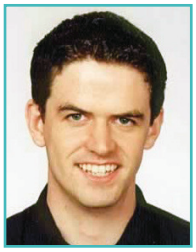
Utopia Documents and The Semantic Biochemical Journal experiment



Teresa K.Attwood*†, Douglas B.Kell‡§, Philip McDermott*†



James Marsh‡, Steve R.Pettifer* and David Thorne‡

* School of Computer Science
† Faculty of Life Sciences
‡ School of Chemistry, The University of Manchester, UK
§ Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester, M1 7DN.

Introduction

Recent technological advances have led to the accumulation of data on an unprecedented scale. Adding to this information overload is the advent of desk-top sequencing, with machines capable of delivering terabytes of data per hour. The problem is, the rush to increase the amounts of information we collect does not in itself bestow a miraculous increase in knowledge. For information to be usable, it needs to be stored and organised in ways that allow us to access it, to analyse it, to annotate it and to relate it to other information. Unfortunately, to date, we have failed to store and organise much of the rapidly accumulating information (whether in databases or documents) in rigorous, principled ways, so that finding what we want and understanding what's already known become increasingly exhausting, frustrating and costly experiences.

Scientists for whom these problems have become especially acute are database curators. Today, the largest protein sequence database in the world is UniProtKB [1]. UniProtKB currently contains >9 million entries, of which ~500,000 have been contributed by its manually-annotated component, Swiss-Prot [2]. By inspecting thousands of articles and hundreds of other database entries, it has taken 23 years for the Swiss-Prot curators to annotate about half of these sequences – indeed, Bairoch estimates that this gargantuan task has involved 600 person years of effort [3]! The difficulties faced by the curators are enormous: with ~25,000 peer-reviewed journals publishing ~2.5 million articles per year, this equates to something like two new papers appearing in Medline every minute [4]. Consequently, it is impossible for curators to keep abreast of developments, and more and more difficult for them to find relevant papers, or to locate relevant facts within them. It isn't really surprising, then, that Bairoch should opine, "It is quite depressive to think that we are spending millions in grants for people to perform experiments, produce new knowledge, hide this knowledge in a often badly written text and then spend some more millions trying to second guess what the authors really did and found" [3].

The life of curators would be a lot easier if articles could become better conduits to their underlying research data. In fact, it has already been argued that the distinction between an online paper and a database is already diminishing [5]. Nevertheless, it is clear that much more needs to be done to make the data contained in research articles more accessible.

In this Letter, we briefly outline a new development with Portland Press Ltd., the so-called Semantic *Biochemical Journal* experiment [6]. Behind this 'experiment' is a new software tool, Utopia Documents, which builds on the Utopia suite described in previous EMBnet.news articles, and elsewhere [e.g., 7-9]. Here, we provide a sketch of these new developments, in order to provide a taster of what can be achieved through academic-journal-publisher collaboration.

The "experiment"

Utopia is a software suite that semantically integrates visualisation and data-analysis tools: its most recent component, Utopia Documents, brings document-reading and document-

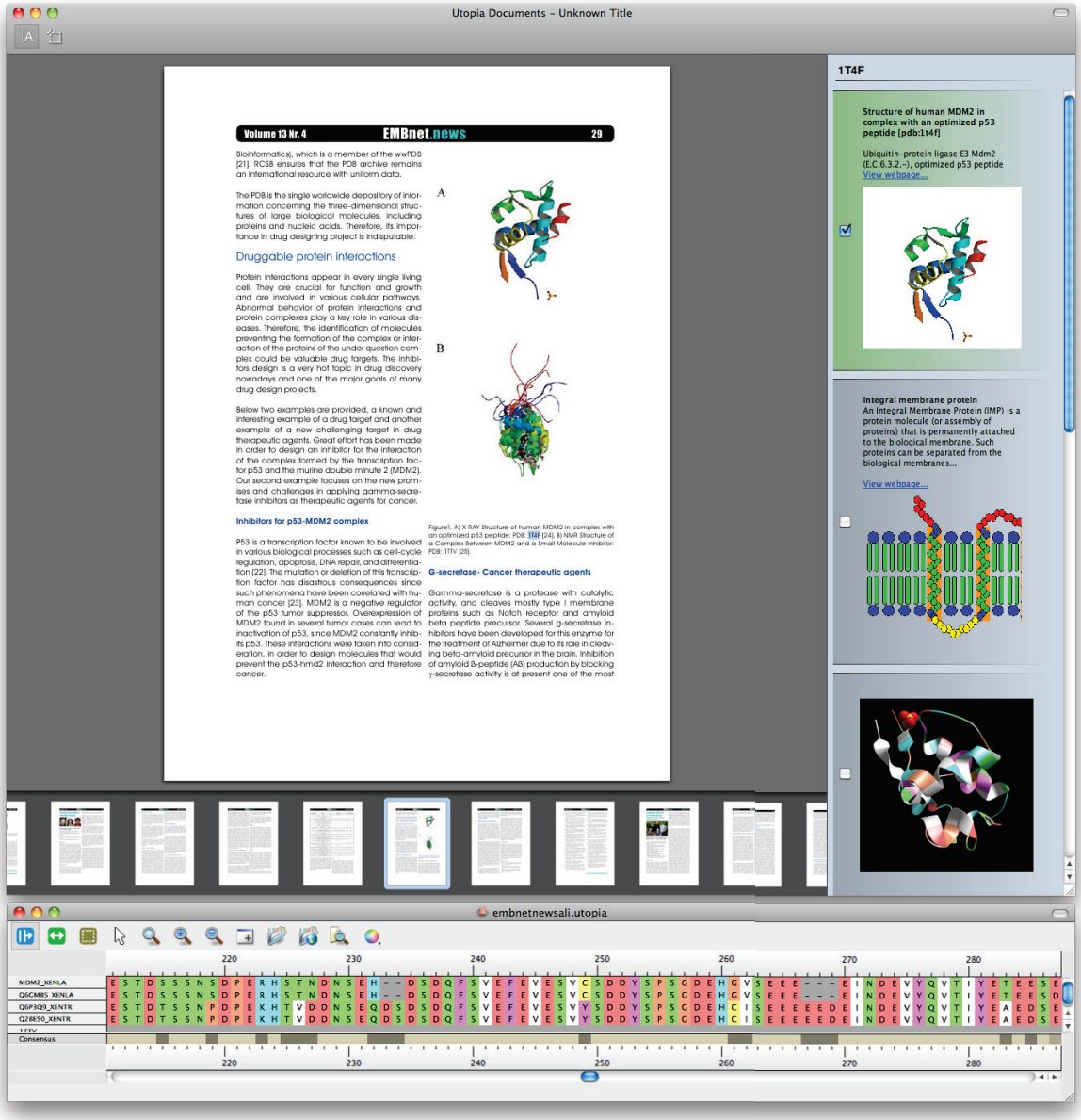


Figure 1. Composite screen-shot illustrating some of the features of Utopia Documents. Dominating the Figure is a page from EMBnet.news (volume 13, issue 4, page 29). Clicking on terms of interest in the text (e.g., 1T4F - the highlighted term half-way down the second column of text) provides definitions from various online databases, dictionaries or thesauri, and uses RDF-linked data to infer and retrieve related information. Here the reader has accumulated information about the 1T4F molecule from the PDB, a definition of 'membrane protein' from Dbpedia, and an interactive visualisation of 1T4F using Utopia's Ambrosia molecular viewer, as well as a related protein alignment, viewed using the CINEMA alignment editor. Hence, from a single page in an article, access is gained to information from databases, from online dictionaries and encyclopaedias and to interactive analysis tools, without having to leave the context of the PDF document.

management utilities to the suite. The aim of the *Semantic Biochemical* Journal (BJ) experiment was to use Utopia Documents to make the content of BJ electronic publications and supplemental data richer and more accessible. To achieve this, Utopia was integrated with in-house editorial and document-management workflows, allowing the BJ editors to mark up article content prior to publication.

The Utopia Document PDF-reader creates unique fingerprints of document contents as they are rendered onscreen, identifying key typographical and bibliometric features (authors, references, etc.). Its innovation lies in being able to turn static features of a document into objects that can be linked, annotated, visualised and analysed interactively. In so doing, the document is transformed from a digital facsimile of its printed counterpart into a gateway to related knowledge, providing readers with focused interactive access to analysis tools, external resources and the wider literature.

As part of the experiment, the journal editors have marked up papers in the December 2009 issue of the BJ using Utopia Documents. Aspects of these articles relating to protein sequence and structure analysis have been the main targets for mark-up, in the first instance, because this was the functionality built into the original Utopia toolkit. The kinds of additional mark-up currently provided by the software include: links from the text to external Websites (e.g., to databases like UniProtKB, PDB [10]) and InterPro [11]; term definitions from ontologies and controlled vocabularies; extra embedded data and materials (images, videos and the like); and links to interactive tools for sequence alignment and 3D molecular visualisation. Utopia does not itself provide any domain-specific functionality for processing or analysing data, but relies on external Web services – these are accessed via plug-ins whose appearance in the software interface is mediated by a 'semantic core' (which can be customised to any subject area by incorporating the relevant discipline-specific ontologies).

Reliance on external Web services is both a strength and weakness of the system: whereas it allows greater flexibility for customising the functionality of the suite, it also depends on the reliability of the external services it exploits – if these become unavailable (e.g., owing to routine maintenance or some kind of faulty operation), their functionality becomes unavailable to Utopia. These issues afflict *all* systems that rely on Web services, but are mitigated to some extent by the establishment of a Web-service registry, which systematically monitors and provides status reports on its registered services [12].

Future work

Utopia Documents is still at an early stage of development and there is much more work to be done. As the system is readily customisable, we plan to extend its scope, especially to encompass chemical biology – here, in particular, we plan to explore collaborations with the Royal Society of Chemistry, who have done pioneering work with their Prospect software (http://www.rsc.org/Publishing/Journals/ProjectProspect/). We are also embarking on exploratory discussions with Nature and Elsevier, and various pharmaceutical companies, in order to deliver bespoke mark-up and document-management solutions for these companies.

Another possibility that we're keen to explore is the use of Utopia Documents to mark up issues of EMBnet.news, which could add value at a critical time, as EMBnet.news becomes a *bona fide* peer reviewed publication – see Figure 1.

Find out more

The Semantic Biochemical Journal was formally launched on 10 December 2009. To gain further insights into the status of the project, and to better appreciate how this might benefit EMBnet.news in future, we encourage readers to view articles in volume 424(3) of the BJ (http://www.biochemj.org/bj/424/3/default.htm?S=0), and especially to read the launch article, Calling International Rescue: knowledge lost in literature and data landslide! (http://www.biochemj.org/bj/424/0317/4240317.pdf) - a preview of this interactive paper is given in the following video: http://www.youtube.com/watch?v=rIOgARI1a3E. Utopia Documents itself is available for download from http://getutopia.com/.

Funding

Utopia Documents has been funded by the European Union (EMBRACE, grant LHSG-CT-2004-512092), the Engineering and Physical Sciences Research Council (Doctoral Training Account), the Biotechnology and Biological Sciences Research Council (Target practice, grant BBE0160651), and

Portland Press Limited (The Semantic Biochemical Journal project).

References

1. The UniProt Consortium (2009) The Universal Protein Resource (UniProt). Nucleic Acids Res 37: D169-D174.
2. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31: 365-370.
3. Bairoch A (2009) The future of annotation/biocuration. Nature Precedings doi:10.1038/npre.2009.3092.1.
4. Hull D, Pettifer SR, Kell DB (2008) Defrosting the Digital Library: Bibliographic Tools for the Next Generation Web. PLoS Comput Biol 4: e1000204.
5. Shotton D, Portwin K, Klyne G, Miles A (2009) Adventures in semantic publishing: exemplar semantic enhancements of a research article. PLoS Comput Biol 5: e1000361. Doi:pcbi.1000361.
6. Attwood TK, Kell DB, McDermott P, Marsh J, Pettifer SR, Thorne D (2009) Calling International Rescue – knowledge lost in literature and data landslide! Biochem J 242: 317-333.
7. Pettifer, S., Attwood, T.K., McDermott, P., Sinnott, J. and Thorne, D. (2007) UTOPIA: User-friendly Tools for OPerating Informatics Applications. EMBnet.news 13: 19-24.
8. Sinnott JR, Pettifer SR, Attwood TK (2004) Introduction to the CINEMA5 sequence alignment editor. EMBnet.news 10(3).
9. Pettifer, S., Thorne, D., McDermott, P., Marsh, J., Villeger, A., Kell, D.B. & Attwood, T.K. (2009) Visualising biological data: a semantic approach to tool and database integration. BMC Bioinformatics 10: S19.
10. Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE, Berman HM (2006) The RCSB PDB information portal for structural genomics. Nucleic Acids Res 34: D302–D305.
11. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn R, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJA, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37: D211-D215.
12. Pettifer S, Thorne D, McDermott P, Attwood T, Baran J, Bryne JC, Hupponen T, Mowbray D, Vriend G (2009) An active registry for bioinformatics web services. Bioinformatics 25: 2090 - 2091.t