

数理解析研究所講究録電子化の取り組み

岸本 典文

京都大学 数理解析研究所

kishimoto.norifumi.7c@kyoto-u.ac.jp

Digitalization of RIMS Kôkyûroku

Norifumi Kishimoto

Research Institute for Mathematical Sciences, Kyoto Univ.

概要

京都大学数理解析研究所にて発刊している数理解析研究所講究録を 2006 年度より電子化し、公開している。冊子体のスキャンに始まり、OCR 処理、電子化のスピードアップ、入稿原稿および入稿方法の電子化と改善を重ねてきた。本稿ではこれまでの電子化の取り組みを報告する。

1 はじめに

数理解析研究所（以下、「当研究所」という）では、共同研究の記録として、数理解析研究所講究録（以下、「講究録」という）を刊行している。1964 年の開始以来、毎年数十巻を刊行し、2016 年には 2000 巻が刊行された。

2006 年度からは、『講究録』の電子化・オンライン化についての方針 [1] に基づき、講究録を電子化し、京都大学学術情報リポジトリおよび、当研究所ホームページにて公開している。単に冊子をスキャンするだけでなく、OCR 処理や電子化のタイミングなど、より研究に資する形式を検討し、改善を重ねてきた。本稿ではその内容を報告する。

2 刊行済み講究録の電子化（2006 年～）

講究録の電子化以前、その内容を確認したい場合は、刊行された冊子を直接閲覧するしか方法がなかった。ただし、各巻の目次だけは当研究所ホームページに公開していた。

2006 年度より、2005 年度までに発刊された講究録を順に遡って電子化し、京都大学学術情報リポジトリへ登録を行った。また当研究所 Web サーバにも電子化したデータを置き、ホームページで公開していた目次のデータにリンクを貼る形で、公開を始めた。

2.1 メタデータの作成

電子化した論文をリポジトリに登録する際、論文自体の情報（著者、タイトル、言語、など）をメタデータ（図 1）として登録する必要がある。冊子から必要な情報を拾い、手入力で完成させた。

```
@NCID AN00061013
@VOLN 1963
@YEAR 20151000
@ACCN
@ATTR P
@TITL 教育数学の一側面
@TITY キョウイク スウガク ノ イチソクメン
@TITE Aspects of educational mathematics
@AUTH 高橋, 博
@AUTY タカハシ, ヒロシ
@AUTE Takahashi, Hiroshi
@AFFN 京都大学理学部
.
```

図 1 メタデータの例

2.2 論文のスキャンと OCR 処理

原稿は、冊子体を裁断して PDF 形式でスキャンした。解像度は 600dpi とし、スキャン後のデータに OCR 処理を行った。OCR 処理は、数式を認識し、TeX の構文に変換する株式会社デジタルノーツの OCR ソフト [2] を使用した（図 2）。これにより、数式の検索が行いやすくなった。

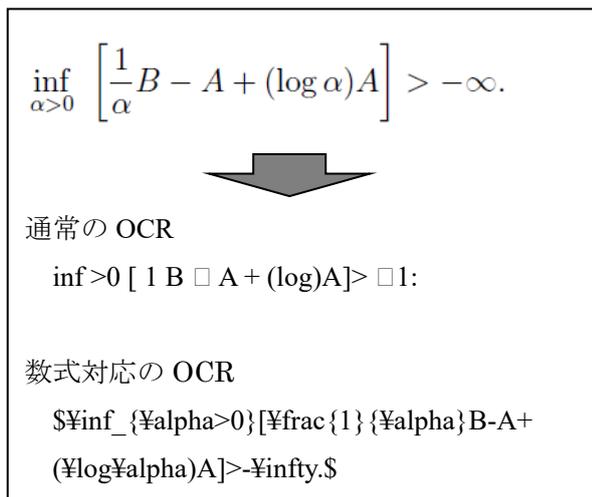


図2 数式対応の OCR

2.3 ホームページ更新処理の半自動化

メタデータは構造化された扱いやすいデータであるため、ホームページで電子化した講究録を公開する際にも利用している。メタデータの内容を読み取って、HTML ファイルを生成し、公開用の PDF データとともに Web サーバへアップするという一連の作業をツールで行うようにした (図3)。

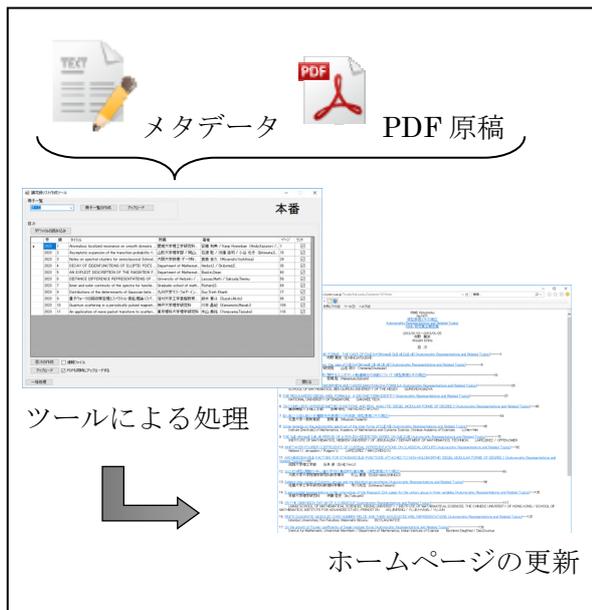


図3 ツールによるホームページ更新処理

2.4 課題

第1巻まで遡って電子化し、公開することができた。しかし、2006年度から最新巻は、コストを下げるため100巻ごとにまとめて電子化、公開していた。そのため、冊子の発刊から電子化まで最大で2年程度待たなければならない場合もでてきた。

3 速報ファイルの公開 (2012年～)

最新巻発刊から電子化まで時間がかかるとい

う不便を緩和するため、講究録の印刷前のデータを当研究所ホームページにて速報ファイルとして公開することとなった[3]。

3.1 印刷前データの収集

講究録の原稿はすべて紙で入稿していたが、速報ファイルとして公開するにあたり、著者から電子データの提供も依頼することにした。

3.2 課題

PDF データの場合、フォントの埋め込みがなければ、閲覧環境によって見た目が変わる場合があり、著者の期待通りの仕上がりとならない可能性がある[4]。そのため、講究録の印刷業者への入稿は以前と変わらず紙で行っていた。結果、速報ファイルとして直ちに論文の公開ができるようになったが、その分作業量が増えてしまった。

また、速報ファイルはあくまでも発刊前の稿であり、冊子体をスキャンした正式版の公開は従来通り100号ごとのままだった。

4 電子入稿の開始 (2018年～)

講究録の著者は、速報ファイル用に電子データを、冊子用に紙原稿をそれぞれ入稿しており二度手間と言える状態だった。また、冊子体の刊行から、冊子をスキャンした電子化データの公開には時間差があった。これらを解消するため今年度より、著者からは PDF 形式のデータのみを入稿してもらい、そのデータをもとに、当研究所ホームページおよび、京都大学学術情報リポジトリにて公開するよう改めた。これにより、冊子体のスキャン作業も今後なくなる予定である。

4.1 アウトライン化データで原稿確認

従来、講究録の原稿はすべて紙で入稿していたため、原稿と発刊された冊子はそのまま同じ見た目だった。ところが PDF データの場合、フォントの埋め込みがなく、閲覧環境に指定のフォントがない場合、フォントの置き換えが起これ、見た目が変わってしまう。そこで、入稿された PDF データを当研究所にてアウトライン化し、それを印刷後のイメージとして、著者に確認してもらうことにした (図4)。

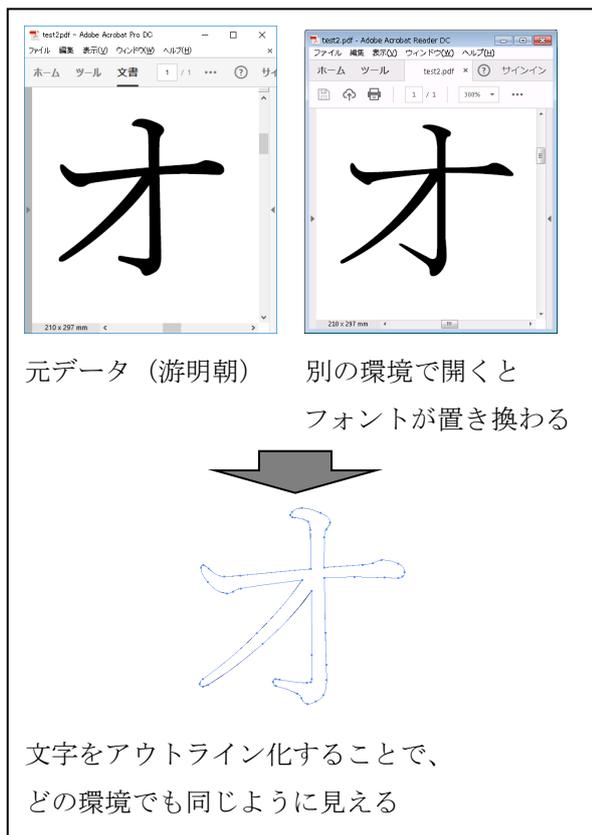


図 4 環境による見え方の違いとアウトライン化

4.2 入稿、確認用の Web フォームとクラウドストレージ

原稿が紙からデータに置き換わるため、入稿方法も郵送から Web フォームに変更した (図 5)。フォームからアップされたデータはクラウドストレージから確認することもできる (図 6)。

図 5 原稿入稿フォーム

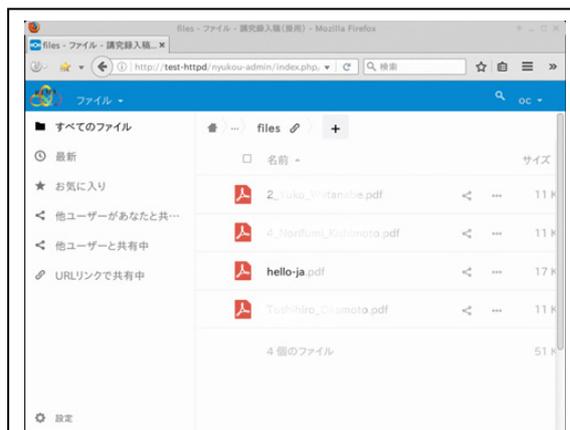


図 6 クラウドストレージから原稿を確認

4.3 課題と今後の展開

電子入稿は始まっており、今後 2019 年の初め頃からは電子入稿された原稿が講究録として発刊される予定である。現状は特に大きな問題はないが、まだ経過の観察中である。

今後はメタデータの作成をより効率的に行う支援ツールの作成を検討している。

5 まとめ

今回、2006 年度から始まった講究録電子化の取り組みを振り返った。過去の論文の電子化に始まり、新刊の電子化、そして入稿プロセスをも電子化することができた。ICT の進歩に合わせ、より研究に資する形を模索し、改善を重ねてきた研究所スタッフの努力の歴史と言えるだろう。今後とも利用者からのフィードバックを得ながら、付加価値の向上と効率化を図っていきたい。

謝辞

日頃、電子化業務に携わっている数理解析研究所共同利用掛、図書掛の皆様には感謝の意を表す。

参考文献

- [1] 『講究録』の電子化・オンライン化についての方針
<http://www.kurims.kyoto-u.ac.jp/~kenkyubu/news/kokyuroku-densika.html>
- [2] 株式会社デジタルノーツ
 ドキュメント電子化サービス(OCR 事業)
<http://www.digitalnauts.co.jp/solution/ocr.html>
- [3] 数理解析研究所講究録 速報ファイル
<http://www.kurims.kyoto-u.ac.jp/ja/publi-kokyuroku.html>
- [4] 奥村晴彦,黒木裕介, [改訂第 7 版]LaTeX_ε 美文書作成入門, PP.317-318, 技術評論社, 2017 年.