独自公開の Hokkaido Mathematical Journal も含めている。

相関を可視化し図 2 に示す。詳細な解析は後の機会に譲る。このような手法は科学技術政策研究所発行の [12, 13] において Thomson ISI 社のデータベース ESI を用いた共引用解析として実施されている。

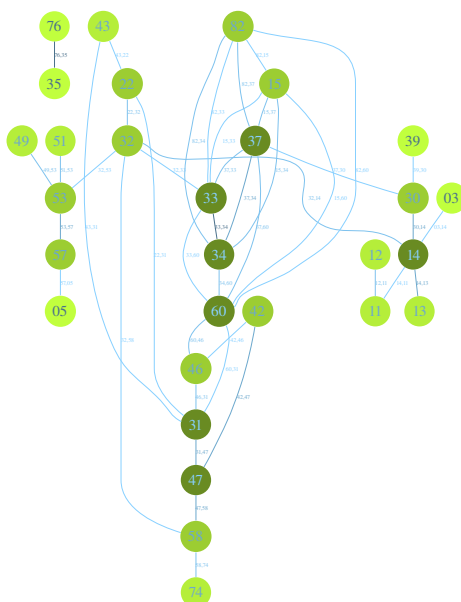


図 2: Project EUCLID から 2004 年の東北大、東工大、名古屋大の各ジャーナルについてメタデータを抽出

5 今後の展開

EPrints は e-print サーバとして開発が続いているソフトウェアであり、それに従う形での運用は既に述べたように比較的容易である。セミナー情報サーバ、メタデータサーバとしての運用は予定しているだけであるが、設定次第で可能であり、実現すれば総合的な研究情報サーバとしての見通しをつけられる。

OAI-PMH で提供するメタデータ形式を複数種類に対応させることで、MathNet [9] や JuNii [8] といったメタデータ関連プロジェクトとの相互運用を可能にできる。

最近のいわゆる blog の普及に伴い、メタデータは確実に普及し知名度を増している。その中で主役に近い役割を果たしているものは更新通知に用いられる RSS というメタデータの形式である。これには数種類の版があり議論の多い部分でもあるが、メタデータの主要部にはダブリンコアを採用している。

更なる展望として、更新通知を RSS で行いメタデータの完全な交換には OAI-PMH という方針が考えられる。

A メタデータ交換プロトコル OAI-PMH の概観

A.1 リクエストの概要

OAI-PMH は HTTP 上の GET メソッドへエンコードしたリクエストによって系統的にメタデータを取得するプロトコルである。CGI スクリプトによって実装されることが多い。最新版は 2.0 である。

リクエストは次の形式をとる。

```
http://server.domain/cgi-script?verb=request&request-options
```

主なリクエストを次に挙げる。

1. Identify: リポジトリの情報を取得。
2. ListSets: 選択しうる主題を取得。
3. ListMetadataFormats: メタデータ形式を取得。
4. ListRecords: 実際にメタデータを取得。オプション metadataPrefix は必須。

ListMetadataFormats によってメタデータ形式を取得した後、次のように ListRecords リクエストを発行すれば系統的にメタデータを取得できる。メタデータ形式 oai_dc は必須であり、これを使うならば ListMetadataFormats の発行は必ずしも必要ではない。

```
http://server.domain/cgi-script?verb=ListRecords&metadataPrefix=oai_dc
```

参考文献

- [1] <http://www.ams.org/msc/>
- [2] <http://dublincore.org/>
- [3] Open Archives Initiative, <http://www.openarchives.org/>
- [4] <http://www.eprints.org/>
- [5] <http://www.dspace.org/>
- [6] Stephen Pinfield, Mike Gardner and John MacColl, *Setting up an institutional e-print archive*, Ariadne. Issue 31, March-April, 2002,
(邦訳: <http://www.nii.ac.jp/metadata/oai-pmh/eprints/>)
- [7] Budapest Open Access Initiative, *A Guide to Institutional Repository Software*,
<http://www.soros.org/openaccess/software/>
(邦訳: http://www.nii.ac.jp/metadata/oai-pmh/osi_ir_software_guide.html)
- [8] <http://ju.nii.ac.jp>
- [9] <http://www.math-net.org>
- [10] <http://citebase.eprints.org>
- [11] <http://coe.math.sci.hokudai.ac.jp/literature>
- [12] NISTEP REPORT No.82 科学技術の中長期発展に係る俯瞰的予測調査 急速に発展しつつある研究領域調査 平成 15 年度調査報告書
- [13] NISTEP REPORT No.95 我が国における科学技術の状況と今後の発展の方向性
- [14] EUCLID 全体 <http://metron.math.sci.hokudai.ac.jp/search-new/navi>
日本の数学教室発行雑誌 <http://metron.math.sci.hokudai.ac.jp/search-new/navi/jp>