We live in interesting times. Portents of impending catastrophe pervade the literature, calling us to action in the face of unmanageable volumes of scientific data. But it’s not so much data generation per se, but the systematic burial of the knowledge embodied in those data that poses the problem: there is so much information available that we simply no longer know what we know, and finding what we want is hard – too hard. The knowledge we seek is often fragmentary and disconnected, spread thinly across thousands of databases and millions of articles in thousands of journals. The intellectual energy required to search this array of data-archives, and the time and money this wastes, has led several researchers to challenge the methods by which we traditionally commit newly acquired facts and knowledge to the scientific record. We present some of these initiatives here – a whirlwind tour of recent projects to transform scholarly publishing paradigms, culminating in Utopia and the Semantic Biochemical Journal experiment. With their promises to provide new ways of interacting with the literature, and new and more powerful tools to access and extract the knowledge sequestered within it, we ask what advances they make and what obstacles to progress still exist? We explore these questions, and, as you read on, we invite you to engage in an experiment with us, a real-time test of a new technology to rescue data from the dormant pages of published documents. We ask you, please, to read the instructions carefully. The time has come: you may turn over your papers...

Key words: dynamic document content, interactive PDF, linking documents with research data, manuscript mark-up, mark-up standards, semantic publishing.

INSTRUCTIONS TO READERS

Before reading any further, we are going to ask you to download a piece of software. Together, as we journey through this article, we will test the software [a new PDF document reader, called Utopia Documents (UD)] in different scenarios. You are, of course, free to read on without installing the software; however, for those of you reading the PDF version of this article, seen through the lens of UD, much more functionality will be revealed and the test will become tantalizingly more interesting.

To install UD, please visit the abstract page for this article (at www.BiochemJ.org), or http://getutopia.com/. The installation process is straightforward: simply follow the link to the website, and the guidance notes there will talk you through the software installation for your platform of choice.

Once you have successfully downloaded UD, you are ready to read on. As you do so, look out for the UD logo: ☛. This is used to draw your attention to interactive features, pinpointing where to click on particular icons. During the test, the story will unfold gradually and the interactive features will grow in complexity. We invite you to explore the increasing functionality at your leisure (for more adventurous, full documentation is available from the installation site).

INTRODUCTION

New technologies that promise to transform our lives excite us, but often come with unanticipated side-effects. Just think about life before email, laptop computers or mobile phones and it’s clear that as much as they’ve improved some aspects of our lives, they’ve made significant demands on us in others: e.g. to learn how to use yet another new gadget, to navigate yet another new interface, to cope with the daily bombardment of (often irrelevant) communications – in short, to control the technology before it controls us. Getting the balance right can be a struggle.

The life sciences have not been immune from these effects. Technological advances have led to the accumulation of data on a scale unthinkable only a couple of decades ago, promising to revolutionize how we ‘do’ biology and to have dramatic impacts on our understanding of such processes as gene expression, drug discovery, and the progression and treatment of disease [1,2]. Yet the metaphors of doom used to describe the phenomenal pace of data acquisition (from data floods [3], deluges [4,5], surging oceans [6] and tsunamis [7], to icebergs [8,9], avalanches [10], earthquakes [11] and explosions [12]) betray a deep concern: despite the early warnings, we appear to have been caught unprepared, and the resulting torrent of information has all but

Abbreviations used: BJ, Biochemical Journal; COHSE, Conceptual Open Hypermedia Services Environment; DOI, Digital Object Identifier; GO, Gene Ontology; GPCR, G protein-coupled receptor; HTML, HyperText Mark-up Language; IUPAC, International Union of Pure and Applied Chemistry; NTD, Neglected Tropical Diseases; OBO, Open Biomedical Ontologies; PDB, Protein Data Bank; PDF, Portable Document Format; PLoS, Public Library of Science; PMC, PubMed Central; PTM, post-translational modification; RSC, Royal Society of Chemistry; SDA, Structured Digital Abstract; STM, Scientific, Technical and Medical; UD, Utopia Documents; XML, eXtensible Mark-up Language; XMP, eXtensible Metadata Platform.

1 To whom correspondence should be addressed (email teresa.k.attwood@manchester.ac.uk).
burst our databases [13,14]. Desperate as things may seem, this is probably just a prelude to further troubles ahead, with ‘desk-top sequencing’ becoming a reality, and the latest machines delivering terabytes of data per hour. Faced with this onslaught, standard laboratory information-management systems will be unable to cope, a situation that has been likened to “taking a drink from a fire hose” [5].

Beyond the information-management headaches [8] and nightmares [15], however, lies a deeper problem. Merely increasing the amounts of information we collect does not in itself bestow an increase in knowledge. For information to be usable, it must be stored and organized in ways that allow us to access it, to analyse it, to annotate it and to relate it to other information; only then can we begin to understand what it means; only with the acquisition of meaning do we acquire knowledge. The real problem is that we have failed to store and organize 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 exhausting, frustrating, stressful [7] and increasingly costly experiences.

Let’s consider, for a moment, an activity for which these problems have become especially acute – the annotation of biological data for deposition in a database. There are now probably thousands of bio-databases around the world. One of the best known of these is Swiss-Prot [16], the manually annotated component of UniProtKB [17]. By contrast with UniProtKB, which currently contains more than 9 million entries, Swiss-Prot will soon contain 500,000 protein sequences, of which around half have been annotated by a team of curators that has devoted 600 person years to the task over a 23 year period [18] – an incredible human effort. They achieved this by reading thousands of articles and visiting hundreds of other databases, and carefully distilling out Swiss-Prot-relevant facts. The difficulties faced by the curators are legion: with something like 25,000 (increasingly specialist [19]) peer-reviewed journals publishing around 2.5 million papers each year, in the life sciences alone this effectively equates to two new papers appearing in Medline each minute [20] (see Figure 1). It is consequently both impossible to keep up with developments, and progressively more difficult either to find pertinent papers or to locate new facts within them. Each newly published paper is thus now cast adrift and essentially lost at sea. Little wonder that Bairoch should lament, “It is quite depressing to think that we are spending millions in grants for people to perform experiments, produce new knowledge, hide this knowledge in an often badly written text and then spend some more millions trying to second guess what the authors really did and found” [18].

The tasks of curators would not be quite so daunting were it possible to connect easily from articles to their underlying data-sets. True, supplementary data are more commonly being made available with publications, but this is usually a supporting subset rather than all the experimental data: journals simply do not have the capacity to archive all research data described in the articles they publish, and universities are only now beginning to consider the practicalities of how they might undertake this task themselves. For now, then, navigating between data and published descriptions of these data remains a formidable challenge, because the data are arriving in an “unorganized, uncontrolled and incoherent cacophony . . . None of it is easily related, none of it comes with any organizational methodology . . . [and the data are being] produced at greater and greater speed . . . Faster and faster, more and more and more”; and the truth is, without structure, data are mere babble [7].

The crux of the problem is the lack of organizational principles. The failure of online databases to interoperate seamlessly with each other, and with the literature, is ultimately a matter of standards, or lack of them [21,22]. Online databases, and online journals, were designed to be accessed by humans, not by machines; but the proliferation of databases and journals now makes the need for efficient machine-access imperative. If databases had standard interfaces and standard methods for scripts to access their contents, many of the problems of gathering and integrating information from diverse sources would evaporate [23].

On the other hand, contributing to the problems is the state of the literature itself. In the wake of organism-specific gene-naming cultures, the post-genomic literature descended into nomenclature chaos; faced with the task of rationalizing gene names across organisms, the amusement value of names like ken, T-shirt, hedgehog, cap ‘n’ collar, and so on, palls. It is precisely this kind of mess that spurred projects to develop meaningful ontologies [24–31] to help standardize how we describe biological entities. Coupled with standard, structured approaches for marking up journal articles, the fruits of these painstaking endeavours could, in future, position us to link articles not only to each other, but also to databases and other online resources [11]. The importance of being earnest in our approaches to such problems, in the way we write about our data, is crucial if we are to make sense of the complexities [32]. Without such approaches, our literature is in danger of giving way to yet more of what Kerr has described as “touchy-feely text and psychobabble” [33].

It is clear that scientific articles could become much better conduits for the publication of research data [34,35]. Indeed, it has been argued that the distinction between an online paper and a database is already diminishing [36]. Nevertheless, much more needs to be done to make the data contained in research articles more machine-readable, a sentiment endorsed in the 2007 Brussels declaration on Scientific, Technical and Medical (STM) publishing (http://www.stm-assoc.org/public_affairs_brussels_declaration.php), which commits STM publishers to “change and innovation that will make science more effective.” This commitment will challenge publishers to embrace all the potential of modern Web(2.0) technologies, including blogs, wikis, Really Simple Syndication (RSS) feeds and so on [11,37–39], ultimately to provide more lively, interactive access to their content, and to save our journals from becoming incurably dull.
[40]. “The time has come,” as O’Donnell asserts, “to grab back our ‘literature’ and for editors to restore journals to their readers” [41]!

In the present review, we examine some recent initiatives to make published biomedical texts more machine-readable, and hence more dynamic, interesting and informative. In particular, we outline a variety of projects involving academic-journal collaborations: these are the first seedlings of much-needed community–publisher engagement, which we hope will blossom into more and wider alliances to tackle the very difficult problems involved. We also introduce a new development with Portland Press Limited, the so-called Semantic Biochemical Journal (BJ) experiment, illustrating how much can be achieved through appropriate collaboration, yet recognizing how much remains to be done. Reflecting on the considerable opportunities that lie ahead, we conclude with an international call to arms to embrace the future of digital publishing together.

GRABBING BACK OUR LITERATURE

In the sections that follow, we examine a variety of projects that challenge us to change the way we think about the scholarly literature, and to embrace new ways of interacting with it. These projects promise to transform how we access and extract the knowledge embedded in scientific articles. We discuss the advances that have been made, some of the problems these approaches help to solve, and the obstacles to progress that still exist.

Ontologies for biomedical literature

To formalize how we describe biological entities and convert published biomedical information into machine-readable data, accessible to search engines and to algorithmic processing, several groups have developed ontologies and controlled vocabularies for biomedical texts: these are now numerous, but include, for example, the RNA Ontology [26], the Sequence Ontology [25], the Cell Ontology [27], the Systems Biology Ontology [30] and, probably the best known, the Gene Ontology (GO) [24]. To bring order to these proliferating initiatives, and better support biomedical data annotation and integration, the Open Biomedical Ontologies (OBO) Foundry was set up to unify these diverse resources [28].

Building on these endeavours, various Web-based tools have been developed to render such machine-readable information more generally useful to the community. One of the broadest of these, COHSE (Conceptual Open Hypermedia Services Environment) runs as a portlet: this allows users to select an ontology, then adds relevant hyperlinks to target pages (see Figure 2), matching the ontology terms to those pages and propagating links to further pages [42,43]. Extensions to COHSE (including text-mining components to improve linking opportunities, and integration of workflows and services as possible link targets) are planned, but the current public version provides relatively limited functionality and is not yet sufficiently mature for some practical applications – for instance, it does not...
Figure 3 Illustration of Prospect mark-up in part of a Molecular BioSystems article

Terms found in the source ontologies, which may be toggled on or off via the greyed-out Tools and Resources toolbar to the right of the page, are highlighted in different colours: e.g. pink highlights denote compound terms, which link out to diagrams of their structures, synonyms, Simplified Molecular Input Line Entry Specification (SMILES) nomenclature, etc.; yellow highlights link to definitions from the Gold Book; blue highlights are biomedical terms and green highlights are chemical terms, both of which link out to relevant definitions, synonyms and ontologies. Fragments of linked webpages are overlaid on this Figure as ‘callouts’. (http://www.rsc.org/Publishing/Journals/ProjectProspect/). The extract from Molecular BioSystems ([48]; Koenigs, M.B., Richardson, E.A. and Dube, D.H. (2009) Metabolic profiling of Helicobacter pylori glycosylation. Volume 5, 909–912; http://dx.doi.org/10.1039/b902178g) has been reproduced by permission of The Royal Society of Chemistry.

allow direct navigation to specific data (such as biomolecular sequences) via its life-science ontologies [44]. The long-term vision of projects like this, and of the OBO Foundry in particular, is that all biomedical research data should ultimately form a single, consistent, machine-accessible whole (see also http://www.bio2rdf.org). Realizing this goal will not be easy: the challenge will be to provide sufficient flexibility for scientific advances to flourish within a sufficiently robust and principled framework for unification to be feasible.

Blogs for biomedical science

In recent years, ‘web logging’ (blogging) has emerged as a widespread social phenomenon. With > 100 million blogs on the Internet, and a new blog appearing every half second, blogging is now recognized as a vehicle of unprecedented power for information dissemination [46]. The scientific community is in the process of catching up with these developments, and there are now ~ 1200 blogs dedicated to scientists and their conversations.

Against this background, publishers have begun to appreciate the potential of blogs to engage more interactively with their readers, to promote discussion of their journal content and to stimulate peer review. Consequently, many of the major journals now have their own blogs; some have several. Notable here is the series of blogs from Nature Publishing Group, including: Nascent, Indigenus, Methagora, Nautilus, Spoonful of Medicine, The Sceptical Chymist, The Great Beyond, The Niche, The Seven Stones and others.

The proliferation of the Nature blogs is a testament to the popularity of this medium for discussing and advancing science. Some journal blogs are doing less well, however, and attract little or no traffic. With so many to choose from, the problem is partly in knowing that a particular blog exists and partly in knowing which are the most worthwhile to read; other barriers to take-up include the activation energy required to visit individual blogs on a regular basis, and the disruption this causes to researchers’ work patterns.

Nevertheless, blogging has clearly captured the imaginations of hundreds of scientists and, as the ‘blogosphere’ becomes noisier, it is likely to need increasingly artful hooks to seduce the research community to engage with it more meaningfully.

Project Prospect and the Royal Society of Chemistry

With Project Prospect, the Royal Society of Chemistry (RSC) has played a pioneering role in introducing meaning (semantics) to published content [47] and creating computer-readable chemistry.
Some of their journals, such as Organic and Biomolecular Chemistry and Molecular BioSystems, now offer enhanced HyperText Mark-up Language (HTML) versions of articles, marked up by their editors using the Prospect software. Accessed via a tool-box (the ghostly silhouette on the right-hand side of the article in Figure 3), features available for mark-up include compound names, bio- and chemical-ontology terms, and terms from the Gold Book [the International Union of Pure and Applied Chemistry (IUPAC) Compendium of Chemical Terminology] – marked-up terms appear as colour-coded highlights within the text. Clicking on the highlights provides relevant definitions from the Gold Book or from the Gene [24], Cell [27] and Sequence Ontologies [25], together with GO identifiers and InChI (IUPAC International Chemical Identifier) codes, lists of other RSC articles that reference these terms, synonym lists, links to structural formulae, patent information and so on.

Prospect mark-up significantly enriches RSC journal articles, making navigation to additional information trivial and increasing the appeal to readers, but this is just a start. More work is needed to extend the scope of the work to other subject areas, to include more extensive linking (e.g. to databases and experimental data) and to add other Prospect services. The system is currently limited to HTML, and it will be interesting to see how readily the project principles can be extended to the rest of RSC’s journals and to its [Portable Document Format (PDF)] e-book collection.

The ChemSpider Journal of Chemistry

The ChemSpider Journal of Chemistry is another experiment set up to demonstrate the added value that Web technologies can offer in terms of enriching published information. The Journal spans a range of chemistry-related subjects, including chemical biology, chemo-informatics and molecular modelling. Its articles are marked up using the Chemistry Markup And Nomenclature Transformation Integrated System, ChemMantis. ChemMantis identifies and extracts chemical names, converting them into chemical structures using name-to-structure conversion algorithms and dictionary look-ups in the ChemSpider chemistry database (which provides access to almost 21.5 million unique chemical entities); it also marks up a range of other chemical entities, including chemical families, groups, elements and reaction types; where appropriate, the terms are linked to their Wikipedia definitions (see Figure 4). A facility is also provided to allow readers to comment on individual articles.

The current ChemSpider Journal of Chemistry website lists a dozen articles, the majority of which were published in March 2009. No further papers have appeared since the acquisition of ChemSpider by the RSC in May 2009; the status of this particular online experiment therefore appears uncertain.
Mutagenesis performed to identify ephrin receptor binding sites revealed four regions (residues 11 to 15, 44 to 51, 100 to 108, and 147 to 151) important for the activation of receptor signaling [53]. With the exception of residues 149 to 152, functionally flexible predictions were made outside of these binding hot spots. This shows that our predictors are not trained to predict binding sites, but rather regions where flexibility accommodates different conformational states.

Comparison to Protein Disorder Predictors

Several protein disorder predictors were compared to see which best predicted disorder hotspots. Disorder predictors that considered temperature factors or missing residues in crystal structures, such as DISORDER [54], are a neural network trained for the prediction of disorder motifs in proteins. To take advantage of information embedded in algorithms that, respectively, use running properties of hydrophobicity and net charge. IUPRED [65] uses concepts of pairwise interaction potentials observed in globular proteins to make assignments for each residue. Finally, NORSF [66] assesses regions based on low confidence predictions for secondary structural elements.

Some overlaps are expected with disorder predictions because FFRs may be disorder dependent on the conformational state of the protein. Otherwise, we expect little correlation since disorder predictors generally aim to identify structural disorder and regions with low propensity to form ordered units. Potential functional roles were not considered in their design, although these regions are suggested to be important for protein-protein recognition after examining positively classified sequences [65]. With the exception of the arc repressor where predicted results exhibited significant overlap, Wiggle and Wiggle200 have been found to target regions that were not otherwise identified by disorder predictors.

For arc repressor (SAB), the hinge region connects all residues involving recognition loop, catalytic loop, and so forth (Figure 9). The predicted flexible regions for the arc repressor were identified by disorder predictors. While Wiggle predictors did not move for P3 loop ends (Figure 9), it identified the minor groove locations.

Current disorder prediction tools failed to identify these hydrophobicity and net charge contributions to amino acid recognition (RCNN).

The FEBS Letters experiment

The FEBS Letters experiment was a pilot collaborative study involving the journal editors, an initial small group of authors and the curators of the MINT interaction database [50]. The broad aim here was to integrate data published in scientific articles with information stored in databases [51], but with a pragmatic focus on protein–protein interactions and post-translational modifications (PTMs); making all published biological data instantly machine-readable was clearly not possible [52]. The experiment hinged on adopting the concept of the Structured Digital Abstract (SDA). The idea of the SDA is simply to provide a mechanism for capturing an article’s key facts in a machine-readable, eXtensible Markup Language (XML)-coded summary, in order to make them accessible to text-mining tools [21].

For the purpose of this experiment, key protein interaction and PTM data were collected from authors via an Excel spreadsheet and structured so as to include: descriptions of the nature of the experimental evidence; characteristics of the participating protein partners; details of the biological roles of proteins in the interactions; expression levels; the PTMs required for interaction, or that result from it; unique protein identifiers with links to MINT and UniProtKB [17]; definitions drawn from the Human Proteome Organization (HUPO) Proteomics Standards Initiative’s Molecular Interaction Controlled Vocabulary; and so on [53] – a typical SDA is shown in Figure 5. By the nature of the project, the parameters of the experiment were well-defined, and most of the captured relationships point to MINT entries; were it to be widely adopted, however, the system has been designed to readily generalize to other databases of protein interactions or other biological relationships.

The experience of handling the first seven manuscripts was reported in 2008 [53]. The authors of only five of these papers chose to participate, most of whom had relatively few problems with the SDA and required minimal assistance; but one author had major difficulties and needed substantial help from the MINT curators to complete the spreadsheet. During the next 10 months, to February 2009, SDAs appeared in 90 FEBS Letters papers [34], pointing to a rather slow uptake within the community. Ultimately, if the experiment were judged to have been successful, it was intended that these SDAs would form an integral part of Medline abstracts. However, this development has yet to materialize, and the future of SDAs is unclear.

PubMed Central and BioLit

BioLit is a suite of open-source tools designed to integrate open literature with biological databases [55]. As a proof-of-concept, the tools have been implemented using a subset of papers from PubMed Central (PMC), structural data from the Protein Data Bank (PDB) [56], and terms from various biomedical ontologies.

BioLit allows full-text (or excerpts of full-text) articles to be included directly in a database, and permits metadata (PDB identifiers and GO terms) to be added to such articles. The system works by mining the full text for terms of interest, indexing the terms identified, and delivering them as machine-readable XML-based article files. To make these files human-readable, a Web-based article viewer displays the original text with the...
Rescuing knowledge lost in literature and data

Impact of Environment and Social Gradient on Leptospira Infection in Urban Slums

Renato B. Reis 1,2, Guilherme S. Ribeiro 1,2, Ridvalda D. M. Felzemburgh 1, Francisco S. Santana 1,2, Blamir Mohr 1, Atirid X. T. O. Melendez 1, Adriana Queiroz 1, Andréa C. Santos 1,3, Romy R. Ravines 1, Wagner S. Tasaninn 3,4, Marília S. Carvalho 5, Miriam G. Reis, 1, Albert I. Ko 1,5

Abstract

Background

Leptospirosis has become an urban health problem as slum settlements have expanded with leaps and bounds at the hand of lack of population-based information in urban slums. To estimate the prevalence of leptospiral infection and identify risk factors for infection,

Methods and Findings

We performed a community-based survey of 3,171 slum residents from Salvador, back in 2008. The prevalence of leptospirosis antibodies was 15.4% (95% confidence interval [CI] 14.1, 16.7), with significance levels (p < 0.05). The presence of titres of 1:320 and above were independent risk factors. An increase of 0.18% in infection risk.

Conclusions

Deficiencies in the sanitation infrastructure where slum inhabitants reside may help in the control of leptospiral infections. The elimination of these factors may need to address the social factors that prevent improving sanitation.

Figure 7 The PLoS NTD article marked up using the system developed by Shotton et al. [34]

Users may select from the coloured tabs at the top of the page to reveal entities of interest in the text: here, the protein (purple), disease (red), habitat (green) and organism (blue) tabs have been chosen. Organism terms are linked to uBio, a community initiative to create a comprehensive catalogue of the names of all (living and once-living) organisms (e.g. overlaid here as a ‘callout’). Reproduced from [58]; Reis, R.B., Ribeiro, G.S., Felzemburgh, R.D., Santana, F.S., Mohr, I., Melendez, A.X., Queiroz, A., Santos, A.C., Ravines, R.R., Tassinari, W.S. et al. (2008) Impact of environment and social gradient on Leptospira infection in urban slums. PLoS Neglected Tropical Diseases 2, e228.

Public Library of Science (PLoS) Neglected Tropical Diseases (NTD)

In another interesting adventure in semantic publishing, Shotton et al. [34] chose an article in PLoS NTD as a target for enrichment. The criteria for selecting this particular article included the fact that it contained various different data types (geospatial data, disease-incidence data, serological-assay results, and so on) presented in a variety of formats (maps, bar charts, scatter plots, etc.); moreover, it was available in an XML format, published under a Creative Commons License – the article could therefore be modified and re-published.

The semantic enhancements added to the article include: live Digital Object Identifiers (DOIs) and hyperlinks; mark-up of textual terms (disease, habitat, organism, protein, taxon, etc.), with links to external information resources (see Figure 7); interactive figures; a re-orderable reference list; a document summary, with a study summary, tag cloud and citation analysis; mouse-over boxes for displaying the key supporting statements from a cited reference; and tag trees for bringing together semantically related terms. Augmenting these enhancements are both downloadable spreadsheets containing data from the tables and figures, enriched with provenance information, and examples of ‘mashups’ with data from other articles and Google Maps. In addition, a ‘citation typing’ ontology was implemented to allow compilation of machine-readable metadata relating both to the article and to its cited references [29].
The enhancements described in this study are platform- and browser-dependent and are confined to a single article. However, in the hope of stimulating more general take-up of their ideas, the authors assert that what they achieved was not “rocket science”, but was accomplished using standard mark-up languages, ontologies, style sheets, programmatic interfaces, and so on. They recognize, nevertheless, that their exemplar was manually intensive, and that to bring the approaches they espouse into mainstream publishing protocols will require greater degrees of automation.

Elsevier Grand Challenge

In 2008, to stimulate further efforts to improve the way scientific information is communicated and used, Elsevier announced its Grand Challenge of Knowledge Enhancement in the Life Sciences. The focus of the contest was to develop tools for semantic annotation of journals and text-based databases, to improve access to, and sharing of, the knowledge contained within them: in short, to change the way that science is published.

The winners of the contest developed a tool (Reflect) that addresses the routine need of life scientists to be able both to jump from gene or protein names to their molecular sequences, and to understand more about particular genes, proteins or small molecules encountered in the literature [44]. With a single mouse click, Reflect tags such entities when they occur in webpages; it does this by drawing on a large, consolidated dictionary (containing 4.3 million small molecules and >1.5 million proteins from 373 organisms) that links names and synonyms to source databases. When clicked on, the tagged items invoke pop-ups (see Figure 8) displaying brief summaries of key features (domain structures, small-molecule structures, interaction partners, etc.), and allow navigation to core biological databases like UniProtKB.

Reflect was optimized for speed rather than accuracy—inevitably, therefore, there are errors in the tagging. As part of their ongoing system developments, the authors plan to address this problem by implementing mechanisms for community-based, collaborative editing of some of the information provided by Reflect, and especially to allow correction of some of its errors.

The system is currently accessible to users directly via the Web, and as Firefox or Internet Explorer plug-ins; in future, programmatic access via Web services might also be possible, obviating the need for users to install browser plug-ins.

Liquid Publications

A rather different slant on the problem of dissemination and re-use of scientific knowledge is offered by the Liquid Publication Project, a European initiative partnered by Springer Verlag [60]. The intention here is for publications to become fluid entities, created in a collaborative and evolutionary fashion over time, in much the same way as open-source software is developed;
there are also parallels here with successful social/collaborative annotation models such as Wikipedia.

This project aims to exploit emerging Web technologies to spur a transition away from traditional ‘solid’ scientific papers (which crystallize fragments of scientific knowledge at a point in time) to Liquid Publications, which may adopt multiple shapes, evolve continuously and are enriched by multiple sources. The idea is to promote early circulation of innovative ideas, to optimize the processes by which researchers create, assess and disseminate knowledge, and to stimulate publishers to offer more advanced services (including the maintenance of scientific social networks, automatic notification of new contributions in certain areas, social bookmarking, collaborative authoring, blogging and reviewing) – to become “the yahoo, flickr, digg and delicious of the publication world” [60].

It is hard to assess how far the project has progressed towards achieving these goals. By definition, there is no current solid publication summarizing the work; and the Liquid Document available on the project website (version 2.3), itself an evolution of a previous paper (which argues “why the current publication and review model is killing research and wasting your money” [61]), was last updated in 2007. Like water, therefore, the impact of Liquid Publications is difficult to grasp.

Are we there yet?

Although the initiatives outlined above may differ slightly in their specific aims, they are nevertheless reflections of the same overall aspiration – to make the data and knowledge sequestered in the literature more readily accessible and re-usable. The results, to date, are encouraging, and it is interesting to see the common themes that have emerged: most are HTML- or XML-based, providing hyperlinks to external websites and term definitions from relevant ontologies via colour-coded textual highlights. But these are only first steps towards much more far-reaching possibilities, and new ideas and new tools are clearly still needed.

Lynch, for example, imagines a future in which there exists a wide range of specialized visualization tools for various forms of structured data [37]. It would be useful, he suggests, to be able to toggle between a rendered image and its underlying data-set, or between a published table of numerical values and their graphical representation, perhaps like the scenario shown in Figure 9.

In a similar state of reverie, Bourne has a vision in which journals provide software for visualizing and interpreting their published content, obviating the need for specialized knowledge in handling esoteric tools; he envisages such software ultimately allowing various forms of basic analysis (simple statistical tests, principal-component analysis, and so on), making new levels of comprehension possible [36,63]. More specifically, he asks us to imagine reading a description of a molecule’s active site in a paper, being instantly able to access its atomic co-ordinates, and to imagine reading a description of a molecule’s active site in a paper, being instantly able to access its atomic co-ordinates, and thence to explore the interactions described in the paper, perhaps something like the scenario illustrated in Figure 10?

These concrete initiatives and wistful imaginings bear witness to the yearning within the community for more productive ways of interacting with the literature. In 2005, Bourne asked, “Is the technology available to support the next steps and is the scientific community ready for such a change?” [36]. An important step forward would be to assign standard identifiers, not only to papers, as we do now, but also to their authors [65] and to the biological objects the papers describe. An outcome of such an approach would be the ability to find all papers that reference, say, a particular sequence motif [36,63]. More specifically, he asks us to imagine reading a description of a molecule’s active site in a paper, being instantly able to access its atomic co-ordinates, and thence to explore the interactions described in the paper, perhaps something like the scenario illustrated in Figure 10?

Figure 9

Lynch imagines being able to toggle between a published table of numerical values and their graphical representation

why publications have benefited so little from the opportunities offered by such infrastructure is probably cultural [38]: simply, the community has grown up with static manuscripts, and most electronic articles are still delivered in unexpressive, semantically limited forms, like PDF or HTML [37], which some authors accuse of impeding the progress of scholarship.

To gain the most from electronic articles, and especially from dormant document archives, semantic mark-up of content is clearly necessary. But retrospective addition of semantics to legacy data is complex, labour-intensive and costly. A balance must therefore be found between the degree of automation it is possible to introduce to the process, and the degree of cultural change it is reasonable to expect in a research community that has not hitherto considered the relationship between data and published articles, and has hence not been concerned about providing the semantic context necessary to unite them. In the long-run, it is to be hoped that the benefits of semantic mark-up, and the availability of the right tools, will together help to seed this much-needed cultural change: compare and contrast, for example, the pages shown in Figure 12.

What is clear is that new technologies will emerge (and indeed, are already emerging) to promote a fundamental shift away from how scholarly communication currently works [69]. A key driver of this change will be realization of the benefits that accrue from having more explicit links between articles and the data and concepts they describe [70]. Processes that will particularly profit from such links are peer review and the dissemination of (reliable) knowledge. Were a paper to become an interactive interface to its underlying data, it could, for example, facilitate further research across multiple articles and databases, and lead more easily to the discovery of errors; combined with suitable social technologies for community commentary, a published paper could at the same time act as its own self-correcting record. This would be an especially powerful development, as the extent to which peer review of an article extends to its underlying data is generally not at all clear, and current mechanisms for data correction, updating and maintenance are not synchronized with those for managing the literature [37]. Thus, as Antezana points out, reported ‘facts’ may be incomplete, incorrect or simply false, and new knowledge may refute ‘accepted’ information [10]. Unfortunately, however, we have no way of knowing what the error rates in the literature or in biological databases actually are, or indeed what are the rates of propagation of those errors between databases and papers, and vice versa. The ramifications of new tools and technologies that could support the discovery of errors and inconsistencies, which could allow us to track and to consistently record the evolution of the current state of our knowledge, are therefore potentially profound. Consider, for a moment, the example illustrated in Figure 13.

Sharing knowledge is at the philosophical root of scientific scholarship, and our publishing systems were designed to help us do this. But Wilbanks asserts that, in the aftermath of the “earthquake of modern information and communication technologies”, we are not sharing information efficiently: we need infrastructures that facilitate knowledge sharing and integration, rather than mere Web publishing [11]. He bemoans the lack of standardized mechanisms to connect knowledge, which means that, “we can’t begin to integrate articles with databases” not least because “the actors in the articles (the genes, proteins, cells and diseases) are described in hundreds of databases.” Solving this
will not be easy; much of it, he warns, will be “very, very hard. But the current system is simply not working” [11].

While there is a sobering degree of truth in these comments, we believe that growing awareness of the issues, coupled with a community-wide desire for progress, has stimulated some promising developments. Let’s take a closer look, in the next section, at a new initiative from Portland Press Limited.

The Semantic Biochemical Journal experiment

The Semantic Biochemical Journal (BJ) experiment was a collaborative project involving the BJ editorial staff and the developers of Utopia [73], a software suite that semantically integrates visualization and data-analysis tools with document-reading and document-management utilities. The principal aim of the project was to make the content of BJ electronic publications and supplementary data richer and more accessible. To achieve this, Utopia was integrated with in-house editorial and document-management workflows, allowing copy editors to mark up content prior to publication; this removed the mark-up burden from submitting authors, and ensured rigour and consistency from the outset.

The UD reader works by creating unique fingerprints of document contents as they are rendered onscreen, identifying key typographical and bibliometric features (authors, figures, references and so on). But the real innovation lies in being able to turn static images, tables and text into objects that can be linked, annotated, visualized and analysed interactively. The additional data are overlaid rather than embedded in the documents, leaving their provenance and integrity intact; this means that features can

Figure 11  Bourne imagines being able to find all papers that reference a particular sequence motif described in a paper

In this 2008 Biophysical Chemistry article [68], Illingworth et al. describe the GXXG motif characteristic of the LanC (lanthionine synthetase C)-like proteins (a), and also reference them elsewhere in the literature (b), including their appearance in main cycle, whose three-dimensional structure was determined by Li et al. [67], and in the putative G protein-coupled receptor (GPCR) GCR2 [68] (c). For readers viewing this article using UD, to bring life to this image and visualize the GXXG motifs, click on the UD logo. Reproduced from Illingworth, C.J.R., Parkes, K.E., Snell, C.R., Mullineaux, P.M. and Reynolds, C.A. (2008) Criteria for confirming sequence periodicity identified by Fourier transform analysis: application to GCR2, a candidate plant GPCR? Biophysical Chemistry 133, 28–35, Copyright (2008), with permission from Elsevier; and from Gao, Y., Zeng, Q., Guo, J., Cheng, J., Ellis, B. E. and Chen, J.-G. (2007) Genetic characterization reveals no role for the reported ABA receptor. GCR2, in ABA control of seed germination and early seedling development in Arabidopsis. The Plant Journal 52, 1001–1013 with permission from Wiley-Blackwell.

© 2009 The Author(s)
The author(s) has/have asserted their right to be identified as the author(s) of this work in accordance with the Copyright, Designs and Patents Act 1988. This article has been made available in this electronic format by Portland Press Limited (www.portlandpress.co.uk) and is copyright the author(s). It may be cited, reproduced (when necessary for personal use or in full) and disseminated in hard copy format, for educational or non-commercial purposes, on the conditions that the Portland Press Limited logo and the author(s) name(s) are displayed or cited on any copy and the author(s) and Portland Press Limited are referred to as the original publisher. The full terms and conditions of use are available online at http://www.portlandpress.co.uk/page/termsandconditions.html.
be reliably associated with any version of a file, even one that has lain unread on a laptop for many years. In this way, the electronic document is transformed from a digital facsimile of its printed counterpart into a gateway to related knowledge, providing the research community with focused interactive access to analysis tools, external resources and the literature.

For the purposes of this experiment, all the papers in the current issue of the BJ have been marked-up by the Journal’s copy editors (as will subsequent issues). For practical reasons, features relating to protein sequence and structure analysis have been the main targets, because this was the functionality built into the original Utopia toolkit [74]. At the time of writing, the additional mark-up provides: links from the text to external websites (including major databases such as UniProtKB [17], PDB [56] and InterPro [75]); term definitions from ontologies and controlled vocabularies; extra embedded data and materials (including images, videos and so on); and links to interactive tools for sequence alignment and three-dimensional molecular visualization. Utopia does not itself provide any domain-specific functionality for processing or analysing data, but relies on external services; these are accessed via plug-ins whose appearance in the software interface is mediated by a ‘semantic core’ (the core can be customized to any subject area by incorporating the relevant discipline-specific ontologies).

Reliance on external Web services is a strength of the system, in the sense that it allows greater flexibility for customizing the functionality of the software (obviating the need for the developers to second-guess all current and future potential user needs); it may also be a weakness, however, because when those external services become unavailable (e.g. owing to routine maintenance or faulty operation of some kind), their functionality also becomes unavailable to Utopia. Such issues (which afflict all systems that rely on Web services, not just Utopia) are mitigated to some extent by the establishment of a Web-service registry, which systematically monitors and provides feedback on the status of its registered services [76].

As with other projects outlined in the present review, UD is still at an early stage of development and there is much more work to be done. As the system is readily customizable, we plan to extend its scope, for example, to systems and chemical biology, and to the medical and health sciences, as many of the requisite chemical, systems biology, biomedical, disease and anatomy ontologies are already in place and accessible via the OBO Foundry.

Another challenge concerns a feature of UD that allows readers to append notes or comments to articles, and how this is developed in future. There are at least three different scenarios to consider here: (i) a reader might wish to make a ‘note to self’ in the margin, for future reference; (ii) a reviewer might wish to make several marginal notes, possibly to be shared with other reviewers and journal editorial staff; and (iii) a reader might wish to append notes to be shared with all subsequent readers of the article (e.g. because the paper represents an exciting breakthrough or because
it contains an error) without having to establish a personal blog or to write a formal Letter to the Editor. These scenarios involve different security issues, and work will be needed to investigate and establish appropriate ‘webs of trust’.

For now, to gain further insights into the status of the Semantic Biochemical Journal experiment, we encourage readers to view the PDFs of other articles in this BJ issue (and subsequent issues) through the animating lens of UD.

**DISCUSSION**

The PDF debate

In recent years, the literature has seen the value of PDF as a mechanism for digitizing the printed page rather hotly contested. PDF, although easy for humans to read, is not regarded as an efficient medium for gathering information, nor for sharing, integrating and interacting with knowledge; it is considered...
semantically limited by comparison with XML, and antithetical to the spirit of the Web [11,34,35,37,77].

Notwithstanding the critics, PDFs are still the dominant means of dissemination of scientific papers. For the human reader, they are like ‘electronic paper’ – they generally inherit the standard typesetting conventions of the original journal and hence feel ‘natural’ to read. People also like to have their own copies of documents, which can be read offline, with the added comfort of knowing that the PDF won’t disappear even if its originating website does.

Adobe’s PDF has therefore become the de facto standard for document dissemination (although technically a proprietary format, it is sufficiently open to be supported by all platforms). It supports basic annotation and hyperlinking (within a document, and to external sources), and also allows inclusion of metadata. Interestingly, earlier this year, the Charlesworth Group, working with Nature Publishing Group, completed a project to incorporate Xtensible Metadata Platform (XMP) metadata within Nature’s online PDFs (the metadata include article titles, author details, keywords, images, DOIs and so on; http://www.nature.com/press_releases/charlesworth.html). This has the dual advantage of presenting scholarly information both in a human-readable form and in a format accessible to software applications. However, although all new Nature research articles will contain embedded XMP metadata as they are published, there are no plans for retrospective mark-up of the Nature archives. Moreover, as the metadata are embedded at the point of publication, they are effectively as fixed as the original PDF and are unavailable for future modification. This is in contrast with the approach taken with UD, which vivifies the static PDF document by overlaying dynamic, customizable metadata, in turn adding evolvable, interactive content to the underlying file. As mentioned above, this system also yields the potential for sharing community comments and annotations on any document (past and present), storing them on a common server and making them accessible to future semantic Web applications.

Clearly, the technology to add value to PDF documents, whether with links to websites, links to interactive analysis tools or to live online commentaries or blogs, is with us now; the time is therefore ripe to exploit it. On a technical level, the ultimate goal is effective ‘knowledge management’ [11,78]; on a human level, it is to deliver to the research community a tangible way not only to understand but to interact with scholarly data. As such, digital documents are like ‘electronic paper’ – they generally inherit the standard typesetting conventions of the original journal and hence feel ‘natural’ to read. People also like to have their own copies of documents, which can be read offline, with the added comfort of knowing that the PDF won’t disappear even if its originating website does.

Achievements and challenges

The projects outlined in the previous sections bear witness to the growing momentum, fuelled by community pressure, to tackle these issues, to get more out of digital documents and especially to facilitate access to underlying research data. The projects differ a little in scale and focus; all are, in some sense, experimental. They therefore present opportunities to learn what has worked best, what hasn’t worked so well, and why. They also serve as valuable models, revealing what more needs to be done and what obstacles still exist before we can realize the goal of truly integrated literature and research data.

The RSC have taken pioneering steps with Prospect and ChemSpider. The content mark-up they have achieved looks set to become richer and wider in scope, and will doubtless extend to more of their own published content over time. The application of BioLiTo a subset of PMC articles also looks promising but, as with the FEBS Letters experiment, in its original implementation it links only to a single database – to be optimally useful, these initiatives would need to become richer and wider in scope, and will doubtless extend to more of their own published content over time. The application of BioLiTo a subset of PMC articles also looks promising but, as with the FEBS Letters experiment, in its original implementation it links only to a single database – to be optimally useful, these initiatives would need to embrace many more biomedical tools and resources.

Shotton’s project [34] with PLoS NTD was, in some ways, more ambitious in scope. Despite being limited to a single article, the semantic enhancement provided was found to be a labour-intensive exercise. To render their approach more cost-effective, Shotton recognized the need for greater levels of automation, and he pointed to tools like Reflect to help ease manual mark-up burdens. However, Reflect and similar tools that use named-entity recognition are error prone [79,80]. For now, then, a balance has to be found between the degree of automation necessary to make semantic enrichment feasible and the degree of manual intervention necessary to ensure rigour and consistency of mark-up. As a trivial illustration, look more closely at the definition Reflect gives to OMP in Figure 8 – Olfactory Marker Protein. Ironically, directly above the pop-up, the correct expansion of the acronym is given in the original text – Outer Membrane Protein. What is simple to spot by eye is much harder to achieve computationally. Issues of this type are the scourge of text-miners, and there are no perfect solutions.

Why semantic mark-up is hard

Most of the projects mentioned in the present review have exploited fairly traditional text-mining methods, in conjunction with controlled vocabularies and ontologies, to provide a springboard from marked-up entities within published texts to external webpages. As such, they come with all the limitations of current text-mining tools in terms of precision; they also bring an overhead to readers in terms of having both to identify and to correct errors – having to know that an error really is an error is perhaps one of the biggest pitfalls. Moreover, as Fink and Bourne point out for BioLiTo, the mark-up these approaches provide is not truly semantic, in terms of inferring relationships [55]. This is partly because most electronic articles are delivered in what are considered to be fixed, semantically limited forms (PDF and HTML) [37,82], but partly also because genuine semantic mark-up is hard – it is labour intensive; it requires significant financial investment; it demands adoption of, and adherence to, common mark-up standards; and, perhaps most difficult of all, it involves cultural change.

The philosophy embodied in UD is to hide from authors and readers as much of the underlying complexity as possible, to avoid requiring them to change their existing document-reading behaviours, and to present no additional barriers to publication. But, like the other work discussed in this review, UD is also an experiment. The success of the experiment will ultimately depend on several factors, including whether the barriers to adoption are
sufficiently low; whether the approach is found to add sufficient value; whether the cost of the approach is sustainable; and whether entire communities can be galvanized to move forward and work together.

The cost of doing it

The FEBS Letters experiment involved a significant time investment on the part of journal editors, MINT curators and co-operating authors – the harder authors found it to engage with the mark-up process, the greater the burdens that fell to curators. The RSC’s experience with project Prospect was also labour intensive, involving collaboration with text-miners and the input of skilled, in-house domain-specialists, with sufficient breadth of expertise to understand XML, to edit, mine, mark-up and ‘user-friendlify’ the final results. Shotton estimates that his own experiment with one PLoS NTD article required ten person-weeks of effort (although, with the learning phase behind them, the exercise could doubtless be repeated more swiftly) [34]. Similarly, the Semantic Biochemical Journal experiment involved close collaboration with BJ editorial staff, and more than 2 person-years of technical effort to build the necessary infrastructure to make future mark-up relatively trivial. Overall then, these experiments have not been cheap.

The price of not doing it

If the cost of semantic publishing seems high, then we also need to ask, what is the price of not doing it? From the results of the experiments we have seen to date, there is clearly a need to move forward and still a great deal of scope to innovate. If we fail to move forward in a collaborative way, if we fail to engage the key players, the price will be high. We will continue to bury scientific knowledge, as we routinely do now, in static, unconnected journal articles; to sequester fragments of that knowledge in disparate databases that are largely inaccessible from journal pages; to further waste countless hours of scientists’ time either repeating experiments they didn’t know had been performed before, or worse, trying to verify facts they didn’t know had been shown to be false. In short, we will continue to fail to get the most from our literature, we will continue to fail to know what we know, and will continue to do science a considerable disservice.

What we’ve learned

It is clear from these experiments that the way ahead must involve genuine collaboration between life scientists, computer scientists, bio- and chemo-informaticians, database curators, publishers, learned societies, librarians and many others – the necessary advances in current publishing practices cannot be achieved in isolation. Although necessary proofs of principle, the problems will not be solved by linking a single database to a single article, by linking a single database to several articles, or by linking several databases to a single issue of a single journal; nor will they be solved by developing and protecting proprietary mark-up tools and ontologies. The real challenge concerns the need for interactions between all databases, all journals, and all research data, and will involve the commitment of entire communities.

The pace of progress will ultimately be determined by the extent to which the research and publishing communities can be persuaded to work together to promote new data standards and to build new, open ontologies; it will also depend on the extent to which publishers are prepared to engage with technology providers to evolve their traditional roles in scholarly communication towards knowledge-management solutions, and in turn, on the extent to which authors are prepared to evolve their habits in line with the ongoing publishing revolution.

A call to arms

Learned societies, publishers and their editorial boards are well placed to champion the standards for manuscript mark-up necessary to drive effective knowledge dissemination in future, and to garner community support for those standards. To this end, the sponsors of the International Association of Scientific, Technical and Medical Publishers and of societies such as the Biochemical Society, the International Society for Computational Biology and the newly-formed International Society for Bioiocuration would substantially help in taking the next steps forward, as would dialogues with the publishers and curators whose journals and databases have been the focus of the experiments outlined in the present review. There are likely to be many other stakeholders, with vested interests in their own domains of knowledge. It will therefore be essential to stimulate constructive discussions and collaborations among all the relevant players. The seeds of these much-needed debates could be sown, perhaps, on the various society and community discussion boards, on prominent blogs (e.g. http://blogs.bbsrc.ac.uk/), and on journal commentary pages, or placed on the agenda at International meetings. As Seringhaus and Gerstein point out [21], it’s important not to rush at this, but to consider the issues carefully. The benefit of getting it right could be a cost-efficient investment in a new type of knowledge landscape, one that better serves the needs of new millennium readers, authors and publishers – it’s a potential win, win, win situation, if we build on the foundations together.

ACKNOWLEDGEMENTS

We are grateful to Harry Mellor and Martin Humphries for introducing us to staff at Portland Press Limited. We thank Audrey McCulloch, Andy Gooden, John Day and especially Rhonda Oliver for having the courage and tenacity to support our vision and for their, at all times, patient and positive collaboration. We also thank Pauline Starley and the editorial team for their hard work in marking up the current issue of BJ.

FUNDING

The development of Utopia Documents has been supported by the European Union (EMBRACE) (grant number LHSG-CT-2004-512092); the Engineering and Physical Sciences Research Council (Doctoral Training Account); the Biotechnology and Biological Sciences Research Council (Target practice) (grant number BB/D016065/1); and Portland Press Limited (The Semantic Biochemical Journal project).

REFERENCES


© 2009 The Author(s)
which permits unrestricted non-commercial use, distribution and reproduction in any medium, provided the original work is properly cited.

The author(s) has paid for this article to be freely available under the terms of the Creative Commons Attribution Non-Commercial Licence (http://creativecommons.org/licenses/by-nc/2.5/)

© 2009 The Author(s)


Casati, F., Giungiathial, F. and Marchese, M. (2007) Publish and perish: why the current publication and review model is killing research and wasting your money. ACM Ubiquity 8


Received 21 September 2009; accepted 29 September 2009
Published on the Internet 10 December 2009, doi:10.1042/BJ20091474

© 2009 The Author(s)
The author(s) has paid for this article to be freely available under the terms of the Creative Commons Attribution Non-Commercial Licence (http://creativecommons.org/licenses/by-nc/2.5/) which permits unrestricted non-commercial use, distribution and reproduction in any medium, provided the original work is properly cited.