ワークショップ “Towards a Digital Mathematics Library 2009” への参加報告

行木孝夫
北海道大学大学院理学研究院数学部門
nami@math.sci.hokudai.ac.jp

1 概要

国際会議 CICM2009 のワークショップとして開催された “Towards a Digital Mathematics Library 2009” へ参加し，日本の DML(DML-JP) に関する口頭発表を行った．昨年 7 月に Birmingham で開催された DML2008 に続くもので，直接には 2 回目のワークショップである．参加者は会議全体で 50 名程度であった．以下，Digital Mathematics Library を DML と略記する．

プログラムは web ページ [1] を参照．Organizer はチェコの DML(DML-CZ) の中心にいる Petr Sojka であり，九大数理の鈴木昌和氏が委員の一人であった．

1.1 各国の DML

Digital Mathematics Library (DML) なる概念それ自体は 2002 年頃から (シリアルズクライシスの深刻化との関係もあって) 提案され，各地でドメスティックなプロジェクトが形成されてきた経緯がある．基本的には，大規模出版社のスコープに入らない数学系出版物を各地域のイニシアチブで電子化するものであった．初期は International Mathematical Union(IMU) の委員会である CEIC によって主導されてきたが，仏の numdam.org，独の www.emani.org 等の主要なグループがリードする形になりつつある．

各国の DML を講演資料 Thierry Bouche (NUMDAM),“Ongoing Work on a European Virtual Library of Mathematics General presentation”[2] より抜粋した．米独仏は別格として，欧州各国の DML は日本の大規模大学の数学系ジャーナル 1 タイトルに相当する規模である．EuDML は FP6 への提案が採択されたと聞いた．小規模な DML を統合して EuDML とするプランは，小規模なジャーナルとプラットフォームが多数存在する日本と状況は類似していると考えられる．

US J-STOR (260,000 items), Project Euclid (100,000), CMS (4,000)

Asia DML-JP (30,000 items), China ??

Europe EuDML (190,000 items)

Germany ERAM/JFM, GDZ, ELibM (85,000 items)

France Gallica-Math, NUMDAM, CEDRAM, TEL (50,000 items)

Poland ICM/BWM (13,000 items)

Portugal SPM/BNP (2,000 items)

Spain DML-E (5,000 items)

Czech Rep. DML-CZ (11,000 items)

Russia RusDML (13,000 items)

Bulgaria BulDML (2,500 items)

Serbia No formalised project (3,700 items)

Switzerland SwissDML (5,000 items)

1.2 日本の位置

筆者は数学会の情報化委員会のもとに DML-JP を試験的に運用してきた．昨年度および今年度については国立情報学研究所の SPARC Japan 事業から支援を受けている．2005 年以来 RIMS 研究集会として紀要の電子化に関わる研究集会を開催してきたが，その成果の一つでもある．

DML-JP には約 30,000 論文を捉えており，規模では米国を除き独仏に次ぐ位置にある．うち 15,000 件程度が Project Euclid であり，残り 15,000 件程度が機関リポジトリの貢献である．

数学分野において特定の研究分野に限定すると，全世界の論文数のうち 10-5% を日本発の刊行物 (紀要/ジャーナル) がカバーする．該当する研究分野における日本の研究機関所属の研究者が書く論文数の割合は 15-10%であることを考えると，[5] の 3 節「我が国の学術情報発信の今後の在り方について」に示された「国立情報学研究所の調査では，平成 12 年には，我が国の研究者は国際的に流通している学術論文の約 12 パーセントを生産しているが，そのうち約 80 パーセントは海外の学術雑誌に掲載されており」という状況を逆転しているものである．DML の構築を通じて得られた数値であり，数学における特徴として意識してよい事実であろう．

意外に思われた点として，日本語コンテンツの電子化を推進することが AMS の参加者からも強く推奨されたことを挙げられる．この場合の日本語コンテンツとは，Math. Reviews に登録されているコンテンツであるが，日本数学会発行の和文誌「数学」と京大数理研発行の「数理解析研究所講究録」である．前者は岩波書店にて検討され，後者は京大の機関リポジトリにおいて電子化されている．後者について Math. Reviews の該当レビューからフルテキストへリンクを提供することも DML-JP のスコープの一つである．この種のメタデータの利用について，AMS との連携を開始したい．

1.3 DML の背景

DML に関する共通認識を改めて確認する．これも，[2] より抜粋した．数学の研究資料に関わる一般的な事項だが，数学以外では全く適合しない．従って，機会を捉えて主張し続けるべき事項でもある．

査読された数学の論文は決して無用になることはない．新たな結果が出ることはあっても，その基盤になっている．他の分野に対しては，後年になってから応用されることが多い．従って，全てを参照できるネットワークが必要である．注意深くアーカイブ，インデックスされ，長期に渡って保存するものでなければならない．

即時的な情報交換は preprint として arXiv 等が担っている．学術誌への出版までには数年の遅れが生じる．参考文献の 50% は 10 年以上経過した論文，25% は 20 年以上経過した論文である．

紙媒体なら図書館間の相互貸借や複写依頼制度で実現しているともいえるのだが，電子的には未だに実現していない．

2 大規模デジタルリポジトリとの関わり

近年の機関リポジトリをはじめとするデジタルリポジトリの展開に伴い，DML への影響も少なくない．基調講演には J-STOR と Euclid, Math. Reviews からの招待講演が設定されていた．

2.1 J-STOR

John Burns: デジタルリポジトリとしての J-STOR は，アーカイブをコミュニティへ積極的に公開している．「公開」の意味は API の開発やアノテーションプロジェクトへの支援などである．J-STOR は約 40,000,000 ページを保持し，参考文献リストの提供などの高度なリポジトリを形成するが，それを有効に利用するのはコミュニティの責任である．

2.2 Project Euclid

David Ruddy: Project Euclid はシリアルズクライシスへ対抗し数学に関する学術出版を支援するものであった．Mellon 財団による支援から開始．1999-2002 頃の発展段階ではカレントを重視したが，現在ではバックファイルも保持する．トータルでは 97,000 論文, 70% がオープンアクセスである．

2.3 Mathematical Reviews

Patric Ion: Math. Reviews のエントリーは時間の 3 乗に比例して増加し，平均の共著者数も増加している．Mathematics Subject Classification 2010 [4] に際しては新機軸を積極的に導入した．Watt による新統計量も導入し，自動分類への取り組みも実施．引用指数 MCQ は 5 年単位のインパクトファクターである．Math. Reviews の提供する統計値は数学に適したものでなければ意味がない．MathML3 の策定へは積極的に関与した．文献に付随するファクトデータへの対応などは今後の課題である．

3　トピックス

一般講演で発表のあった話題について概観する．興味深い内容もある一方，すぐに利用することは難しい内容も多い．数式検索と MathML への変換技術など，相互に関連する内容も多いこともあり，広範囲な知識が要求される．

3.1　リファレンス等の同定とリンキング

出版段階での参考文献抽出．可能な限りのリンク (MR,DOI,···) 生成．問題点は，ミスタイプ．OCR 誤認識，不明確な巻号表記，翻訳タイトルなどにある．文字ベースの Levenshtein 距離および同種の距離はこれらを補完しうる．

3.2　数式検索

MathML をベースとする数式検索に関わる講演が二件あった．TEX だけでは web 時代に対応しきれないのだが，MathML の普及も一般へは進んでいないために利用シチュエーションが問題である．

3.3　TEX から MathML への変換

ツールとしては arxmlive,LaTeXML,NUMDAM-CEDRAM による Tralics などである．TEX から XML+MathML+metadata を生成したい．Tralics は自動的に PDF,XML を生成する点が長所．LaTeXML は厚くサポートしているが処理速度は遅い．TeX4ht はパイオニア的ツールである．TtM は速いが認識率が悪い．Hermes は評価不能．

TEX の機能と MathML の機能とは等価ではない．例えば，eqnarrary*は MathML に存在しない．¡table¿による実装と¡mtable¿による実装が選択肢である．

決定版は存在しないから，それぞれの長所をを統合するべきである．

3.4　タグ付与 PDF の生成

TEX からタグ付与 PDF を生成することで，生成した PDF から XML を生成できる．メタデータを別に持つ必要がない．

4　おわりに

Digital Mathematics Library を構築するとは，単なる web ページを作成することに止まらず，個別技術を統合する広い視野が必要となることを痛感したワークショップであった．未だに電子化の問題を抱える個々のジャーナルについては，電子化の済んだジャーナル等の経験を活かす場が必要だとも考えられる．同時に，膨大な日本のジャーナルについては，Math. Reviews でさえも書

誌に混乱が見られることもあり，AMS との密な連絡を伴う系統的な活動が必要と考えられる．そのためには，日本の DML 構築に責任を持つ何らかのイニシアチブが必須となるだろう．

参考文献

[1] Towards a digital mathematics library (7-9 Jul. 2009, Grand Bend, Canada) http://www.fi.muni.cz/ sojka/dml-2009-program.html [Accessed:2009/08/01]

[2] Thierry Bouche, “Ongoing Work on a European Virtual Library of Mathematics General presentation” `http://www.fi.muni.cz/~sojka/dml-2009-bouche-projects.pdf` [Accessed:2009/08/01]

[3] Takao Namiki, Hiraku Kuroda, Shunsuke Naruse, *Experimental DML over digital repositories in Japan*, http://arxiv.org/abs/0907.3826v1

[4] MSC2010, `http://msc2010.org/` [Accessed:2009/08/01]

[5] 学術情報基盤の今後の在り方について（報告）平成 18 年 3 月 23 日科学技術・学術審議会学術分科会　研究環境基盤部会・学術情報基盤作業部会