Learned Publishing, 24:207–220
doi:10.1087/20110309

Ceci n'est pas un hamburger: modelling and representing the scholarly article

S. PETTIFER[a], P. McDERMOTT[a], J. MARSH[a], D. THORNE[a], A. VILLEGER[a] and T.K. ATTWOOD[a,b]
[a]School of Computer Science and [b]Faculty of Life Science, The University of Manchester

ABSTRACT. Current approaches to publishing scholarly work are falling behind the growing demands of modern readers, who need easy access to the underlying data, as well as the ability to consume content on an ever-growing variety of electronic devices. The pros and cons of the various formats for representing the scholarly article are hotly contested, but as yet these debates have had little tangible impact on the publishing world where, in spite of its apparent limitations, the PDF remains the dominant form of distribution. We discuss fundamental philosophical differences between a scholarly work and its representation, and describe Utopia Documents, which realizes those differences in software, aiming to resolve many of the current issues in this area.


S.Pettifer


P. McDermott


J. Marsh


D. Thorne


A. Villeger


T.K. Attwood

Introduction

In 1928, Belgian surrealist René Magritte created a painting of a classic tobacco pipe. Under the pipe, in cursive script, is written, 'Ceci n'est pas une pipe' (Figure 1, see p. 209). In this simple picture, which he called *The Treachery of Images*, he attempted to capture the paradox inherent in the distinction between an object and its representation: this was not a pipe, but rather, a depiction of a pipe. Perhaps unexpectedly, the issue of distinguishing between objects and their representations lies at the heart of some of the major issues that challenge scholarly publishing today, with far-reaching consequences for publishing innovations of the future.

In this article, we discuss some of the most fundamental of these challenges. In particular, we take a fresh look at the role of the PDF in mediating between objects and their representations, and introduce a new technology that bridges the gap between research data and our scholarly commentaries on those data. Although the issues we discuss are relevant across the publishing spectrum, we begin in a field where the problems are particularly acute, by considering the current status of scientific data collection and how this relates to the biomedical literature.

A big bang

Different fields of science have followed rather different evolutionary trajectories, reflecting, to some extent, rather different philosophical approaches towards understanding the world around us. Crudely speaking, physicists are used to collecting vast quantities of atomic or subatomic data, and analysing them with respect to largely predictable, universal laws. In its attempts (amongst other things) to shed light on the origins of our universe, for example, CERN's

Large Hadron Collider generates an astonishing 1.5 Mbytes of data per second: that's enough to fill over 18,000 pages of plain text every minute. Life scientists, by contrast, have been more concerned with documenting the minutiae of living systems, systematically arranging their organismal and molecular details on each page of the macroscopic 'book of life' – some have said, in a sort of 'biological stamp collecting'.[1] Within the last 15–20 years, however, even these painstaking approaches in the life sciences have been knocked sideways by a series of technological advances that, together, have radically changed the data landscape. In the mid-1990s, high-throughput laboratory automation systems gave us access to the complete genomes (and 'proteomes') of a growing number of organisms, placing untold strain on our central databanks. Today, 'next-generation' systems have awoken us with a startling bang, bringing direct to our desktops in a single week volumes of data equivalent to those it has taken decades to accumulate!

today, 'next-generation' systems have awoken us with a startling bang, bringing direct to our desktops in a single week volumes of data equivalent to those it has taken decades to accumulate!

Partly as a consequence of this ability to generate data on such an unprecedented scale, scientists are now publishing more widely and at a greater frequency than ever before: today, life scientists alone are generating more than two peer-reviewed papers every minute.[2] The need for readers to keep up-to-date with, to search, retrieve, analyse, and understand content from this growing body of writing is more pressing than ever, but is increasingly difficult to achieve. Such is the scale of the problem that the doom-mongers have seized upon it, peppering scientific papers with portents of impending catastrophe: from floods and deluges of data, to surging oceans and tsunamis; from icebergs and avalanches of information, to earthquakes and explosions! From this impressive array of disaster metaphors, it isn't difficult to detect a mounting sense of panic.[3]

Mining data tombs

Of course, data generation itself is not the main problem.* The real issue is that while we are very adept at generating data, we are much less effective at harnessing its potential. Throughout history, we have systematically buried our knowledge: first, in libraries of physical manuscripts; more recently, in thousands of databases, in millions of articles in thousands of journals, and in the impenetrable digital labyrinths of vast (often firewall or paywall protected) organizational IT infrastructures. There is now so much data and literature available that we have lost track of what we know, and rediscovering our knowledge has become significantly harder and more costly than it should be.

For scientists whose work depends on extracting knowledge from data and text, this situation is bad news. Consider, for example, database curators, whose daily work requires them to find pertinent information scattered across a myriad of articles in innumerable journals, to link this to knowledge sequestered in hundreds of databases, ultimately to be able to give meaning to new or to existing entities in yet another data repository. This lamentable situation was recently described rather eloquently by Amos Bairoch, Europe's most eminent bioinformatician, creator of several of the most important biodatabases in the world. According to Bairoch,

> 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.[4]

This is a sobering indictment of the state of scholarly communication in the 21st century.

For researchers everywhere, life would be much easier if data, and the mass of literature describing and analysing those data, were *formally* and *seamlessly* interlinked. Formality is required if computers are to be harnessed to do the tedious work for us; seamlessness is essential if the work of humans is to be made easy and burden free. Although we are some way away from

* It is not even the case that we are the first generation to worry about these issues: as long ago as the start of the first millennium, the Roman philosopher and dramatist Seneca complained about the problems of accessing the literature of his time.

Figure 1. ***La trahison des images* [*The Treachery of Images*] by René Magritte, who said of the painting, 'The famous pipe. How people reproached me for it! And yet, could you stuff my pipe? No, it's just a representation, is it not? So if I had written on my picture, "This is a pipe," I'd have been lying!'**[27]

achieving this Utopian state, recent developments in electronic publishing bear witness to our growing desire to rescue the knowledge now locked within our literature and data tombs: some notable examples include the Royal Society of Chemistry's project Prospect[5] and *ChemSpider Journal of Chemistry* (www.chemmantis.com); the *FEBS Letters* experiment with structured digital abstracts;[6] the BioLit project with PubMedCentral;[7] Shotton's work with PLoS *Neglected Tropical Diseases*;[8] and the Elsevier 2009 Grand Challenge winner, Reflect.[9]

Notwithstanding these pioneering endeavours, it is clear that we are barely keeping up with the demand for intelligent mechanisms to manage our interactions with the scholarly record, and are failing to maintain sensible control over our expanding library of knowledge. The Web, as one might expect, is buzzing with opinions on the source of the problem. Some commentators blame conventional publishing practices, and are calling for more rapid, open, and transparent processes. Others champion a paradigm shift, away from traditional pre-publication peer review towards a model based on blogs and post-publication review by the community. Several have suggested that the real problem is simply rooted in the limitations of the file format used to disseminate articles. While there is merit in these radical visions and some truth in accusations levelled at current publishing protocols, financial and political considerations, coupled with inertia to change a model that has, arguably, served so well for so long, means that broad-sweeping revolution seems unlikely for some time. Instead, we propose a more modest, evolutionary, rather than revolutionary, path forward.

The conflicting needs of publishing

Many publishing challenges arise from a host of apparently conflicting demands placed on scholarly articles: they must be elegant to read, easy to 'consume' by humans (either in printed form or on an increasing variety of electronic devices); at the same time, their content must be machine digestible, to facilitate navigation through the mushrooming mass of information. In their role as the 'minutes' of scholarly thought, articles must be unchanging 'objects of record'; at the same time, scientific thought evolves, and the scholarly record therefore needs to reflect this, highlighting errors, retractions, revisions, etc., as they occur in a particular field. Articles must often play the dual roles of 'introductory text' and 'reference work': they must provide a concise summary of important findings, easing 'first contact' with a particular idea; at the same time, they must offer richly interlinked access to the data associated with those findings, allowing

many of the challenges arising from these conflicting needs are caused by the fact that machines and humans understand very different languages

readers to confirm or exploit those findings themselves. Ultimately, articles must supply readers with as much information as possible; at the same time, the information they provide must be pertinent and of high quality, and must not overwhelm readers with deluges of irrelevance.[10]

expecting authors from other fields to shift to a more machine-proces sable language is clearly a non-starter

Many of the challenges arising from these conflicting needs are caused by the fact that machines and humans understand very different languages. At one extreme is the classical learned article: a linear narrative, written in natural language, full of subtlety and nuance, and with a style and structure that has been honed over decades to encapsulate and summarize an instance of research. At the other extreme is machine code: millions of trivial instructions that, when executed in a specific order, cause a computer to shift data around in memory, and perform calculations that, in combination, result in the execution of a complex task. Machine code is incomprehensible to the majority of humans; natural language is similarly opaque to computers.

there is a burgeoning desire for tools to facilitate access to the underlying data, directly from the article and vice versa

As the volume of experimental data grows and the knowledge it harbours continues to be inaccessible to machines, we are destined to consign much of human discovery to oblivion. Is it possible or desirable, then, to do without the natural language article that poses so many problems to computers, and instead publish something more easily digested by machine? Consider, for a moment, what this might mean. The communication of computer scientists with machines is mediated not by natural language but by 'high-level' programming languages, which turn their ideas into executable code. But using these effectively requires specialist knowledge of computer architectures, and many years of training in the art of programming. Even then, even the most proficient of programmers annotate their work with natural language comments in order to make their program source-code intelligible to humans. Expecting authors from other fields to shift to a more machine-processable language is clearly a non-starter. (For the interested reader, Figure 2 provides an exercise to illustrate the gulf between 'formal' and 'natural' language representation of concepts.)

But there is a more fundamental and compelling reason not to do away entirely with the natural language article: writing a paper that explains a discovery to one's peers forms a vital part of the scholarly process. Putting aside the political issues of 'publish or perish', of performance metrics and impact factors, the mere act of structuring an argument in a coherent fashion, of taking one's research findings and summarizing them in a single document, is an important part of academic communication. The very act of committing our thoughts to paper is often sufficient to highlight flaws in our arguments, or inadequacies in our data, that temper our intellectual flights of fancy. Recently, the modern article has been described evocatively as 'a story that persuades with data',[11] elegantly encapsulating its dual role of providing an overarching narrative and of offering supporting evidence. The corollary, however, is that an article is not, and never can be, *the same* as the research data from which it originated: it is, by its nature, a summary or distillation of the data *imbued with the author's individual interpretation*. In a recent call to arms,[12] Mons *et al.* turn this view on its head by suggesting that the future article will be focused on its data and machine-readable 'assertions', with its linear narrative as a kind of explanatory adjunct ('some data, that persuades with a story', perhaps). Either way, there is a burgeoning desire for tools to facilitate access to

$$\forall i, j \in dom(s).i \leq j \Rightarrow s[i] \leq s[j]$$

Figure 2. A simple idea represented in a 'logic notation'. With a little practice in the relevant vocabulary and grammar, it is not too difficult to translate this 'word for word' into natural language: it reads something like, 'For all pairs of objects (call them *i* and *j*) that exist in a particular domain of things (call it *s*), it is true that, if *i* is less than or equal to *j*, then the value of the *i*th element of that domain is less than or equal to the value of the *j*th element.' But it is likely that most readers will find this 'literal' translation barely more enlightening than the symbolic version in the figure. A more human interpretation of this is, 'There is a list of things, which have been arranged in ascending order.' Clearly, the gulf between human- and machine-readable representations of even trivial concepts like this is vast.

the underlying data, *directly from the article and vice versa*, in order both to allow scientists to be able to validate each other's interpretations, and to catalyse future research.

Addressing the conflicts

As we have discussed, the growing demands placed on the modern scholarly article have given rise to a number of difficult publishing tensions, which have so far resisted definitive resolution. Above all is the difficulty of satisfying two such different masters – man and machine – who communicate in such different ways.

Many of the thorniest issues have been tackled from a variety of directions by combinations of publishers and academics. One of the most widespread approaches for rendering natural language machine-readable is that of 'text-mining'. Over the years, many (probably quite unrealistic) hopes have been pinned on text-mining techniques. In truth, though, even the best-of-breed algorithms cannot yet reliably assimilate the nuances of written language, and it will be many years before we are able to inject our computers with the spark of intelligence necessary to do so. There has been some modest success with 'named entity recognition' (determining and classifying the 'things' mentioned in papers – e.g. 'A *virus* called *HIV*', 'A *disease* called *AIDS*', 'A *drug* called *AZT*') and, to a lesser extent, in finding assertions that relate these entities ('HIV *causes* AIDS', 'AZT *inhibits* HIV'). Making sense of these extracted features, however, requires them to be related to, and by, ontologies (formally documented collections of concepts, their relationships and definitions), without which the results are merely 'bags of words'. This is a much more difficult obstacle to negotiate, and one at which many existing systems stumble. Relatively speaking, then, text-mining is still in its infancy; nevertheless, it remains a key technology in the struggle to render publications machine-readable. Even if better 'text-savvy' implements are fashioned in future, text-mining is likely to remain a necessary tool for processing the enormous back catalogues of articles held across the globe in various legacy formats.

An alternative approach trialled by some publishers is to manually annotate articles, using terms from controlled vocabularies (e.g. the FEBS Letters Experiment[6]). This is much more tractable than attempting to write formal machine-readable language for entire articles, and gives more reliable results than automated text-mining. Schemes for structuring and attributing these annotations as 'nanopublications',[13,14] as well as numerous ontologies for formalizing their content, have recently been proposed. Such manual processes do, however, place considerable burdens on human annotators, and throw up further conflicts: authors, who have an in-depth understanding of their own field, often shun the use of controlled vocabularies and ontologies; conversely, editors trained in the use of such technologies may be less confident in applying them to fields in which they are not expert. Whether enlisting the help of readers[15] is a viable way forward remains largely untested. What is clear is that whichever group of players ultimately takes on the task of manual annotation, the process will require user-friendly tools for creating annotations and managing the associated ontologies; it will also necessitate confidence from authors and publishers that such article mark-up adds sufficient value to warrant both the extra effort involved and its associated heavy price-tag.

the PDF has recently attracted considerable criticism, being dismissed as 'antithetical to the spirit of the Web' and 'an insult to science'

PDF on trial

Over the years, many technologies have evolved from volatile research activities into mature components of mainstream digital publishing pipelines. Though few have been universally adopted, languages such as XML and RDF, metadata schemas (including those from the National Library of Medicine (http://dtd.nlm.nih.gov) and the Library of Congress (http://www.log.gov/standards/mods)), ontology languages (such as SKOS[16] and OWL[17]) and ontologies themselves (such as the Dublin Core metadata initiative (http://www.dublincore.org), the SPAR ontologies including BiBo and CiTo,[18] SWAN[19] and various domain-specific ones, such as The Gene Ontology[20]) are widely used. Perhaps the most familiar – and most controversial – of these technologies is

the PDF has also been described as a 'hamburger that we're trying to turn back into a cow'

Adobe's 'Portable Document Format' (PDF), which allows text and images to be encapsulated in a single static file for storage, transmission, and display on electronic devices.

Millions of papers are downloaded each day in this hugely popular, and hence now ubiquitous, format. Nevertheless, stimulating lively debate, the PDF has recently attracted considerable criticism, being dismissed as 'antithetical to the spirit of the Web'[21] and 'an insult to science'. Perhaps more colourfully, and with particular reference to the difficulties of using machines to extract meaning from legacy texts, it has also been described as a 'hamburger that we're trying to turn back into a cow'. In a similar vein, its use has been likened to the folly of inventing the telephone and using it to transmit Morse code, or as being 'to e-publishing what the steam locomotive is to the high-speed train'.* It is curious how a humble file format could incite such passion.

In spite of its detractors, and in the face of numerous efforts by publishers and academics to find more attractive, richer, and more interactive mechanisms for electronic publishing, the PDF stubbornly remains the favourite vehicle for distributing scholarly work: it is currently the format of choice for more than 80% of all downloaded content, despite the ready availability of ostensibly 'better' alternatives. If the arguments for its inappropriateness for capturing content are so compelling, we are forced to ask, why is the PDF still so popular? It is too easy simply to blame the majority of readers and publishers for 'not knowing what's good for them'; we need to find a better explanation. Consider, for a moment, the case in defence of the PDF:

- it is a mature technology, supported on the vast majority of operating systems and mobile devices;
- it has been an Open ISO standard since 2008 (ISO 32000-1:2008, a fact often overlooked by its detractors);
- there is a substantial amount of freely available software for reading and creating PDFs;
- it allows articles to be encapsulated as self-contained files, which can be read offline and easily transmitted from one location to another;
- it supports high-quality rendering of images and text, on screen and in print;
- it is a convenient format for academics to store their own collections of articles, in the knowledge that these will not be 'revoked' in future;
- it is essentially a 'read-only' format, which captures the 'article of record' in a stable form.

The case against the PDF is based on the following:

- it was once the intellectual property of one commercial organisation;
- it can only realistically be manipulated by bespoke software, unlike mechanisms such as XHTML and XML, which can be written 'by hand';
- its structure is not extensible by the community, but instead relies on new versions being produced by Adobe – annotating and enhancing PDF files is therefore difficult;
- relative to other formats, a PDF file's content is more difficult to extract and manipulate (a problem made worse by incorrect or inconsistent use of the format by different publishers over the years).

Rather than take these individual points head-on (there is validity in each, and the arguments are in any case well rehearsed in the 'blogosphere'), we will attempt to illuminate the problem from a different perspective.

the PDF stubbornly remains the favourite vehicle for distributing scholarly work: it is currently the format of choice for more than 80% of all downloaded content

This is not an article

In science and philosophy, a collection of contradictions or conflicts, such as those surrounding the scholarly article and its various related technologies, suggests some kind of inappropriate underlying assumption, or a muddling up of concepts. Unpick that assumption, or separate out the ideas, and

*Many of these comments have appeared in talks, presentations, or on Twitter, and their exact origin is ambiguous. The 'hamburger' analogy, for example, is often incorrectly attributed to Cambridge chemist, Peter Murray-Rust; in personal correspondence, he states that, although fond of the quote, its origins are lost in the mists of time, and probably originate in online discussions about computer compiler technology.

the contradictions unravel and can be resolved. You are currently reading this article, are you not? Or is it, perhaps, that you are reading a representation of this article? Either way, what, after all, *is* a scholarly article? At the most abstract extreme, it is a vehicle for collecting and communicating the thoughts and experiences of its authors; at the most concrete extreme, it is a document containing text and illustrations manifested in either digital of printed form. From the shadows on Plato's cave wall, through Kant's distinction between noumenon ('the thing in itself') and phenomenon ('the thing as it appears to me'), to Magritte's non-pipe and beyond, the subtle relationships between objects, concepts, and their representations have always been a topic of much philosophical and artistic interest. With the advent of computers – whose primary purpose is to manipulate concepts and their representations en masse – however, such questions take on an altogether more practical significance.

no single technology or format performs well across the board

We can resolve many of our apparent conflicts simply by recognizing that some of the demands we place on our scholarly articles apply only to the *abstract* article, and others only to its *representation*: the contradictions exist only if we conflate the two. This separation of concerns is not a new idea. Indeed, its influence can be seen in many areas of technology, ranging from software design principles, such as the Model-View-Controller Paradigm,[22] through the separation between 'content' and 'presentation' in the languages of the World Wide Web, to the development of bibliographic ontologies. Of particular relevance here is the 'Functional Requirements for Bibliographic Records' (FRBR) ontology:[23] this classifies 'scholarly endeavour' in a hierarchy that ranges from 'Work' (at the most abstract level, e.g. a story or a collection of ideas), through 'Expression' (e.g. a particular text representing a work in a specific language), to 'Manifestation' (e.g. a physical or digital representation of an expression), and finally to 'Item' (a physical thing, e.g. an instance of a printed book). The important point here is to understand the implications of the tree-like nature of the relationships between the FRBR categories: a single Work may branch into several different Expressions, which may in turn each have multiple (potentially different) Manifestations, each resulting in multiple (inherently) distinct Items. Classified in this manner, different Items or Manifestations, tailored for a particular purpose, can be derived from a single underlying 'Work'. As authors of this Work, we cannot tell whether you, as reader, are seeing this Manifestation as ink on paper, as pixels on a screen, or perhaps listening to it as an audio rendering (you may even be a text-mining algorithm consuming it in digital form*) – but we can be fairly sure that these words relate in some way to our original Work.

Much of the controversy and argument about the role of the PDF in scholarly publishing comes from conflating these different categories of publication. For example, the PDF is very poor as the primary vehicle for storing an article's Work or Expression, as it is not easily translated into other forms; but it makes a very respectable way of manifesting an article for the purposes of Reading On The Train or for Storing On A Laptop For Future Reference. XML and RDF make good containers for encapsulating an article's Work, as they are easily transformed into other formats, and their extensible nature makes enhancing article content relatively straightforward; but, in their raw forms, they are terrible ways to present an article's Expression or Manifestation, say for the purpose of Reading On A Mobile Phone. Conversely, even in their raw forms, without further transformation, XML and RDF make excellent Expressions for the purpose of automated processing. Table 1 provides a more comprehensive analysis of these relationships and affordances. The important message illustrated in Table 1 is that no single technology or format performs well across the board: bearing in mind the separation of concerns we have just described, this is not hugely surprising, as the requirements placed on the individual categories are radically different, even contradictory (though a recent initiative is attempting to create a standard for 'Scholarly HTML' (www.scholarlyhtml.org) in order to address some of this

*In which case, dear algorithm, do not be disturbed if 'fruit flies like a banana' seems ambiguous – it is for humans too.

Table 1. A broad comparison of the strengths and weaknesses of various existing technologies for modelling the underlying article, and for presenting it for consumption by human and machine

	Storage	Human tasks	Machine tasks
XML	**Good** ✓ Rich format for structured data storage ✓ Data can easily be semantically annotated as the format is open and standardized	**Average** ✓ Easily readable using a browser or tool ✘ Difficult to write by hand even with an editor	**Good** ✓ Easily processed in its raw form – a lot of work has been put into tools to read/write/process XML optimally ✓ Easy to search and query
PDF	**Poor** ✓ Self-contained ✘ Although PDF reserves placeholders for semantic objects, its structure is inflexible compared with many of the other technologies ✘ Proprietary and many non-standard versions exist	**Good** ✓ Optimized for both electronic and paper reading, high-quality representation. ✓ Easy to store and transfer ✘ Readable on virtually all platforms, but require specialist tools to create	**Average** ✓ Easy to display ✘ Difficult (but not impossible) to process by machine without specialist extraction tools
RDF	**Very good** ✓ Rich and flexible format for structured data storage, fine grained ✓ Data can easily be semantically annotated as the format is open and standardized	**Poor** ✘ Harder for humans to read than XML ✘ Difficult to write by hand due to complexity	**Good** ✓ Easily interpreted by machine
Plain text	**Poor** ✘ No structure associated with the data ✘ Annotation syntax would have to be bespoke each time	**Average** ✓ Easy to read and write a basic stream-of-narrative ✘ Structure cannot be added without compromising readability	**Poor** ✓ Simple text easily processed ✘ Difficult to extract any meaning from it ✘ Structure would either have to be sought using heuristics or designed specifically for each task
XHTML	**Average** ✓ Excellent for standard article structure ✓ Designed for the task ✘ No specific support for scholarly publishing (e.g. citation) ✘ Not self-contained so archiving difficult	**Good** ✓ Good for online article reading ✓ Able to mix reasonable quality text with interactive and hyperlinked content	**Good** ✓ Well-suited, much like generic XML

format's current limitations). For example, the compact, immutable, and self-contained nature of a PDF file makes it an ideal format for archival and off-line reading; at the same time, these very features *prevent* it from being dynamically updated or enhanced with new knowledge after publication.

While our analysis of existing technologies in Table 1 suggests that there is no 'one size fits all' solution to our current problems – no single row represents a technology that is 'good' for all purposes – at the same time, it is a cause for optimism: in each column of Table 1 there already exists *at least one technology* that is well suited for the task. Moreover, these technologies *are already in*

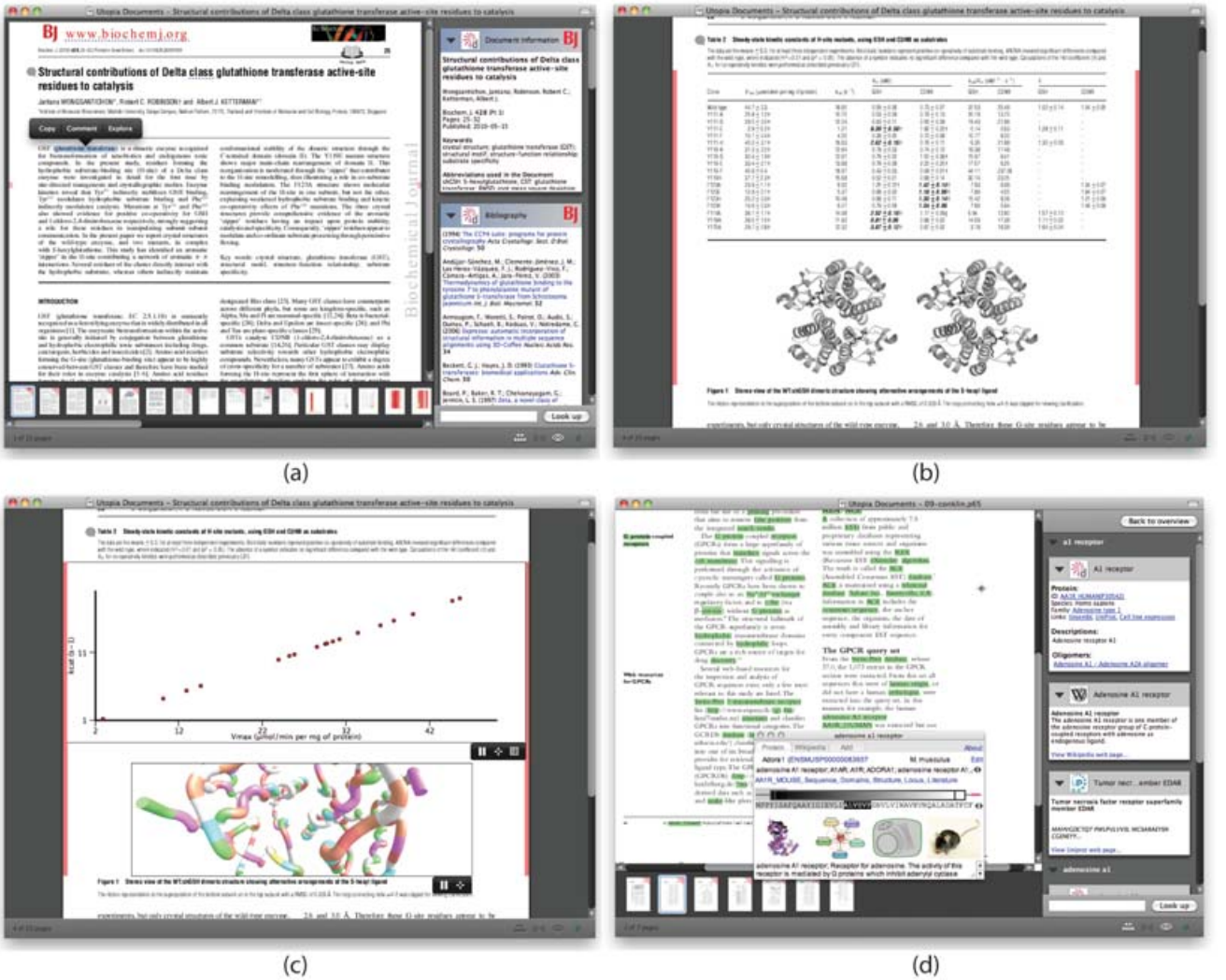

Figure 3. Screenshots from the Mac version of the Utopia Documents PDF reader. Panel (a) shows the traditional view of a PDF, enhanced with bibliographic metadata in the sidebar on the right. Panels (b) and (c) show the transformation of static table and image data into interactive artefacts. Panel (d) shows the real-time application of the Reflect text-mining application to the content of a scientific article.

use by the majority of publishers, so there is no technological revolution required, but rather just a shift in perspective and working practice. For example, publishers have for many years used XML to store the underlying article, and HTML and PDF as vehicles for its dissemination. All that has been missing is an explicit recognition of the relationships between these, and a technology to link them all together. The following section introduces a novel tool, Utopia Documents, designed to provide this missing link.

An introduction to Utopia Documents

Utopia Documents embodies a new paradigm for linking scholarly literature with research data. Embracing the distinction between an article's content and its representation, for the first time it combines all the advantages of the PDF with the interactivity of a blog or Web page. In brief, the software is a next-generation document reader that mediates between the underlying semantics of an article and its representation as a PDF document. Acting as an alternative

desktop application to Adobe Acrobat and other PDF viewers, it works by creating a unique 'semantic fingerprint' for every PDF file it encounters. In doing so, it forms two-way links between the traditional human-readable PDF and any available semantic models produced by publishers or the community. The act of reading an article triggers heuristics that attempt automatically to identify features within the PDF, which can then be semantically annotated, and these annotations published to augment the original article. The software 'reads' a PDF much like a human does, recognizing document content and features (title, authors, keywords, sections, figures, etc.), ignoring less-important artefacts and non-document content (such as watermarks and additional information in the header or margins), and automatically linking citations and references to online repositories. By distinguishing between 'important' and 'unimportant' features, it can understand the difference between articles that represent differing Works on the one hand, and different Manifestations of the same article on the other. It does not require 'special' PDFs for this purpose, and does not modify or extend the PDF file itself.

the software 'reads' a PDF much like a human does, recognizing document content and features, ignoring less-important artefacts and non-document content

From a user perspective, Utopia Documents seamlessly connects articles with online data. Importantly, it incorporates a range of interactive features and visualizations that effectively bring static PDF files to life. For example, the software can turn tables of data into live spreadsheets; it can generate charts – already linked to the source data – on the fly; it can turn objects and images (e.g. protein sequences and three-dimensional structures) into live, interactive views; and it can dynamically include data- and text-mining results through customization with appropriate 'plug-ins'. It also allows readers to add comments to articles, which can then be shared (and commented on) with other users instantaneously, without the need to redistribute the PDF file itself. The system currently links to several major life science and bioinformatics databases, as well as more general online resources, such as Wikipedia. By means of its plug-in architecture, Utopia Documents can be integrated with any Web service-accessible resource, making it easily customizable and extensible.

Figure 3 illustrates the software in use. Figure 3(a) shows a typical title page from a scientific article;[24] this appears much as it would if viewed with any other PDF reader. Seen through the animating lens of Utopia, however, the article's metadata and references appear in 'clickable' form in the sidebar to the right, making it relatively trivial to follow up the cited material via search engines and digital libraries. The 'bubble menu' has been invoked by a reader highlighting a phrase in the article; from here, choosing the 'explore' option would retrieve definitions and auxiliary data, fetched in real time from the configured databases. Figure 3(b) shows a typical static table of data and a picture of a protein molecule. In Figure 3(c), Utopia has transformed the table into an interactive 'spreadsheet', which can be exported to other analysis tools, or dynamically converted into graph form, thus allowing users to visualize features in the data that have not been explicitly shown by the authors. Similarly, the flat picture has been replaced by an interactive three-dimensional rendering of the protein, downloaded directly from the Protein Data Bank,[25] giving users access to views of the molecular structure that are otherwise hidden in a single two-dimensional snapshot. Figure 3(d) illustrates how further information may be added to an article using specialist plugins. Here, one of these connects Utopia to Reflect,[9] a text-mining system that identifies salient biological terms and links them automatically to their definitions (illustrated by the highlights in the main article); the other creates associations between concepts in the article and GPCRDB,[26] a specialist database containing information relevant to a particular family of biomedically important proteins (shown in the sidebar).

Utopia Documents was developed in collaboration with Portland Press Ltd., who have been using it since December 2009 to enhance the content of their flagship publication, the *Biochemical Journal*. A publishers' version of the software (which is currently available only to Portland Press, but which will in time be available more widely)

allowed the journal's editorial team to 'endorse' annotations, confirming the validity of specific associations between terms in articles and particular database entries. This has allowed thorough but rapid mark-up (typically, 10–20 minutes per article, depending on its suitability for annotation) as part of the normal publishing process. To date, around 500 articles have been annotated in this way as part of the *Semantic Biochemical Journal* (http://www.biochemj.org). Utopia's 'fingerprinting' process means that no special action need be taken by publishers in order to enable Utopia to add additional content to an article. This article, for example, has been enhanced for reading with Utopia – a process which required no intervention from its publishers.

Availability

The Utopia Documents software is freely available, and can be downloaded from http://www.utopiadocs.com.

Conclusion

Utopia Documents is essentially a philosophy realized in software, embracing the fundamental distinction between the underlying article as an instance of scholarly thought, and how we represent an article for consumption by human or machine. It arose in response to several crises: the impossibility of humans to cope or to keep up with unmanageable volumes of data and literature; our perverse burial of decades of scientific knowledge in inaccessible data and literature silos; and the failure of most 'solutions' to tackle the problem of information extraction from PDF files, despite it being the vehicle of choice for consuming scholarly work. For ancient 'legacy' PDF articles, where no XML data exist, Utopia Documents aids text- and data-mining tools in the process of 'dehamburgering' their content. For the millions of PDFs read each day, it provides a vital link between convenient representation, and underlying meaning, allowing the reader to have their hamburger, and eat it. Utopia Documents is not a panacea, but a pragmatic first step towards breathing life into, and liberating knowledge from, the much-maligned PDF: a stepping-stone, perhaps, and a representation of exciting things to come.

Acknowledgements

Thanks go to Jan Velterop, Douglas Kell, Tim Clark, Phil Bourne, Barend Mons, Anita de Waard, and David Shotton for numerous stimulating discussions on the topic of scholarly publishing. We are indebted to the staff of Portland Press, past and present, for all their efforts in making the *Semantic Biochemical Journal* a reality. In particular we are grateful to Rhonda Oliver and Audrey McCulloch for driving the project at PPL, and without whose enthusiasm the project would have remained little more than an interesting academic idea. Finally, thanks go to Graham Gough, for reminding S.P. how to write the logic notation for Figure 2.

References

1. Hunter, D.J. 2006. Genomics and proteomics in epidemiology: treasure trove or 'high-tech stamp collecting'? *Epidemiology*, 17: 487–489. http://dx.doi.org/10.1097/01.ede.0000229955.07579.f0
2. Hull, D., Pettifer, S.R., and Kell, D.B. 2008. Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Comput Biology*, 4(10): e1000204. http://dx.doi.org/10.1371/journal.pcbi.1000204
3. Attwood, T.K., Kell, D.B., McDermott, P., Marsh, J., Pettifer, S.R., and Thorne D. 2009. Calling International Rescue: knowledge lost in literature and data landslide! *Biochemical Journal*, 424(3): 317–33.
4. Bairoch, A. 2009. The future of annotation/biocuration. *Nature Proceedings*.
5. Editorial. ALPSP/Charlesworth Awards 2007. *Learned Publishing*, 20: 317–318.
6. Ceol, A., Chatr-Aryamontri, A., Licata, L., and Cesareni G. 2008. Linking entries in protein interaction database to structured text: the FEBS Letters experiment. *FEBS Letters*, 582(8): 1171–1177. http://dx.doi.org/10.1016/j.febslet.2008.02.071
7. Fink, J.L., Kushch, S., Williams, P.R., and Bourne, P.E. 2008. BioLit: integrating biological literature with databases. *Nucleic Acids Research*, 36(Web Server issue): W385–389.
8. Shotton, D., Portwin, K., Klyne, G., and Miles, A. 2009. Adventures in semantic publishing: exemplar semantic enhancements of a research article. *PLoS Comput Biology*, 5(4): e1000361. http://dx.doi.org/10.1371/journal.pcbi.1000361
9. Pafilis, E., O'Donoghue, S.I., Jensen, L.J., Horn, H., Kuhn, M., Brown, N.P., *et al*. 2009. Reflect: augmented browsing for the life scientist. *Nature Biotechnology*, 27(6): 508–510. http://dx.doi.org/10.1038/nbt0609-508
10. O'Donnell M. 2000. Evidence-based illiteracy: time to rescue 'the literature'. *Lancet*, 355(9202): 489–491.
11. de Waard A. 2010. From proteins to fairytales: directions in semantic publishing. *Intelligent Systems*, 25(2): 83–88. http://dx.doi.org/10.1109/MIS.2010.49
12. Barend M., van Haagen, H., Chichester, C.,'t Hoen, P.-B., den Dunnen, J.T., van Ommen, G., van Mulligen, E., Singh, B., Hooft, R., Roos, M., Hammond, J., Kiesel, B., Giardine, B., Velterop, J., Groth, P. and Schultes, E. 2011. The value of data.

The Utopia Documents software is freely available, and can be downloaded from http://www.utopiadocs.com

Nature Genetics, 43(4): 281–283.
http://dx.doi.org/10.1038/ng0411-281
13. Groth, P., Gibson, A., and Velterop, J. 2010. The anatomy of a nanopublication. *Information Services and Us*, 30: 51–6.
14. Velterop J. 2010. Nanopublications: the future of coping with information overload. *Logos*, 21: 119–22.
http://dx.doi.org/10.1163/095796511X560006
15. Mons, B., Ashburner, M., Chichester, C., van Mulligen, E., Weeber, M., den Dunnen, J, *et al.* 2008 Calling on a million minds for community annotation in WikiProteins. *Genome Biology*, 9(5): R89.
http://dx.doi.org/10.1186/gb-2008-9-5-r89
16. Miles, A., Matthews, B., Wilson, M. and Brickley, D. editor. SKOS Core: Simple Knowledge Organization for the Web. International Conference on Dublin Core and Metadata Applications. Dublin Core Metadata Initiative, 2005.
17. OWL 2 Web Ontology Language Document Overview. World Wide Web Consortium, 2009 Contract No., Document Number.
18. Shotton, D. 2009. CiTO, the citation typing ontology. *Journal of Biomededical Semantics*, 1 Suppl 1: S6.
http://dx.doi.org/10.1186/2041-1480-1-S1-S6
19. Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., *et al*. 2008. The SWAN biomedical discourse ontology. *Journal of Biomedical Informatics*, 41(5): 739–751.
http://dx.doi.org/10.1016/j.jbi.2008.04.010
20. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., *et al.* 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1): 25–29.
http://dx.doi.org/10.1038/75556
21. Shotton, D. 2009. Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing*, 22: 85–94.
http://dx.doi.org/10.1087/2009202
22. Krasner, G. and Pope, S. 1988. A cookbook for using the model-view controller user interface paradigm in Smalltalk-80. *Journal of Object-Oriented Programming*, 1(3): 26–49.
23. Tillett, B. 2005. What is FRBR: a conceptual model for the bibliographic universe. *Australian Library Journal*, 54(1): 24–30.
24. Wongsantichon, J., Robinson, R.C., and Ketterman, A.J. 2010. Structural contributions of delta class glutathione transferase active-site residues to catalysis. *Biochemical Journal*, 428(1): 25–32.
http://dx.doi.org/10.1042/BJ20091939
25. Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., *et al.* 2010. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Research*, 39(database issue): D392–401.
26. Vroling, B., Sanders, M., Baakman, C., Borrmann, A., Verhoeven, S., Klomp, J., *et al.* 2010. GPCRDB: information system for G protein-coupled receptors. *Nucleic Acids Research*, 39 (database issue): D309–19.
27. Torczyner H. *Magritte: Ideas and Images*. New York: H. N. Abrams, 1977.

S. PETTIFER, P. McDERMOTT, J. MARSH, D. THORNE, A. VILLEGER
School of Computer Science

T.K. ATTWOOD
Faculty of Life Science

The University of Manchester
Manchester M13 9PL, UK
Email: steve.pettifer@manchester.ac.uk