数理解析研究所講究録 1463 巻 2006 年 21-27

Funkcialaj Ekvacioj の遡及電子化中間報告

高山 信毅　　鈴木 昌和
NOBUKI TAKAYAMA　　MASAKAZU SUZUKI
神戸大学　　九州大学
KOBE UNIVERSITY　　KYUSHU UNIVERSITY

Abstract

数学論文誌の高度な電子化を目指して、現在神戸大学の高山研究室と九州大学の鈴木研究室で協力して進めている Funkcialaj Ekvacioj の遡及電子化の計画と経過についてのべる。

1 Funkcialaj Ekvacioj の電子化の歴史

Funkcialj Ekvacioj（略して FE という）は日本数学会函数方程式論分科会が編集し，神戸大学が発行している函数方程式に関する国際専門誌である．1958 年に福原満洲雄（東大理学部教授），南雲道夫（阪大理学部教授），佐藤徳意（神戸大理学部教授）により創刊された．毎年 1 巻 (3 冊分) が発行され，2005 年現在で Vol. 48 に達しており国際的にみても比較的歴史のある雑誌である．

FE の電子化プロジェクトは 2001 年にはじまった．この初期の時期には TeX 原稿から latextohtml, dvips, dvipdfm などを用いて pdf, ps, HTML 形式の論文を作成し，実験的に公開した (44/vol3). 2004 年からは J-stage での運用も開始し PDF や J-stage 用のメタデータ (引用文献の情報など) はレタープレスが作成するようになった．この年の夏に全バックナンバーの scan をおこなった．FE を裁断機できれいに裁断してから，Xerox の Fax/Scanner/Printer 複合機 DocuPrint 230 で片面ずづ tif 形式で scan し，同じく Xerox の文書ソフトウエア DocuWorks を用いて傾き補正，および片面ずつ scan したデータの再並び替えをおこなった (足立，野呂 がいろいろ工夫した)．また DocuWorks を用いて tif 形式から PDF 形式のファイルも作成した．

さてこの時点から FE は J-stage および 神戸大学数学のサイトの 2 個所で運用することとなった．どうして一カ所だけにしなかったかの理由は以下のとおりである．

1. 将来的に予算面での問題が発生したときのための保険．
2. J-stage のユーザ認証システムはこの時点で不十分．
3. 電子出版はまだまだ流動的な仕組みである．ビジネスモデルの考察や技術の展開を考えていかないといけない．自分達でサイトを運用するのは，自分達の勉強になるであろう．

さて，2005 年は過去分の XML 書誌情報の作成と，検索用のテキスト情報を埋め込んだ PDF ファイルの作成を九大・鈴木を中心とする数学文書情報処理に関する研究グループ InftyProject[1] と協力して進めている．以下はその記録である．

2 計画の概要

現在，欧米で急速に進められている多くの数学論文誌の電子化プロジェクトで行われているのは，

1. 現存する主要な数学論文誌を (多くの場合，第 1 巻から現在まですべて) スキャニングして PDF 化し，WEB で閲覧できるようにする．
2. 各論文の書誌情報と文献表の書誌情報を抽出し，MathSciNet などのデータベースとリンクを結ぶ．
3. 論文のテキスト部の OCR による認識結果を Hidden Text の形で埋め込み，検索可能にする．

[1] http://www.inftyproject.org

という3点に要約される．この作業を高精度で効率的に方法が求められているわけであるが，FE 電子化においては高山の提案により上記 InftyProject と協力して電子化を行うこととした．単に，既存の遡及電子化と同等のことを実現する雑務として捉えるのではなく，現在もっている技術と数学者が獲得できる主要な外部資金源である科学研究費程度の予算規模を踏まえた上で，電子化された数学論文コンテンツの将来の利用拡大に備えて，現時点で何がで得きるかを考え，電子化技術の研究を行いながら FE の遡及電子化を実現する考えである．具体的には，上述の3点に加えて，差しあたり 2005 年度は

4. 各論文の節のタイトルや定義・定理・命題などを抽出した目次生成．
5. 将来の数学公式集の研究材料としての利用を考えた Displaystyle の数式の抽出．
6. Hidden Text には数式部分も LaTeX で認識結果を格納．

を行う．これらの情報の自動抽出と，修正インターフェース，これらの情報を付加した PDF 生成，XML 化を行いながら，数学論文誌電子化技術の蓄積をする．

現在，作業中の FE のデータは第1巻 (1958)〜第４５巻 (2002), 917 論文で，総ページ数は 15043、内訳は

第 1巻〜第 10 巻 : 1699 ページ
第 11 巻〜第 20 巻 : 2588 ページ
第 21 巻〜第 30 巻 : 3326 ページ
第 31 巻〜第 40 巻 : 5498 ページ
第 41 巻〜第 45 巻: 1938 ページ (第４３巻なし)

で, 言語別論文数は

英語：８７６，　仏語：３２，　独語：５，　露語１，　エスペラント語：１

となっている．2005 年度作業分は 2006 年 4 月に Web 公開する．

3　要素技術

数学論文誌の電子化作業は，次の４ステップに大別できる．

1. スキャニング（画像データとしての電子化）
2. 文字や数式の認識（OCR)
3. 情報抽出（書誌情報，キーワード，文書論理構造など）
4. 抽出した情報の構造化 (XML 化)

この節では，これらの点について注意するべきことや技術の現状などについて述べる．

3.1　スキャニング

最終的に高精度な電子化データを得るために，認識技術が重要であることは論を待たないが，実際の作業に最も大きな影響を与えるのがスキャニングであると言ってよい．如何に高度な技術を用いても，最初に得られるスキャン画像が低品質で有れば，その認識結果に多くは期待できず，人手による多大な修正コストがかかってしまう．

出来る限り状況良く保存されたきれいな原稿を用意することが先ず重要で，汚れを落とし，書き込みなどが有れば事前に消しておくことが大切である．製本されている論文誌は綴じ代を裁断して，高精度の自動給紙装置を用いてスキャニングする．大量のデータのスキャニングに安価なスキャナを用いると，給紙トラブルによるしわや極端な傾きの原因となる．ヘッドの汚れにも注意を払う必要がある．大量のデータのスキャニングには最新のコピー機を利用するのが良いであろう．

数学の論文誌の電子化では多くの場合，白黒2値のスキャニングで十分であろう．カラー画像はスキャニングにも，後の認識処理にも時間がかかり，コスト的に見合わない．必要な場合は特別なページのみ別処理するのが賢明であろう．また，電子データをディスプレイで閲覧することだけを目的とし，印刷ページが通常の文字しか含まないような場合は 300DPI 程度で画像データを作成する場合もあるが，閲覧者が論文を印刷して読むことも想定に入れた場合や，数学の論文誌のように小さな文字が出現する数式を含む場合は 600DPI が適当であり，欧米の数学論文誌電子化プロジェクトでは多くの場合，この解像度を用いている．

3.2 OCR

海外で高精度のOCRとして大きなシェアを占めているのは ScanSoft 社製の OmniPage と Abby 社の FinReader というソフトであり，共に一般のテキスト文書については非常に高い認識精度をもっている．しかし，認識精度を上げるために通常は言語情報が積極的に利用されていて，その為か数式の前後や複数言語が混在した引用文献表の部分などでは誤認識が増加する傾向がある．勿論，数式の部分は意味不明な記号列になったり，無理に誤ったテキスト文字列に変換されていたりする．

他方，InftyProject で開発している OCR ソフト"InftyReader" は数式の部分を認識して出力する実用ソフトとしては，現時点では世界的に見ても唯一であると考えられるが，テキスト部分の認識率についてはまだまだである．比較的ノイズの少ないページ画像に対しては，欧文テキスト部の認識率は国産の市販ソフトとほぼ同水準の認識率に達しているが，上述のような欧米の OCR に比べると少々見劣りがすると言わざるを得ない．

そこで，InftyProject では InftyReader の認識と任意の市販の OCR の認識結果を統合して出力する手法の開発を行っていて，FE の電子化ではその手法を試みている．実験では欧米の多くの数学論文誌電子化プロジェクトで使われている FineReader で認識し，その出力結果を InftyReader の認識結果と DP マッチングを利用しながら照合して，テキスト領域では原則として FineReader の認識結果で置き換え，その結果を原画像と照合して明白な矛盾が有る場合は InftyReader の認識結果を採用する方法を取っている．OCR の精度を向上させるために，複数の OCR の結果を組み合わせて Voting などによって認識結果を決める方法が非常に効果があることが良く知られている．上述の InftyReader と FineReader の認識結果を組み合わせる方法は，更に複数の市販 OCR の認識結果を参照する形に拡張することも容易で，テキスト部分については将来的に非常に高精度な電子化ができると期待している．しかしながら，数式部の認識については，残念ながら競合するソフトウェアがなく，同じような効果を出そうとすると，自前で異なる複数の手法や学習データを用いて記号認識や数式構造認識を実装する必要があり，苦しいところである．

3.3 情報抽出

現在，OCR の結果から抽出している情報は

1. 各論文の書誌情報（タイトル，著者，ページなど）
2. 引用文献表の各文献の書誌情報（著者，タイトル，雑誌名，巻号，年号，ページなど，詳しくは第 4 節を参照）
3. サブタイトル（section, subsection のタイトル）
4. 定義，定理，補題，命題などの記述部
5. Displaystyle の数式

である．2005 年 11 月の段階で，全論文に対して 1. ～ 3. は自動抽出後，目視による一通りの修正作業までが終了している．4. と 5. については，自動抽出は出来ていて，エリアの修正までは年度内に終了することが出来るが，内部の文字や数式の修正までは今年度の予算では出来ない見通しである．

これら作業の中で，最も困難であったのは文献表の各文献の書誌情報抽出である．後述の高山による MR の自動リンク作成を早く始めたいという事情があって，認識結果のテキストファイルを用いて直接，認識結果の文字・数式[2)]の修正と各書誌情報アイテム（著者名，タイトル，雑誌名，巻号，年号，ページなど）への分解を同時に行ったため，非効率的は作業になってしまった．その作業終了後，改めて Infty のユーザーインターフェース (図 1 参照) を用いて文献表の部分の文字認識修正を行ったが，こちらは比較的短時間に終了することができた．各書誌情報アイテムへの分解も通常のジャーナルの論文を引用している場合は自動抽出も精度が高く，修正も容易であった．最も困難であったのは，単行本や博士論文，大学や研究所のセミナー報告などが大量に引用文献表に含まれていて，その場合の書誌情報をどのように扱うかという判断に時間がかかってしまい，必ずしも統一した扱い方が出来たわけではない．今後の検討課題である．作業に数名のアルバイト（学生など）を雇用して行ったのも問題であった．文字や数式の認識結果の修正作業は，複数のアルバイトで行っても，それほど問題はないが，雑誌名や出版社名，数学者の引用文献表記の慣例などを考慮した「判断」を必要とする書誌情報アイテムの取得は，しっかりした能力をもつ一人の人にやってもらうのが（効率的にも結果の精度の面でも）よいというのが，今回得られた教訓である．

[2)] FE は論文誌の性格上，数式を含むタイトルの論文が多い．

3.4 構造化と PDF 生成

得られたデータを構造化するためには，その柔軟性，データ変換処理のしやすさなどから，現在では XML を利用するのが常識的であろう．その際，基礎的な電子データとしては出来る限り細かい情報も残した形で構造化しておくのが望ましい．現時点での電子ジャーナルライブラリに含める予定がないからといって細かい情報を捨ててしまうと，将来，ライブラリの機能拡張や情報追加などを計画したときに，電子化作業そのものをはじめからやりなおすことになってしまう可能性がある．そのような基礎データとしての XML を構築した上で，現時点でライブラリ構築に必要な XML を基礎データ XML から生成する方法をとることが良いであろう．InftyReader が出力する XML（以下，KML と記す）は，各文字の立体・斜体の区別や太字か否かの情報，画像中での位置座標をはじめ，ページ画像のブロック分割情報などを出来るだけ詳しく記録している．図・表・本文の区別タグや，第 3.3 節で述べた各情報を記録している．KML の仕様については，中川 [1], [?] に解説がある．

PDF の生成は KML から Hidden Text (数式も含む）を Picture 環境を用いて白で埋め込んだ LaTeX のソースファイルを生成し，LaTeX ツール（dvipdfmx など）を用いて行っている．

3.5 ユーザーインターフェース

作業を効率的に行うためには，ユーザーインターフェースが非常に重要である．他方，ユーザーインターフェースは開発コストがかかるのが頭の痛い所である．

また，電子ライブラリ用の PDF を生成する際に，Hidden Text は画像中の対応する場所に埋め込む必要があり，各文字や記号が画像中の位置座表を保持していることが重要である．通常のエディタで認識結果を編集すると削除や挿入をしたときに原画像上の座標が消えてしまうため，Infty のシステムでは特別のユーザーインタフェースを開発して使っている．

参考までに，InftySystem の認識結果修正インターフェースのスナップショットを載せておく．編集をしても原画像上の座標を失わないように工夫されている．

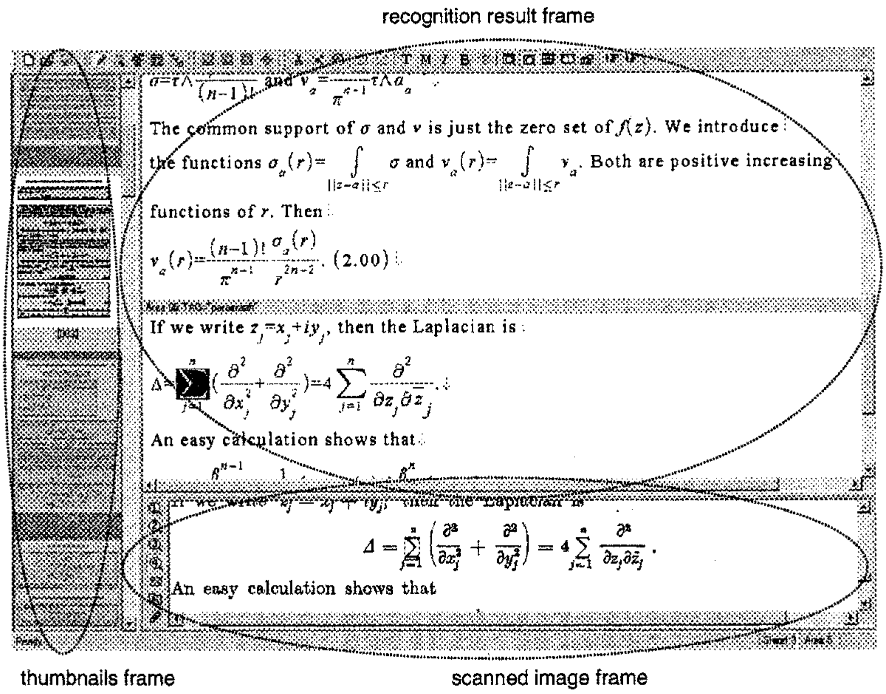

図 1: Infty System のユーザーインターフェース

4 MR の自動リンク作成と XML 編集システム

InftyReader が作成する書誌情報は下記のような形式である.

```
FILE            :fe30-305-332
YEAR            :1987
TITLE           :Studies on the Painlev\'e Equations IV.
                  Third Painlev\'e Equation $P_{III}$
AUTHOR          :OKAMOTO, Kazuo

bibitem         :1
type            :book
author          :Bourbaki, N.
booktitle       :Groupes et Alg\'ebres de Lie, Chapitre 4, 5 et 6
publisher       :Masson, Paris
year            :1981
```

このように "キーワード：値 改行記号" の形式で scan された書誌情報が格納されている. われわれのシステムはまず, AMS の MathSci net に問い合わせて, 参考文献の MR 番号を検索する. このとき多くのデータベースサービスは機械検索をみとめていないと理解するので, チューリングテストの立場をとって検索プログラムを書いた. チューリングテストとは雑にいえば, 対話の相手がプログラムであっても, 対話していて人間とおなじと感じれば知能としてみとめようではないか, というテストである. われわれの検索プログラムは人間があたかもタイプ入力しているように動作する. また時々休憩もする. したがって, 全参考文献の MathSci での検索には一日以上を要するが, MathSci 側からみれば人間とかわりない. また検索した情報は cash しているので, 再度検索プログラムを全情報に対して動作させるときはより高速であるし, MathSci net には余計な負担をかけない. 以下が, MR 番号検索プログラムで処理したあとの書誌情報である. たとえば参考文献 1 に MR 番号が付加されている.

```
FILE            :fe30-305-332
YEAR            :1987
TITLE           :Studies on the Painlev\'e Equations IV.
                 Third Painlev\'e Equation $P_{III}$
AUTHOR          :OKAMOTO, Kazuo
AUTHOR_utf8     :岡本和夫

bibitem         :1
type            :book
author          :Bourbaki, N.
booktitle       :Groupes et Alg\'ebres de Lie, Chapitre 4, 5 et 6
publisher       :Masson, Paris
year            :1981
mr              :MR682756
score           :75
query_string    :Cache/Bo/Bourbaki,Groupes,*,1981,1982
page            :138
```

さてこのように作成した情報は最終的にはやはり人手による修正が必要である. 修正の為に図 2 のように web で修正作業をおこなうための, システムを開発した. このシステムは OpenXM.org プロジェクト (数学ソフトウエアシステムの統合化プロジェクト) の成果 (kan/sm1 の CGI 機能など) を活用して作成されている. OpenXM では OpenXM over HTTP (OX-RFC 104) なる通信プロトコルを開発しており, その一部分として, cgi インタフェースを簡便に書く機能も提供している. これを用いて非常に短い期間で編集機能を実装することができた.

図 2 の上が infty 形式の書誌情報であり下が足立作のスタイルファイルで XML を処理した FE の書誌情報画面である. 修正作業をする人は上の infty 形式の書誌情報を変更し, preview ボタンをおして仕上がりをみる. 最終的にできあがったら, commit ボタンをおして修正情報を書誌情報の CVS サーバに反映させる.

Infty 形式から生成している XML 情報も掲載しておく. XML は複雑な木構造を記述できる強力な言語であるが, 書誌情報は複雑な木構造を必要としないので, infty 形式の書誌情報が読みやすいしメンテナンスが楽である. したがって XML は自動生成とし, infty 形式の書誌情報を基本としている.

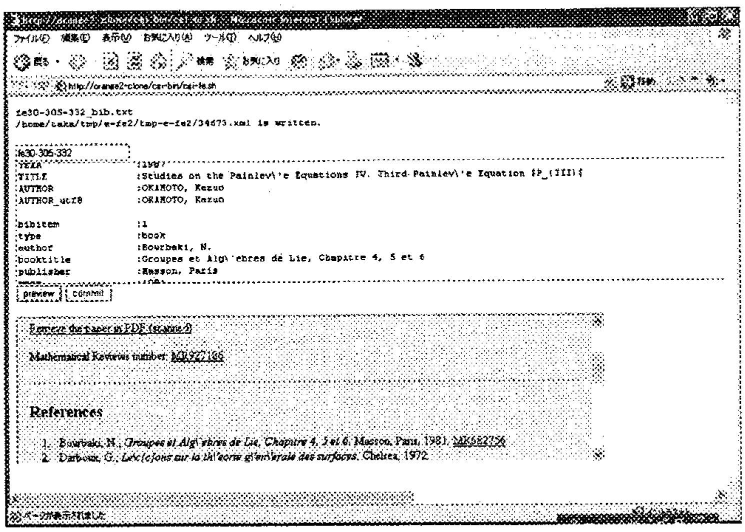

図 2: 書誌情報編集システム

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="f1.xsl"?>
<top_article>
 <mrnumber>MR927186</mrnumber>
 <author>Okamoto, Kazuo</author>
 <author_utf8>岡本 和夫</author_utf8>
 <title>Studies on the Painlev\'e equations. {IV}.
  Third Painlev\'e equation {$P\sb {{\rm III}}$}</title>
 <journal>Funkcialaj Ekvacioj. Serio Internacia</journal>
 <volume>30</volume>
 <year>1987</year>
 <page>305--332</page>
 <url_pdf>
  http://fe.math.kobe-u.ac.jp/FE/FullPapers/vol30/fe30-2-7.pdf</url_pdf>
 <mathsci_link>
  http://www.ams.org/mathscinet-getitem?mr=MR927186 </mathsci_link>
3584<fesi_info>
  <FILE>fe30-305-332</FILE>
  <YEAR>1987</YEAR>
  <TITLE>Studies on the Painlev\'e Equations IV.
    Third Painlev\'e Equation $P_{III}$</TITLE>
  <AUTHOR>OKAMOTO, Kazuo</AUTHOR>
  <AUTHOR_utf8>岡本 和夫</AUTHOR_utf8>
</fesi_info>

<references>
  <book>
    <bibitem>1</bibitem>
    <author>Bourbaki, N.</author>
    <booktitle>Groupes et Alg\'ebres de Lie, Chapitre 4, 5 et 6</booktitle>
    <publisher>Masson, Paris</publisher>
    <year>1981</year>
    <mr>MR682756</mr>
    <score>75</score>
    <query_string>Cache/Bo/Bourbaki,Groupes,*,1981,1982</query_string>
    <page>138</page>
  </book>

  <book>
 以下略
```

このインタフェースは現在広く利用されるようになった PukiWiki 等のユーザインタフェース設計と同じような精神にもとづいているといってよいであろう．つまり PukiWiki では HTML を直接書かない．その代りに独自の簡略されたマークアップ記号を用いて入力し，preview 機能で仕上がりをたしかめる．我々のシステムでも XML は直接書かない．その代りに我々の用途に適した簡略された入力形式であるところの infty 書誌情報形式を入力方法として利用し，Preview 機能で仕上がりをたしかめる．XML を入力するために GUI を用いる方法はいろいろ研究されている．GUI でなく簡略化された入力形式を用いる方法の利点は，入力がある文法をみたすテキストファイルなので，GUI ツール以外も利用できる．この点ではより柔軟性が高く，また論理的思考が得意な数学関係者向けでもあるだろう．また作業の変更履歴が cvs で管理されているので，気楽に変更ができる．実際の作業をやってみてこの "気楽さ" というのが，ユーザインタフェースの大事な部分であることが実感できた．

なお OAI 対応の XML 情報変換プログラムはまだ書いていないが，近日中に書くつもりである．

5 課題

既に述べたように，引用文献表の各文献の書誌情報抽出は最終的には人による判断が必要になる．第 4 節で述べた MathSci Net 検索と，原画像を参照しながら書誌情報アイテムを抽出するプログラムを合体させたユーザーインターフェースがあれば作業効率上効果的であろう．今後，そのようなインターフェースの制作を行っていきたいと考えている．

現在の InftyReader はかなり複雑な構造の行列も認識できるアルゴリズムが組み込まれているが，行列構造を容易に編集するユーザーインターフェースは未だ開発途上である．

ジャーナルを電子化する最大の利点はインターネットを介して簡単に参照できる環境が構築できる点にあることは勿論であるが，その他にも電子化することによって，紙媒体にはない情報や機能を付加することで新しい利用環境の構築が将来的には可能になるであろう．2005 年 7 月の研究集会「紀要電子化とその周辺」では，高山-鈴木-中川の共同発表の中で，中川がそのような可能性の一つの方向性を"Mathematical Knowledge Browser" の形で提案した．その概要は中川 [2] で紹介されている．

6 Reference

参 考 文 献

[1] K.Nakagawa, A.Nomura, M.Suzuki, *Extraction of Logical Structure from Articles in Mathematics,* Mathematical Knowledge Management, 3rd International Conference MKM2004, Bialowieja, Poland, Lecture Notes in Computer Sciences 3119, Springer (2004) pp.276–289

[2] K. Nakagawa and M. Suzuki. Mathematical Knowledge Browser with Automatic Hyperlink Detection. In *Mathematical Knowledge Management, Fourth International Conference, MKM 2005, Bremen, Germany, July 15-17,* To appear in LNCS. Springer, 2005.