PROVIDING NEXT-GENERATION TOOLS FOR SCHOLARS: JSTOR'S ADVANCED TECHNOLOGY RESEARCH GROUP

JOHN BURNS AND KEVIN S. HAWKINS

Abstract *The work of scholars is rapidly changing. As new digital resources and tools are developed, and old tools and resources reinvented for the digital world, the practice of scholarship is quickly adapting to the expectations that content should be accessible from anywhere, that it is a raw material to be manipulated, and that an excess of information is the major challenge facing scholars. Despite these expectations, the current generation of tools are inadequate for emerging scholarly practices. JSTOR's Advanced Technology Research (ATR) group has built and collaborated on a number of software projects and platforms that attempt to provide the next generation of tools for scholars. We provide an overview of these projects.*

CHANGES IN HOW SCHOLARS WORK

Scholars today work in an era of unprecedented access to information, with digitised library collections, online journals, and grey literature available to users nearly instantly. (Print-only literature, meanwhile, is increasingly marginalised because it is less convenient to access than online literature). Users of these online resources vary in their sophistication but make use of them in all phases of the scholarly information lifecycle: identification, locating and acquiring, filtering, distillation, analysis, synthesis, and dissemination. Today's digital repositories usually support only identification and acquisition, primarily of secondary literature. They are rather like an online bookshelf, supporting discovery of material and delivery to human eyeballs.

International Journal of Humanities and Arts Computing 4.1–2 (2010): 141–149
Edinburgh University Press
DOI: 10.3366/ijhac.2011.0013

www.eupjournals.com/ijhac

The brains behind those eyeballs, however, are different than they were a generation ago. As Renear and Palmer[1] point out, immersive reading, while central to the humanities, is not predominant in the sciences, where researchers typically read and compare multiple documents simultaneously and extract salient figures from them. Scholars today keep abreast of an increasing volume of literature – which is so easily available – without necessarily reading it all closely. A shift in reading habits from the print to online worlds has already been observed,[2] with scholars spending more time skimming abstracts than reading whole articles. It is possible that this change in reading style is at least partly a response to the relative poverty of interfaces to repositories of scholarly literature: screen real estate is limited, and on-screen annotation is difficult and distracting compared to using paper. Nevertheless, scholarly practice is changing, and archives of scholarly literature have to respond to changing consumption patterns, to changing content, and to collaborations across disciplines and distance.

Users have become more visually oriented in recent years, and while reliance on graphics is often seen to be a spillover from the 'sound-bite' culture and to indicate a lack of attentiveness, in our opinion it is a rational response to excess information and is a good use of technology that allows for the creation and display of such visualisations as never before. The human brain is extremely effective at discerning visual patterns, and when dealing with vast amounts of content and data, visualisations become the most effective cognitive tool for separating information from the mass of content.

With more data in machine-processable form than ever before, empirical methods of research are gaining ground even in fields of the humanities where they were previously unknown, in part because they were simply impractical. It is now possible to use rigorous statistical methods to test hypotheses in a way that humanities scholars could never do before.

Users are working across disciplines and encountering unfamiliar vocabulary and scholarly practices. With ever more mathematical botanists and theoretical archaeologists, we see scientists increasingly working in the humanities and arts, for instance, by building models of brush technique as a fingerprinting technology for paintings or by using MRI techniques to 'read' furled scrolls or peer through layers of paint. We expect every branch of academic endeavour to become increasingly quantitative, yet we cannot expect every scholar to be an expert in statistics, image processing, mathematical epidemiology (which can describe far more than epidemics), or any of a dozen of abstruse fields.

Since content is increasingly available through multiple channels, repositories will no longer be able to offer merely unique content; instead, they will have to bring content, packaged tools, expertise, and services to their users. Moreover, they will have to partner with their users to mediate expertise and tools across

the community of researchers. Repositories will also become content providers, providing not only derived content like specialised indexes and semantic mark-up but also a platform where scholars can annotate and manipulate data and scholarly literature, moving these private practices into a more public setting. Repositories should ensure that these activities can be performed in a principled, open and high-quality fashion; scholars want to be able to use content from various sources interoperably, manipulate it, and share it with colleagues. The greatest need for tool development is in these areas.

NEED FOR IMPROVED TOOLS

Hardware manufacturers are not developing tools that serve the niche needs of scholars. While large, high-resolution computer monitors have recently become affordable, making it possible at last to compare documents side by side, the e-book and mobile device industries are driven by the consumer market, with its orientation towards reading one text at a time, leaving e-readers inadequate for much of the work of scholars.

There are great opportunities for developing software that will help scholars. In addition to comparing documents, scholars also heavily annotate them, synthesise them, and collaborate around then. None of these functionalities are well-supported today in software, but the first generation of such tools have met acclaim among users. As academic content from digitised manuscripts to yesterday's scholarship become increasingly available online, it's possible to decompose, rearrange, explore, and present this content – and information about it – in ways that reveal structures, events, and relationships. This was simply not feasible in an analogue world. As quantitative data becomes easier to collect, there is great potential for using visualisation tools to make sense of this information.

Scholars need tools for making sense of a corpus of material that cannot possibly be read thoroughly – that is, tools for 'distant reading'.[3] Concordances can be generated on demand for arbitrary sets of texts to show the frequency of use of words or show when concepts and names came into or fell out of use – for example, when did 'consumption' become 'tuberculosis' in newspaper articles in the UK versus in the US?

Any corpus for empirical research will likely need to be assembled from various sources. Repositories of documents and data increasingly include URIs for items in the collection, allowing persistent reference by both users and machines, but unfortunately much content is still not amenable to machine processing because of the format of the content itself. For instance, page images of digitised texts can be referenced by page but not at a lower level such as

a paragraph or sentence. (This problem can be partially circumvented using 'region extraction', such as that allowed with the djatoka image server.[4])

Stable identifiers like URIs and heuristics for disambiguating print citations provide the foundation for new types of citation analysis and citation network exploration. While print citations refer only to previously published work, an online publication can include a dynamically generated list of citations of that work, allowing users to find subsequent work on the same topic. In addition, online publications can more easily integrate errata and retractions, neither of which is easily discovered in the print world. Stable identifiers are also important in recommender systems. While these are becoming more common in subject repositories, they are still rarely found in online journals.

Much data-driven research is unfortunately hindered by copyright law, but tools for 'non-consumptive' research – research that uses documents without requiring their copying and without needing to access the whole document – open up new avenues for scholars. Legal barriers to creating digital resources aside, the cost of digitisation and reformatting is unfortunately still too high for many researchers and institutions to undertake, even on a small scale.

Finally, while the Internet has made communication over great distances cheap and efficient, tools for collaboration across distances are still in their infancy. Scholars share research interests with researchers at other institutions but rely on applications like Skype and Google Docs for real-time collaboration; true online collaborative workspaces are rare. The Internet also provides opportunities for non-specialists to contribute to scholarship, whether through 'volunteer computing' (like SETI@home[5]), open access to scholarly literature, or 'crowdsourcing' to identify objects or solve problems in order to improve algorithms.

THE REPOSITORY AS A PLATFORM

Digital repositories of scholarly content, such as JSTOR[6] and HathiTrust,[7] are uniquely situated in our changing world: they sit at the confluence of scholarly practice, publishing, and practical technology infrastructure. It is entirely compatible with their missions as not-for-profit organisations to act as a testing ground and dissemination vehicle for advanced technologies developed in academia: users create tools for their research needs using the repository's infrastructure, accessed through an application programming interface (API), and repository platforms help share them with the larger research community, including those outside of the discipline of the developer. Exposing such large corpora so that others can build on them requires thorough APIs based on extensible standards such as Search/Retrieve via URL (SRU)[8] and Open Archives Initiative Object Reuse and Exchange (OAI-ORE).[9]

JSTOR was among the first to build the foundations of this model, and its Advanced Technology Research group has led the way in turning the JSTOR repository into a platform for research.

JSTOR ADVANCED TECHNOLOGY RESEARCH (ATR)

In 2007, JSTOR established Advanced Technology Research (ATR), a team dedicated to exploring technologies that could improve the utility of JSTOR collections and provide new tools to the scholarly community. ATR seeks to understand what users of JSTOR are trying to achieve and determine how JSTOR can make it easier for them to do so. More specifically, ATR exists to pose and answer not just the question 'what can we do?' but, more critically, 'what should we do?' and 'how do we deliver the capabilities to our users?'. Its key contribution to the work of JSTOR is assessing technologies (wherever they originate) to see how they are used and how they provide value to the user. It is innovation, not invention, which matters. ATR is interested in helping scholars be productive in any phase of the scholarly information lifecycle, not just when reading scholarly literature in JSTOR or elsewhere.

ATR has two primary foci: external collaborations that yield overall benefit to the scholarly community and internal projects that add or improve functionality in the JSTOR platform. ATR often collaborates with researchers at other institutions, choosing projects that will benefit large numbers of users. While projects are sometimes developed using thematic collections of content, the intent is to ensure that it is, or can readily be, generalised to a broad audience; project-specific solutions are avoided.

ATR'S PROJECTS

Querying the repository: Data for Research (DfR)

ATR's major effort to allow more effective discovery and use of traditional JSTOR journal content is the Data for Research (DfR) beta site,[10] which provides a web interface for non-consumptive research on the JSTOR archive. This workbench, open to all users, exposes the full text and metadata of the entire JSTOR archive (which includes some primary source material in addition to journal content) to allow users to filter and select sets of data by discipline, date, journal title, or particular search terms. From within the DfR interface, users can create graphs showing word frequency over time, enabling users to explore the corpus and spot trends and anomalies. For example, a search for any word spelt with the long 's' (but captured through OCR as 'f') shows a rapid transition around 1800 to use of only the contemporary 's'.

DfR automatically generates keywords from an article result set based on the frequency of a word in the set compared within the corpus as a whole. This feature can lead to interesting discoveries. For example, a search for 'James Joyce' generates 'reviewed author' facets for other prominent literary figures but also for Richard Ellman, who was Joyce's biographer – something not widely known outside Joyce studies.

The DfR website could not possibly provide a sufficiently generalised analytical tool for all scholars: such a tool would be difficult to create, support, and use. Better by far to allow scholars to choose their own tools to use in their own environment. DfR allows users to generate and download datasets in CSV format, which can be opened by Microsoft Excel and other programs commonly used to manipulate and analyse data. DfR allows users to select which information will be included in the export fields. Advanced users can use the DfR API to query the database. The service of DfR, admittedly, is not entirely altruistic: by performing data mining on its own collections, JSTOR can augment its human-created metadata with metadata generated through clustering to make it more discoverable for consumptive uses of the JSTOR archive (access to JSTOR is through a subscribing institution, by which users can access the full text of articles).

The DfR team actively collaborates with researchers doing large-scale corpus studies. JSTOR can provide data to researchers for text and data mining and, in limited circumstances, for studying usage data. For example, David Blei and colleagues use text mining to identify and track topics in literature as they evolve over time.[11] This research helps JSTOR assign keywords to journal articles, organising them into one or more fields and subfields in a more granular way than in the current organisation of journal articles by a single discipline, allowing faceted browsing of search results. Carl Bergstrom and colleagues are exploring new methods of calculating impact factors,[12] and JSTOR hopes to integrate the results of some of their research in the future. MESUR (MEtrics from Scholarly Usage of Resources)[13] uses JSTOR data as it attempts to measure impact of scholarly writing by measuring usage.

CURATION: AUCTION CATALOGUE PROJECT

By the very nature of academia, most advanced research depends on specialised source material that has a limited audience. We expect that repositories will provide their communities of interest with a readily configurable community curation platform that will allow users to 'community-source' the curation of their material. JSTOR has in recent years added non-journal content to its archive, and ATR has developed some tools to make these resources more useful.

For example, ATR staff worked on a project funded by the Andrew W. Mellon Foundation to enable community curation of nineteenth-century art auction catalogues.[14] These auction catalogues contain fielded data in print, with one piece of data (the hammer price) written by hand as a margin note or on an insert. The material is somewhat arcane, filled with abbreviations, antiquated language, and jargon. While the typeset text was adequately OCR'd using Abbyy FineReader Engine 9.0 and handwriting transcribed by a vendor, algorithms were developed to delimit transaction (lot) records from the 100,000 pages of these catalogues.

It was decided to encode manually 266 of the catalogues and then use these to train a machine learning system that would perform initial encoding of the remaining 1,400 catalogues. The system identified lots correctly in more than 90% of cases. All 1,400 catalogues were then made available to users not only for searching and browsing but also to correct lot identification and any transcription or markup errors using an interactive image of the catalogue pages that accepts both on-image drawing and text entry.

DIGITISATION: DECAPOD

While a few thematic non-journal collections, like the auction catalogues mentioned above, have already been added to JSTOR, ATR is developing tools to make digitisation more affordable, increasing the amount of primary source material that could be included in JSTOR and thus keeping print-only materials from being marginalised in the digital world. There is a huge need for affordable, easy-to-operate tools for non-destructive scanning of library and archival materials. The Andrew W. Mellon foundation is funding the Decapod project[15] to build a 'one-click' solution for converting paper to a structured, searchable digital document suitable for immediate posting online. It includes open-source software as well as instructions on how to use consumer hardware to build a look-down scanner. Decapod involves an international team of researchers, including ATR staff, who bring decades of experience in document processing to the project.

One of the design goals of Decapod is to create digital documents with no visible transcription errors resulting from the OCR process (though OCR errors might still exist in the character codes used for searching, leading to unavoidable search failures as with current technology). The software creates a custom font based on characters detected in the page images and embeds this font in the output (currently a PDF file but in the future possibly HTML5 or any other sufficiently expressive format). Use of a custom font allows not only graphic fidelity but also text reflow, allowing users to view the document on various screen sizes with various column layouts, all the while leaving the file size smaller than it would be if the pages were stored as raster images.

Lowering the cost of digitisation, and making digitised resources more useful to users, will allow scholars to use tools for analysis, manipulation, and sharing on materials that until recently were accessible in print format only.

STANDARDS FOR DIGITAL ANNOTATION: OPEN ANNOTATION COLLABORATION (OAC)

JSTOR is also a partner in the Open Annotation Collaboration (OAC),[16] which is defining a set of standards by which annotation clients can interoperate with resources to be annotated. Since each discipline has its own practices and standards (de facto or de jure) for marking up content, a standard annotation service must abstract from these practices and standards to achieve interoperability. As the standard matures, JSTOR will explore how to support the storage and presentation of the annotation service and how to provide clients that allow the annotation of the increasingly diverse content in JSTOR.

CONCLUSION

While scholarly resources in digital form have revolutionised the work of scholars, the tools for creating and using the information are seriously deficient compared to the needs of scholars. JSTOR's ATR attempts to address these deficiencies both by enabling research that takes advantage of JSTOR's huge corpus of scholarly literature, metadata, and usage information and by contributing toward community efforts to provide tools for the creation and use of digital resources.

END NOTES

[1] A. Renear and C. Palmer, 'Strategic reading, ontologies, and the future of scientific publishing', *Science*, 325, no. 5942 (2009), 828–832.

[2] See G. Buchanan's paper, 'The Usability of Digital Documents – a Barrier to Digital Scholarship' in the present volume, also N. Carr, 'Is Google making us stupid?', *The Atlantic*, 302, no. 1 (2008), 56–63.

[3] F. Moretti, *Graphs, maps, trees: abstract models for literary history* (London, 2005).

[4] *Djatoka Jpeg 2000 image server* < http://djatoka.sourceforge.net/ > (18 Aug 2010).

[5] *SETI@home* < http://setiathome.berkeley.edu/ > (14 Aug 2010).

[6] *JSTOR* < http://www.jstor.org/ > (14 Aug 2010).

[7] HathiTrust < http://www.hathitrust.org/ > (14 Aug 2010).

[8] *Search/Retrieve via URL* < http://www.loc.gov/standards/sru/ > (14 Aug 2010).

[9] *Open Archives Initiative Object Reuse and Exchange* < http://www.openarchives.org/ore/ > (14 Aug 2010).

[10] *Data for research, beta* < http://dfr.jstor.org/ > (14 Aug 2010).

[11] D. Blei and J. Lafferty, 'Dynamic topic models' < http://www.cs.princeton.edu/~blei/papers/BleiLafferty2006a.pdf > (14 Aug 2010).

[12] *eigenfactor.org: ranking and mapping scientific knowledge* < http://eigenfactor.org/ > (14 Aug 2010).

[13] *MESUR* < http://www.mesur.org/ > (14 Aug 2010).

[14] *Auction catalogs, beta* < http://auctioncatalogs.jstor.org/ > (18 Aug 2010).

[15] *Decapod* < http://sites.google.com/site/decapodproject/ > (14 Aug 2010).

[16] *Open Annotation Collaboration* < http://www.openannotation.org/ > (14 Aug 2010).