Reuniting data and narrative in scientific articles

Data and narrative are both important for scientific discourse. They are 'united' in the mind of the scientist-author, yet the current publishing process favours narrative at the expense of data, making it hard to recreate experiments, intuitively link to relevant data points outside the article in question, or indeed find associated data sets. Utopia Documents helps to solve this problem by enabling readers of articles to follow leads without the need to flip (and particularly search) through other systems, and so enhances their ability to gain a deeper understanding of the arguments being presented. Utopia Documents also removes the linkability barriers hitherto inherent in the PDF format, by bridging the connection gap with the web.

STEVE PETTIFER
Senior Lecturer
Computer Science
School of Computer Science
University of Manchester

JAN VELTEROP
CEO
Academic Concept Knowledge Limited (AQnowledge)

TERESA K ATTWOOD
Professor of Bioinformatics
Faculty of Life Sciences and School of Computer Science
University of Manchester

LEE HARLAND
Director
ConnectedDiscovery Limited

JAMES MARSH
Research Fellow
School of Computer Science
University of Manchester

DAVID THORNE
Research Fellow
School of Computer Science
University of Manchester

ALEC TUNBRIDGE
Research Assistant and Software Developer
Academic Concept Knowledge Limited (AQnowledge)

Introduction

It is tempting to see the problems faced by contemporary scientists trying to keep abreast of the latest developments in their field as being a new phenomenon – an unwanted side-effect of the power of the web to propagate information across the globe in a matter of seconds. Yet, this is not a new problem: throughout history, scholars have been demonstrably better at creating knowledge than they have at putting in place mechanisms for disseminating, accessing and exploiting it effectively. Issues of egalitarian access to 'the literature', foreshadowing what today we would consider to be matters of 'open access', date as far back as the 'public' libraries of Ancient Rome[1]. Even the term 'information overload' pre-dates the internet revolution by at least a decade[2]. The details and absolute scale of the problem have changed considerably over time: since the move to digital publishing, for example, physical barriers such as 'geographical distance from nearest library' or 'weight of paper to be shipped' are no longer issues. Politics, commercial interests and simple resistance to change remain as constant brakes on progress. There has always, for one reason or another, been more knowledge available to mankind than any one human can access, much less take in.

There is, undeniably, a growing issue of scale here, and one that has been reported and analysed repeatedly in the literature[3,4]. Although the details of various commentaries on the matter differ, their conclusions are consistent: we are publishing more than ever before; the rate of publishing is increasing; and the tools available to make sense of the growing body of knowledge continue to lag behind our need to understand and exploit it. In some senses, the problem of dealing with the overwhelming amount of literature available can be characterized, as Clay Shirky suggests[5], as simple 'filter failure', but the creation of a suitably sophisticated filter for 'scientific relevance and quality' remains elusive, and is the subject of much ongoing research.

"… the tools available to make sense of the growing body of knowledge continue to lag behind our need to understand and exploit it."

Aside from scale, however, a perhaps more fundamental change has occurred in the nature of what is being published in the scientific arena. Whereas early discoveries could be encapsulated in a few pages of largely self-contained prose (Watson and Crick's 'A Structure for Deoxyribose Nucleic Acid'[6], for example, although rather scant on technical detail, communicates the essence of its world-changing discovery in just short of a page), contemporary articles are increasingly reliant on data to make their case. Unlike

the manually observed and recorded data associated with early discoveries, new data sets are often the result of automated processes or computational simulations, and are frequently orders of magnitude larger and more complex than those created by hand in the past. (The life sciences, for example, have embraced a system-wide culture of 'big data' collection, spawning the high-throughput omics revolution.) The modern scientific article has been described as a 'story that persuades with data'[7], a phrase that elegantly brings together the two essential components of modern scientific writing: without the story, data are just a collection of facts with no useful interpretation or real meaning; without data, a story is in danger of being seen by the community as little more than an opinion – a kind of fairy tale even. A further consideration is that, as more data sets are held electronically, the means to verify them change. There is now a real issue with the lack of reproducibility of research[8]: the more data we capture, the more meticulous we must be also to capture and utilize their provenance, in order to allow readers to validate the 'facts' before them.

"… the more data we capture, the more meticulous we must be also to capture and utilize their provenance …"

The mainstream actors in the scientific publishing process have been slow to recognize and embrace this change. For the most part, modern online journals manifest themselves as digital facsimiles of their traditional paper-based counterparts, where the narrative component dominates the associated data. There are exceptions, such as the nascent journal, *GigaScience*[9] that focuses on publishing big-data-led science. However, most scientific studies remain locked into a historical format. Even the terminology used suggests that data are treated as second-class citizens, typically being relegated to 'supplementary' or 'auxiliary' documents, often hosted in *ad hoc* and volatile repositories[10], and published in a similar form, often PDF, to the associated narrative article. This, in part, has led to significant criticism of the PDF as a vehicle for scientific communication. The core issue is that the narrative component of a scientific publication is an essential distillation of the authors' thoughts, to be read and understood by another human. For this, the PDF, when used properly – an important caveat, since sadly, many of the PDFs created today fail to take advantage of features such as structural mark-up, hyperlinking and accessibility that have been present in the format for some considerable time; a limitation of the publishing process, however, not the PDF specification – is a reasonable (if unremarkable) format. In fact, publishers' download statistics attest that the PDF is the pre-eminent format for human interpretation of natural language text, significantly more popular than HTML. The associated data, on the other hand, are increasingly only meaningful with the help of a machine, and publishing *these* in PDF form unnecessarily obfuscates their electronic analysis and re-use. Because the two aspects – readability by humans and readability by machines – are usually conflated, and not just by publishers, it is perhaps understandable that the PDF sometimes attracts scientists' ire.

In selected areas, change is occurring. If an article describes a novel DNA sequence or molecular structure, for example, some journal editors and publishers require the underlying data to have been deposited in an appropriate public database, and that the accession number be cited in the article. Elsevier, for example, has recently announced the adoption of a standard way of referring to databases and records to be encouraged throughout their life science journals[11]. Even so, only a relatively small number of journals have such requirements and, seen in the context of the whole of scientific publishing, such mandates remain as valuable but rather piecemeal solutions to a generic issue that faces the entire community. To complement these domain-specific approaches, numerous more generic data-publishing systems have appeared (e.g. Dryad[12,13,14], Pangaea[15,16], DataCite[17], DataOne[18,19], UK Data Archive[20], encouraging authors to deposit raw data files in return for stable, citable identifiers.

While these generic systems offer long-term storage, accessibility and, in some cases, citability for data sets, the data are stored away from their original context, hampering their future use. They are also less amenable to discovery of data points for individual elements (e.g. finding data on a single gene across different data sets). In some cases, domain-specific repositories have been created that do provide much more tailored indexing (such as ArrayExpress[21] and Gene Expression Omnibus (Geo)[22] for microarray data, ChEMBL[23]

and PubChem[24] for pharmacology). However, it is still challenging for readers of a paper to jump back and forth between the article and these online repositories. The problem is exacerbated when the scientific conclusion is derived from combining analyses across different experimental data sets: say, combining genomics data with *in vivo* and *in vitro* assays. While tools such as the Investigation-Study-Assay (ISA) system[25] provide the means to record the required metadata to connect these together, there is still a lot of work to be done by the user reading the paper to enable them to quickly look up facts and conclusions. The notion that 'reading scientific literature often feels like trying to imagine what the tapestry depicts, while only seeing the back of the embroidery' will doubtless be recognizable to many a scientist.

"… reading scientific literature often feels like trying to imagine what the tapestry depicts, while only seeing the back …"

These problems and others have been the topic of much recent interest, and the subject of numerous workshops and international conventions (the ongoing work of which is represented by the Force11 community[26]). Integrating and broadening these initiatives, continuing to influence the way in which authors construct articles, and improving the mechanisms that publishers use to disseminate them, is likely to be a lengthy process, and one that is as much about changing minds as it is about creating or deploying novel technology. Even putting aside the deadly embrace caused by the interaction between impact factors, funding bodies, publishers and authors, scientists are understandably slow to abandon or modify approaches that have served them well in the past. At the same time, small publishers frequently lack the resources to make radical changes, while larger publishers are often hobbled by the inertia of their own processes.

What is also apparent is the disconnect between the practices of data archiving and the publication of scientific data and narrative. While the content and functionality of experimental data repositories continue to expand inexorably, and the scientific literature grows apace, there is no seamless link between them. Most scientists probably view databases and articles as fundamentally separate entities – they are more likely to be drawn to data via the scientific discourse in an article than directly via a repository or archive. Some articles may have accession numbers here or there, but there is still no intuitive link between them. Data repositories thus seem likely to grow in sophistication, to better serve the communities of informaticians who understand and use them, rather than reaching out to 'average' bench scientists and the publishers who serve them.

Another difficulty of bringing data and narrative together – and connecting narrative with narrative – is the lack of truly standardized vocabularies. The human mind is extraordinarily capable of disambiguating textual uncertainties and convolutions not only from context, but often from a whole range of circumstances that are not represented, or even representable, in such a way that machine interpretation is feasible. Researchers' inventiveness and natural 'sloppiness' with scientific terminology usually demands human reading to extricate the meaning of narratives; keeping up with, and reading, the growing numbers of published articles, however, is increasingly difficult for humans to achieve without the use of computers.

"… a mechanism is needed for bringing data and narrative together in a semantically useful way."

Clearly, a mechanism is needed for bringing data and narrative together in a semantically useful way. This becomes even more important in light of publishers' large back catalogues of PDFs, for which issues of text-data connectivity were not a consideration when the PDFs were made, and which have therefore been largely dismissed as semantic dead-ends.

Utopia Documents

Utopia Documents is a PDF viewer for scientific articles, designed (amongst other things), to (re)connect data and narrative. Focusing, in particular, on papers in the life sciences, medicine and chemistry, it provides convenient mechanisms for readers to look up data relating to articles and their contents without having to swap to other applications in order to be able to do so – this includes data not known or available to the authors, but generated or made available post-publication.

When an article is loaded into Utopia Documents, the software examines its typographic layout, and text and figure content, in order to identify the paper and its features. Various online repositories, including CrossRef and PubMed, are queried to confirm the article's identity, and to retrieve authoritative metadata, which includes article identifiers (PubMed ID, DOI, ISBN/ISSN, etc.), title, author names, and so forth. These, in turn, are used to query other services that return information relevant to the current article. A growing number of such services are queried, including (but not limited to):

- Altmetric (to discover, track and analyse online activity related to an article)
- PLOS Article-Level Metrics (to give readers of PDF versions the same up-to-date information on metrics as those reading the web versions)
- Mendeley (to identify and locate related articles)
- Pubmed Central (for linkable references)
- SciBite (for patents, news, alerts on critical topics in biomedicine)
- ChEMBL/Chemspider (for bioactive, drug-like small molecules)
- PDB (the Protein Data Bank, for protein 3D co-ordinates)
- UniProtKB (the Universal Protein Knowlegebase, for protein sequences)
- AQnowledge (for laboratory materials and supplies).

The point of embedding links to these and other data services in the text of the PDF is to make life considerably easier for readers when having to jump between articles and data repositories; and, because this embedding is done 'on the fly', the links are always up to date, no matter how old the PDF.

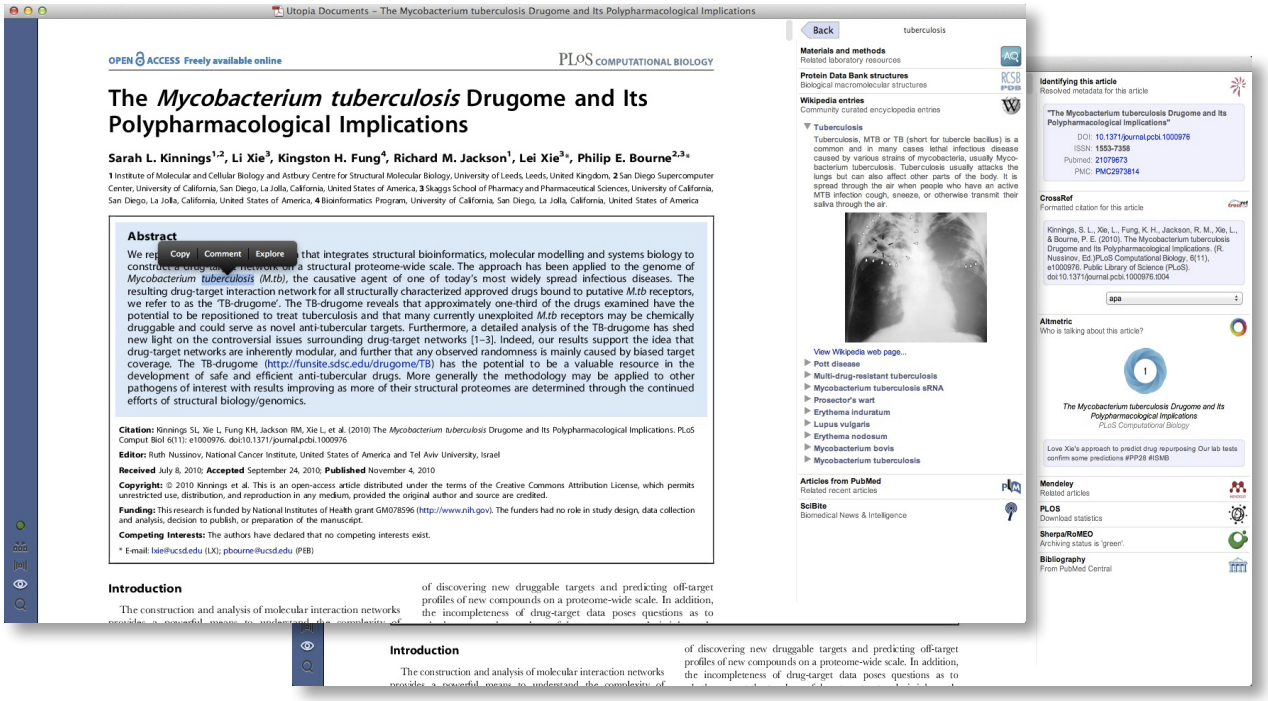

Figure 1. Screenshots of Utopia Documents showing (in front) the results of 'exploring' the term 'tuberculosis', and (behind) information relating to the document as a whole

Apart from the data sources that are queried by default – and from which search results are presented – readers can highlight terms or phrases, and look those up in many other data repositories. The ability to follow leads without the need to flip (and particularly search) through other systems enhances scientists' abilities to gain deeper understanding of the

arguments being presented. For example, finding that a particular gene is over-expressed in a cancer study may take on greater significance if it is connected to a pathway, an inhibitor or even a seemingly unrelated disease, and those connections are made tangible to the reader.

But it is not just linking out to data and information sources that is needed for properly (re)connecting narrative and data. The ability to process and manipulate data, for instance, to view numerical tables in the form of a graph, temporarily generated within an article, or to view a protein structure in rotatable 3D, is of great benefit to researchers and students alike. For open access articles, many of these more sophisticated features can be made available to users without the collaboration of publishers. For content that is not open access, or in case open access publishers wish to include features that they would not otherwise be able to deliver, collaboration with Utopia Documents ensures that they are. PDF versions of articles can offer linking and semantic features every bit as sophisticated, useful and up to date as their HTML counterparts. The Utopia Documents semantic PDF reader bridges the connectivity gap between PDFs and web versions of scientific articles, without modifying or needing changes to the PDF itself. This means that PDFs from back catalogues (with the exception of PDFs consisting purely of page scans: bitmaps) are equally enriched when read with Utopia Documents as those generated now.

"PDF versions of articles can offer linking and semantic features every bit as sophisticated, useful and up to date as their HTML counterparts."

One argument against building an enhanced PDF reader is that 'well, they will just print the PDF out anyway'. But what if you make the reading of the PDF electronically so rewarding that this becomes the poorer relation? In fact, with Utopia Documents most PDF versions of articles are better connected to data sources relevant to them than their HTML counterparts.

References

1. Dix, T K, 'Public Libraries' in Ancient Rome: Ideology and Reality, *Libraries and Culture,* 1994, 29, 282–296.
2. Toffler, A, *Future Shock*,1970, Random House.
3. Attwood, T K *et al,* Calling International Rescue: knowledge lost in literature and data landslide! *Biochemical Journal,* 2009, 424, 317–333; doi:10.1042/BJ20091474 .
4. Talbot, M D, On the impossibility of being expert, *BMJ,* 2010, 341; doi:10.1136/bmj.c6815 .
5. 'It's Not Information Overload. It's Filter Failure', video, Clay Shirky at Web 2.0 Expo NY, 16–19 September 2008 (accessed 19 September 2012).
6. Watson J D and Crick F H C, A Structure for Deoxyribose Nucleic Acid, *Nature,* 1953, 171, 737–738.
7. De Waard, A, Stories that persuade with data, lecture at KMDI at the University of Toronto, 30 September 2010: http://www.slideshare.net/anitawaard/kndi-toronto-panel (accessed 19 September 2012).
8. Keogh, E, Why the lack of reproducibility is crippling research in data mining and what you can do about it. In: *Proceedings of the 8th international workshop on Multimedia data mining: (associated with the ACM SIGKDD 2007)* (MDM 2007), 2007, Article 2, ACM Digital Library, New York, NY, USA; doi:10.1145/1341920.1341922 .
9. *GigaScience*: http://www.gigasciencejournal.com (accessed 17 September 2012).
10. Anderson, N R, Tarczy-Hornoch, P and Bumgarner, R E, On the persistence of supplementary resources in biomedical publications, *BMC bioinformatics,* 2006, 7, 260; doi:10.1186/1471-2105-7-260 .
11. Example from Elsevier's Guide for Authors: http://www.elsevier.com/wps/find/L03_410.cws_home/main (accessed 17 September 2012).
12. Dryad: http://datadryad.org (accessed 17 September 2012).
13. Vision, T, Open Data and the Social Contract of Scientific Publishing, 2010, 60, 330-330; doi:10.1525/bio.2010.60.5.2 .
14. Greenberg, J, Metadata for Scientific Data: Historical Considerations, Current Practices, and Prospects, *Journal of Library Metadata,* 2010, 10, 75–78; doi:10.1080/19386389.2010.520262 .

14. Greenberg, J, Metadata for Scientific Data: Historical Considerations, Current Practices, and Prospects, *Journal of Library Metadata,* 2010, 10, 75–78; doi:10.1080/19386389.2010.520262 .
15. Pangaea:
http://www.pangaea.de (accessed 17 September 2012).
16. Grobe, H, Sieger, R, Diepenbroek, M and Schindler, U. In: *Cool libraries in a melting world: Proceedings of the 23rd Polar Libraries Colloquy*, 2010, Bremerhaven, Klimahaus.
17. DataCite:
http://www.datacite.org (accessed 17 September 2012).
18. DataONE:
http://www.dataone.org (accessed 17 September 2012).
19. Allard, S, DataONE: Facilitating eScience through Collaboration, *Journal of eScience Librarianship* 1, 2012; doi:10.7191/jeslib.2012.1004 .
20. UK Data Archive:
http://www.data-archive.ac.uk (accessed 17 September 2012).
21. ArrayExpress:
http://www.ebi.ac.uk/arrayexpress (accessed 17 September 2012).
22. Gene Expression Omnibus (Geo):
http://www.ncbi.nlm.nih.gov/geo (accessed 17 September 2012).
23. ChEMBL:
https://www.ebi.ac.uk/chembl (accessed 17 September 2012).
24. PubChem:
http://pubchem.ncbi.nlm.nih.gov (accessed 17 September 2012).
25. Investigation-Study-Assay (ISA) Commons:
http://isacommons.org (accessed 17 September 2012).
26. The Future Of Research Communications and E-scholarship (Force11):
http://www.force11.org (accessed 17 September 2012).

Corresponding author:

Dr Johannes (Jan) Velterop, CEO
Academic Concept Knowledge Ltd (AQnowledge), Denham House, Norman Avenue, Epsom, Surrey KT17 3AB, UK
E-mail: velterop@aqnowledge.com

To cite this article:

Pettifer, S et al, Reuniting data and narrative in scientific articles, *Insights*, 2012, 25(3), 288–293, doi: 10.1629/2048-7754.25.3.288

To link to this article:

http://dx.doi.org/10.1629/2048-7754.25.3.288