John Burns; Nigel Kerr

Community Curation and Management of Mathematical Literature

In: Petr Sojka (ed.): Towards a Digital Mathematics Library. Grand Bend, Ontario, Canada, July 8-9th, 2009. Masaryk University Press, Brno, 2009. pp. 17--24.

Persistent URL: http://dml.cz/dmlcz/702552

Community Curation and Management of Mathematical Literature

John Burns and Nigel Kerr

Ithaka/JSTOR, USA

Abstract. JSTOR is one of the primary providers of scholarly mathematics texts, providing access to journals in mathematics and the sciences dating back to the mid 1600's. There is now a critical mass of literature online and the task going forward is as much to provide tools to make it more accurate, more discoverable and more usable as it is to add more material. Often the tool building can be done by collaboration between information retrieval experts and practitioners in the field, irrespective of the subject area. Mathematics, however, provides special problems, since the literature is inaccessible to those outside of the field and so mathematicians are the only ones capable of providing the nuanced understanding of notation and meaning necessary to establish equivalence and relevance; the archivists are simply unable to provide the level of expertise and curation necessary. Mathematics therefore provides a unique opportunity to build tools for the discovery and use of the literature via community contributed curation and management of the material. We argue that this community is exceptional and it needs to define and build unique infrastructures, infrastructures that co-exist alongside existing repositories and allow mathematicians to structure their resources and discourse independently of the holder of the material. We will discuss various programs and projects in JSTOR labs of relevance to the mathematics community, including the Open Annotations and the Decapod projects and we cover the ways in which JSTOR could work with the community to meet their needs.

1 Introduction

Online archives are now vast and are starting to approach the ideal of everything being accessible online. JSTOR's mathematics and statistics collections include around 260,000 articles from 1665 to the present, Project Euclid[1] has over 96,000 articles, NUMDAM has about 27,000 and Göttinger Digitalisierungszentrums has almost 4,000 volumes. Arguably, based on the sheer content count, critical mass is being achieved. Moreover, JSTOR's commitment to longer-term access and archiving makes this collection a stable and reliable resource over time; the same presumably applies to the other major repositories.

Digitization processes that work for more generic content are inadequate for mathematical literature. The digitization process includes image capture, OCR and document structure analysis, and the standard process does not deal

[1] Operated by Cornell University Library and Duke University Press

Petr Sojka (editor): DML 2009, Towards a Digital Mathematics Library, pp. 17–24.
 ISBN 978-80-210-4781-5

well with mathematical scripts, at any level – not the glyph, the semantic nor the structural. Many of the symbols use tiny diacritics, are foreign or invented symbols, use subtleties of emphasis such as slant and inter-character spacing, and attach special significance to two-dimensional layout. Mathematical notation has little in the way of grammar or language models[2], so there is little or no redundancy to use for error reduction. Add into this the tendency of mathematicians to invent or define notation as they go along and you have a recognition problem that is insoluble with any foreseeable technology other than employing rooms full of mathematical "clerks" who could presumably be doing something better with their time.

In JSTOR, and probably more widely, the mathematics journals are digitized and presented just like the other journals,

- Bibliographic metadata is keyed, with the inclusion of occasional (and troublesome) snippets of TEX: JSTOR's application of TEX is limited to isolated layout of data and TEX is used to get a particular appearance, not the identity, of symbols and layout. For example the abstract for [5] contains

  ```
  p<sub>i</sub> sont definis par les limites <tex-math>
  $f_{i}\pm 0,5\sqrt{\chi _{(k-1;\alpha)}^{2}}/\sqrt{n}$
  </tex-math>  et les intervalles pour les P<sub>i</sub>
  par les limites  <tex-math>$F_{i}\pm 0,5\sqrt{\chi _
  {(k-1;\alpha)}^{2}}/\sqrt{n}$</tex-math>, sous la condition
   nł ? <sub>(k-1;?)</sub><sup>2</sup>.
  ```

  which somehow fails to capture the intent of the original author.
- The content is captured at 600 DPI bitonal, with grayscale or color images of contone regions captured and composited in where necessary. This is generally an acceptable resolution except when applied to very small glyphs, which abound in mathematical texts.
- OCR is limited essentially to Latin-1 plain text. For example the following text, again from [5]

  $$\mathbf{Pr}\{\,|\,\mathcal{L}_1\,|\leqslant L,\,|\,\mathcal{L}_2\,|\leqslant L,\ldots,|\,\mathcal{L}_{k-1}\,|\leqslant L\,\}\geqslant 1-\alpha\,. \qquad (2)$$

  was translated as

  Pr { 1, < L, I O2 1 <L, . 1.. , kk-l I < L } > 1 -. (2)

  This is typical, neither the layout of mathematics nor the majority of symbols survive.
- Search and Browse are unaware of any deeper significance to the content: your search strings have to literally be there.

The traditional solution to this problem when digitizing documents has been to represent the symbolic mathematics as images, and to make little effort to add meaningful text behind them. There is little or no general semantics behind most mathematical notation anyway – it encodes a series of assertions or transformations, which, while rigorous in context, draw on natural language

[2] For natural languages a language model eliminates unlikely character sequences

to define their meaning. Although any encoding of the notation should define both the appearance and the intent of the notation, in practice you get, at most, appearance, and often you do not even get that.

Discovery of appropriate content is also problematical, since differing nomenclature, formulations, notational conventions and ambiguities lead to difficulties in finding related materials. In the absence of reliable markup of any form it becomes impossible to rely on anything other than the natural language in the article. We suspect that natural language based text mining will have limited success on mathematical literature, and so in this domain, perhaps more than any other, domain specific approaches would yield substantially better recall[3].

Given the nature of the material, the digitization process is likely to generate errors that cannot be addressed via the normal process of OCR plus manual review of low confidence candidates. The material will have more errors or omissions than more prosaic material.

It therefore appears as if there are a set of problems that currently put the burgeoning digitized literature at risk of being poorly discoverable in online environments. We have inadequate mechanisms deployed that can accurately represent mathematical notation, especially for representing equivalence or similarity of notation, and the expertise necessary for curating and correcting and otherwise managing the collections resides in the community, and almost never with the archives and repositories.

2 Maths Literature is unique, just like everyone else

Of course, every specialization has unique challenges, and perhaps mathematics is no more difficult than, for instance, ancient manuscripts. However, the issues facing the development of digital mathematics libraries are more difficult than more prose oriented fields, and are exacerbated by the use of a particular set of notational conventions. They do however have analogs and parallels to problems encountered elsewhere. JSTOR is working with a wide range of discipline-specific communities and the manner in which those those efforts may directly help the mathematics community is examined here in a little more detail.

There are a number of elements discussed here that, when bought together, can facilitate the emergence of a digital maths library. For example, there are issues relating to the digital re-mastering of paper originals, the effective integration of diverse archives, the ability of the community to curate and manage its (virtual) collection, and the ability to discover the material in the collections and to work with it. JSTOR is investigating tools and resources that address these issues, at least to some extent. By providing a suite of tools and resources we hope the community will use these assets to build and manage

[3] In the information retrieval sense of the word recall is the percentage of relevant documents actually located

not only the collection, but to build tools for digital workbenches that allow more effective use of the material and the scholars' time and resources.

In the following section a range of the activities at JSTOR is discussed, and a vision is presented of how they could contribute to a digital mathematics library.

2.1 Open Annotations Project

The Open Annotations project[4] is working to define standards, mechanisms and reference implementations that permit the association of separately held source material and scholarly commentary. The intent is that an annotation, in the broadest sense, can be attached to any part of any URI addressable resource on the web. JSTOR is collaborating with members the Open Archives Initiative (OAI), Zotero[5] and UIUC Monk[6] teams. Essentially it enables the creation of a parallel structure to a set of existing repositories. Interpretations and annotations can be associated with any arbitrary subpart of the articles in the repositories, independent of the holder of that content. In practical terms archives such as JSTOR need only provide stable URIs with addressable subparts, and all other activities can operate on the annotation databases.

2.2 OAI-ORE resource maps of JSTOR

The Open Archives Initiative has produced a specification for Object Reuse and Exchange (OAI-ORE). ORE provides a solution to the problem of how to represent a resource that is composed of separate web-accessible resources, but where a URI cannot be assigned to this collection. OAI-ORE *aggregations* describe collections of other web resources. In collaboration with Rob Sanderson of Liverpool University JSTOR has constructed ORE resource maps for its collections, and those will be made generally available once the appropriate exposure mechanism is understood. With a combination of the Open Annotations Standard and OAI-ORE, the mechanisms would be in place to construct a digital library resource with only modest and generic access provisions from the content holders.

2.3 The Decapod Project

Thirteen years and a thousand plus titles of experience at JSTOR has proven how difficult it is to find and capture complete runs of journals. In our production

[4] The Open Annotations Project is funded by The Andrew W. Mellon Foundation `http://www.openannotation.org`

[5] Zotero is a Firefox extension to help collect, manage, and cite research sources `http://www.zotero.org`

[6] MONK provides over 500 classic texts along with tools to enable literary research through the discovery, exploration, and visualization of patterns. ... these tools are applied to worksets of texts selected by the user from the MONK datastore. `http://monk.lis.uiuc.edu/`

process the first, unavoidable and most time consuming part of the process is to determine a *title history* and to locate physical copies of the whole title, a process that can take months and occasionally years. Once the material is located digitization encounters more barriers that include the removing of the physical material to an off-site (or offshore) facility, physical damage to the material, expensive manual interventions, and shallow document structure representation. Decapod will address all of these issues and greatly facilitate the capture and generation of usable digital documents from bound or unbound material.

JSTOR's partners in Decapod[7] are the ATRC at the University of Toronto and the IUPR group at the University of Kaiserslautern/DFKI. As noted, the intent is to greatly simplify the digitization process. Michael Doob notes in his paper on Retrodigitization [1] that digitizing of the complete mathematics literature is necessary because of the length of its active and relevant life[8] as reflected in its continuing use as the basis of current research. As Doob also notes, the scanning process tends to be destructive, involving the de-binding of the material. This is sometimes acceptable when the material is commonplace, but many holders of the physical material are reluctant to forgo use of the material, even temporarily, while it is dispatched for scanning and are even less prepared to have it damaged. Any given run of a title is likely to have missing issues, and the costs of finding, and digitizing those missing issues can become prohibitive if it involves physical removal from libraries. Decapod will address these issues by providing an inexpensive attach case sized rig that can be taken into a library and can digitize the material then and there, and emit a fully structured, reflowable PDF or hOCR[9] with embedded fonts that precisely mirror the typeface in the original print.

Even when the document image has been captured the readily available OCR systems do not handle mathematics well, either the commercial systems or the open source ones such as Tesseract[10].

As noted previously, Decapod will generate fonts from the observed glyphs, hence ameliorating the OCR issue. Irrespective of the interpretation of the symbols and the resolution of the original scan, the visual rendition will be accurate and scalable and reflowable. Assuming that enough instances of each glyph occur in the document, the synthesized glyphs will be cleaner than the raw rasterized glyphs, which will, in turn lead to better recognition. Moreover, the existence of these clean, vector glyphs will allow specialized software to recognize likely mathematical notation by its distinctive spatial characteristics

[7] Decapod (`http://www.decapodproject.org`) is funded by the Andrew W. Mellon Foundation

[8] JSTOR is hoping to have a citation analysis tool available by the time of this paper on its "Data for Research" site at `http://dfr.jstor.org`, which should permit precise quantitative measurements of such claims for arbitrary cross-sections of the literature

[9] hOCR is a viewable HTML microformat capable of representing all document and process data

[10] Tesseract was created by Ray Smith of HP Labs, now at Google, and open-sourced by one of the authors of this paper.

without having to deal with all the other aspects of document capture and analysis.

2.4 Community based curation, correction and annotation

Poor quality originals or exotic scripts and language lead to higher error rates during the digitization process. This impacts discoverability and so it is desirable to address this shortfall. JSTOR is investigating various automated ways to reduce transcription errors, but these are currently more speculative and wilt take a while to deliver results. However, in a lower risk, more conservative approach JSTOR is participating in a project to use community curation to address the difficulties of obscure documents.

A collection of 19th century art auction catalogs is being used to test the approach and develop a reusable framework for community based curation[11] The catalogs contain notations in the form of hand-written notes in and around each item-for-sale (or "lot") record that can really only be interpreted by an expert in the field. These annotations represent the hammer (sales) price and other post-publication information, and they need to be interpreted and associated with the lot record. Within the system accredited members of the art-history community can review and correct the annotation boundaries and transcriptions. The transcription may be, but need not be a literal interpretation of the marking. They may also may be more interpretative, for instance by adding provenance data or other identifying information. In that sense there is an analogy with mathematical notation, in that a qualified contributor could enter any interpretative text, including, for instance, LaTeX markup or any appropriately labeled semantic transcription. The framework should not overly constrain the process, since this *is* community curation, and the framework must allow flexibility for the community to establish their own curation standards.

3 Vision

Therefore bringing all of this together, it seems that some combination of these resources could contribute towards digital mathematical libraries..

- OAI-ORE maps that represent a graph that associates original material via citation links, user entered links or machine generated similarity links. The original material could reside on repositories such as JSTOR, NUMDAM &c. The resource maps can reside anywhere and describe any arbitrary combination of articles, sub-parts of articles &c.
- Annotations and transcriptions held on Open Annotation servers (either hosted by the repositories or separately). For mathematics the annotations probably hold they key to providing math-specific capabilities in discovery and understanding, so abstracting them from the repositories seems both

[11] This project is in collaboration with several New York libraries and museums, and funded by the Andrew W. Mellon Foundation

natural and desirable. It is likely that mining the annotations will, in time, become as valuable as mining the texts themselves. Moreover, tit is not obvious at this time that the best notational representations are known, so allowing multiple parallel reference structures to exist allows earlier and richer standardization.
- Community curation resources for mathematical texts, possibly adapted from the auction catalogs prototype or independently implemented, using the lessons learned.
- Decapod will make it cost effective for community members to add new material to the archive, using local staff and inexpensive equipment.
- The synthetic-font elements of Decapod will allow mathematics oriented analysis components to identify and process mathematical notation from completed documents, hence allowing post processing and no involvement with the other elements of the digitization process. Since the Decapod pipeline is decomposable, this part of the process (font generation and document structure analysis) can be applied to existing corpora.

The need for annotation and aggregation capabilities is not peculiar to the mathematics community and so that some sense they are the easier ones to provide. The repositories must continue to provide stable URIs for the articles and URI conventions to access particular parts of the articles. Annotation clients such as Pliny[12] [2], Zotero, and math-specific clients will be needed to allow the markup of the documents and the posting of the annotations to servers.

Discovery and Content-based Similiarity. The issues of discovery and similarity seem, in general, much more difficult to address. There are two quite distinct issues, one of which will go away with time and the other that will be with us forever.

The universe of mathematical literature on paper, i.e. that was not born digital, is distinct and limited, and its accurate transcription is a finite task that can be aided by software and will one day be complete. It is likely that, with good tools, it will be possible to make born analog content as good as if it was born digital – including reCaptchas [4] for mathematics perhaps?

There is a deeper and more challenging issue of describing what mathematical notation actually means, or perhaps more tractably, what it is equivalent to. This would facilitate discovery, and even an incomplete version would advance that cause. There are two levels of description required. Firstly and being addressed is a formal way of expressing the visual appearance and local interpretation of the full range of mathematical notation (for instance MathML, LaTeX, OpenMath, Mathematica &c).

There is then the open question of whether it is possible to to describe the meaning or equivalence of notations. In maths literature the meaning of each notational element is described either in the surrounding text or elsewhere in the corpus (or at least it should be). It seems then, at least in

[12] `http://pliny.cch.kcl.ac.uk/whatIsPliny.html`

principle, one should be able to define relationships such as equivalence or composition between notational elements, perhaps by an exact description, perhaps by reference to common definitions elsewhere. Representation of such generalized relationship graphs is the the raison d'être for RDF and semantic web technologies, so the community should consider developing the tools and standards for representing and analyzing notational relationships. Using OAI-ORE (or equivalent) one can envisage the progressive construction of ever more detailed webs of relationships across the literature, rough and mostly manual at first, that can be augmented, refined and verified as the spread and quality of the network evolves.

4 Conclusion

The bookshelf looks well populated, it is time to start really organizing it and adding tools to the digital workbench; tools that will allow the analysis of the literature, the finding of new relationships and the creation of new knowledge.

Just as a physical library is more than the books, so a digital mathematics library should be more than just a collection of material. It should include the organizational tools, the repair and enhancement tools, the discovery tools and ultimately the synthesis tools to advance the field.

A repository like JSTOR can provide a home for the resources and the supporting services, it can help standardize tools and methods and help spread those tools and methods and practices between communities, it can provide a stable and long-lived platform where new tools can be deployed for the use and benefit of the whole community. It can guarantee the integrity and access to community contributed information, and can facilitate upgrades as standards change.

A repository such as JSTOR should not try to invent or build all of those tools, neither can it or should it dictate the governance of community efforts; it simply cannot provide quantity or quality of human resources necessary to mark-up or transform the content.

That being said, JSTOR is committed to efforts towards making a better digital library and being a partner in enhancing and extending the usefulness and value of the collections that it holds in stewardship.

References

1. Michael Doob: Small Scale Retrodigitization Department of Mathematics, The University of Manitoba, Canada
2. Thinking about interpretation: Pliny and scholarship in the humanities, John Bradley, Literary and Linguistic Computing, Vol. 23, No. 3, 2008. Published by Oxford University Press on behalf of ALLC and ACH.
3. See `http://www.openarchives.org`
4. See `http://recaptcha.net`
5. Confidence Intervals for the Frequency Function and the Cumulative Frequency Function of a Sample Drawn from a Discrete Random Variable, A. Naddeo, Review of the International Statistical Institute, Vol. 36, No. 3 (1968), pp. 313–318