BIOINFORMATICS

Vol. 26 ECCB 2010, pages i568–i574
doi:10.1093/bioinformatics/btq383

Utopia documents: linking scholarly literature with research data

T. K. Attwood[1,2,*], D. B. Kell[3,4], P. McDermott[1,2], J. Marsh[3], S. R. Pettifer[2] and D. Thorne[3]

[1]School of Computer Science, [2]Faculty of Life Sciences, [3]School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL and [4]Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester M1 7DN, UK

ABSTRACT

Motivation: In recent years, the gulf between the mass of accumulating-research data and the massive literature describing and analyzing those data has widened. The need for intelligent tools to bridge this gap, to rescue the knowledge being systematically isolated in literature and data silos, is now widely acknowledged.

Results: To this end, we have developed Utopia Documents, a novel PDF reader that semantically integrates visualization and data-analysis tools with published research articles. In a successful pilot with editors of the *Biochemical Journal* (*BJ*), the system has been used to transform static document features into objects that can be linked, annotated, visualized and analyzed interactively (http://www.biochemj.org/bj/424/3/). Utopia Documents is now used routinely by *BJ* editors to mark up article content prior to publication. Recent additions include integration of various text-mining and biodatabase plugins, demonstrating the system's ability to seamlessly integrate on-line content with PDF articles.

Availability: http://getutopia.com

Contact: teresa.k.attwood@manchester.ac.uk

1 INTRODUCTION

The typhoon of technological advances witnessed during the last decade has left in its wake a flood of life-science data, and an increasingly impenetrable mass of biomedical literature describing and analysing those data. Importantly, the modern frenzy to gather more and more information has left us without adequate tools either to mine the rapidly increasing data- and literature-collections efficiently, or to extract useful knowledge from them. To be usable, information needs to be stored and organized in ways that allow us to access, analyze and annotate it, and ultimately to relate it to other information. Unfortunately, however, much of the data accumulating in databases and documents has not been stored and organized in rigorous, principled ways. Consequently, finding what we want and, crucially, pinpointing and understanding what we already know, have become increasingly difficult and costly tasks (Attwood *et al.*, 2009).

A group of scientists for whom these problems have become especially troublesome are biocurators, who must routinely inspect thousands of articles and hundreds of related entries in different databases in order to be able to attach sufficient information to a new database entry to make it meaningful. With something like 25 000 peer-reviewed journals publishing around 2.5 million articles per year, it is simply not possible for curators to keep abreast of developments, to find all the relevant papers they need, to locate the most relevant facts within them, and simultaneously to keep pace with the inexorable data deluge from ongoing high-throughput biology projects (i.e. from whole genome sequencing). For example, to put this in context, Bairoch estimates that it has taken 23 years to manually annotate about half of Swiss-Prot's 516 081 entries (Bairoch, 2009; Boeckmann *et al.*, 2003), a painfully small number relative to the size of its parent resource, UniProtKB (The UniProt Consortium, 2009), which currently contains ∼11 million entries. Hardly surprising, then, that he should opine, 'It is quite depressive to think that we are spending millions in grants for people to perform experiments, produce new knowledge, hide this knowledge in a often badly written text and then spend some more millions trying to second guess what the authors really did and found' (Bairoch, 2009).

The work of curators, and indeed of all researchers, would be far easier if articles could provide seamless access to their underlying research data. It has been argued that the distinction between an online paper and a database is already diminishing (Bourne, 2005); however, as is evident from the success stories of recent initiatives to access and extract the knowledge embedded in the scholarly literature, there is still work to be done. Some of these initiatives are outlined below.

The Royal Society of Chemistry (RSC) took pioneering steps towards enriching their published content with data from external resources, creating 'computer-readable chemistry' with their Prospect software (Editorial, 2007). They now offer some of their journal articles in an enhanced HTML form, annotated using Prospect: features that may be marked up include compound names, bio- and chemical-ontology terms, etc. Marked-up terms provide definitions from the various ontologies used by the system, together with 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. Articles enriched in this way make navigation to additional information trivial, and significantly increase the appeal to readers.

In a related project, the *ChemSpider Journal of Chemistry* exploits the ChemMantis System to mark up its articles (http://www.chemmantis.com). With the ChemSpider database at its heart, ChemMantis identifies and extracts chemical names, converting them to chemical structures using name-to-structure conversion algorithms and dictionary look-ups; it also marks up chemical families, groups and reaction types, and provides links to Wikipedia definitions where appropriate.

In an initiative more closely related to the life sciences, *FEBS Letters* ran a pilot study (Ceol *et al.*, 2008) with the curators of the MINT interaction database (Chatr-aryamontri *et al.*, 2007), focusing on integration of published protein–protein interaction and post-translational modification data with information stored in MINT and UniProtKB. Key to the experiment was the Structured Digital

*To whom correspondence should be addressed.

***Structured summary*:**

MINT-6173230, MINT-6173253:
TSC22 (uniprotkb:Q15714) physically interacts (MI:0218) with fortilin (uniprotkb:P13693) by co-immunoprecipitation (MI:0019)
MINT-6173217:
TSC22 (uniprotkb:Q15714) binds (MI:0407) fortilin (uniprotkb:P13693) by pull-down (MI:0096)
MINT-6173240, MINT-6173270:
TSC22 (uniprotkb:Q15714) physically interacts (MI:0218) with fortilin (uniprotkb:P13693) by two-hybrid (MI:0018)

Fig. 1. Structured summary for an article in *FEBS Letters* (Lee *et al*., 2008). Three interactions are shown, with their links to MINT and UniProtKB.

Abstract (SDA), a device for capturing an article's key facts in an XML-coded summary, essentially to make them accessible to text-mining tools (Seringhaus and Gerstein, 2007); these data were collected from authors via a spreadsheet, and structured as shown in Figure 1—while clearly machine-readable, this format has the notable disadvantage of being rather human unfriendly.

A different approach was taken with BioLit (Fink *et al.*, 2008), an open-source system that integrates a subset of papers from PubMed Central with structural data from the Protein Data Bank (PDB) (Kouranov *et al*., 2006) and terms from biomedical ontologies. The system works by mining the full text for terms of interest, indexing those terms and delivering them as machine-readable XML-based article files; these are rendered human-readable via a web-based viewer, which displays the original text with colored highlights denoting additional context-specific functionality (e.g. to view a 3D structure image, to retrieve the protein sequence or the PDB entry, to define the ontology term).

A more adventurous approach was taken by Shotton *et al*. (2009), who targeted an article in *PLoS Neglected Tropical Diseases* for semantic enhancement. The enrichments they included were live Digital Object Identifiers and hyperlinks; mark-up of textual terms (disease, habitat, organism, etc.), with links to external data resources; 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. In addition, they provided 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.

To stimulate further advances in the way scientific information is communicated and used, Elsevier offered its Grand Challenge of Knowledge Enhancement in the Life Sciences in 2008. The contest aimed to develop tools for semantic annotation of journals and text-based databases, and hence to improve access to, and dissemination of, the knowledge contained within them. The winning software, Reflect, focused on the dual need of life scientists 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 (Pafilis *et al*., 2009). Drawing on a large, consolidated dictionary that links names and synonyms to source databases, Reflect tags such entities when they occur in web pages; when clicked on, the tagged items invoke pop-ups displaying brief summaries of entities such as domain and/or small molecule structures, interaction partners and so on, and allow navigation to core biological databases like UniProtKB.

All of these initiatives differ slightly in their specific aims, but nevertheless reflect the same aspiration—to get more out of digital documents by facilitating access to underlying research data. As such, it is interesting to see that a number of common themes have emerged: most are HTML- or XML-based, providing hyperlinks to external web sites and term definitions from relevant ontologies via color-coded textual highlights; most seem to ignore PDF as a foundation for semantic enrichment (despite a significant proportion of publisher content being offered in this format). The results of these projects are encouraging, each offering valuable insights into what further advances need to be made: clearly, we need to be able to link more than just a single database to a single article, or a single database to several articles, or several databases to a single issue of a single journal. Although necessary proofs of principle, these are just first steps towards more ambitious possibilities, and novel tools are still needed to help realize the goal of fully integrated literature and research data.

In this article, we describe a new software tool, Utopia Documents, which builds on Utopia, a suite of semantically integrated protein sequence/structure visualization and analysis tools (Pettifer *et al*., 2004, 2009). We describe the unique functionality of Utopia Documents, and its use in semantic mark-up of the *Biochemical Journal* (*BJ*). We also outline the development of a number of new plugins, by means of which we have imported additional functionality into the system via web services.

2 SYSTEM AND METHODS

Utopia Documents was developed in response to the realization that, in spite of the benefits of 'enhanced HTML' articles online, most papers are still read, and stored by researchers in personal archives, as PDF files. Several factors likely contribute to this reluctance to move entirely to reading articles online: PDFs can be 'owned' and stored locally, without concerns about web sites disappearing, papers being withdrawn or modified, or journal subscriptions expiring; as self-contained objects, PDFs are easy to read offline and share with peers (even if the legality of the latter may sometimes be dubious); and, centuries of typographic craft have led to convergence on journal formats that (on paper and in PDF) are familiar, broadly similar, aesthetically pleasing and easy to read.

In its current form, Utopia Documents is a desktop application for reading and exploring papers, and behaves like a familiar PDF reader (Adobe Acrobat, KPDF, OS X Preview, etc.); but its real potential becomes apparent when configured with appropriate domain-specific ontologies and plugins. With these in place, the software transforms PDF versions of articles from static facsimiles of their printed counterparts into dynamic gateways to additional knowledge, linking both explicit and implicit information embedded in the articles to online resources, as well as providing seamless access to auxiliary data and interactive visualization and analysis tools. The innovation in the software is in implementing these enhancements without compromising the integrity of the PDF file itself.

Suitably configured, Utopia Documents is able to inspect the content and structure of an article, and, using a combination of automated and manual mechanisms, augment this content in a variety of ways:

2.1 Adding definitions

Published articles are typically restricted to a defined page count, and are usually written for a specific audience. Explanations of terms that might be useful to newcomers to a particular field are therefore frequently

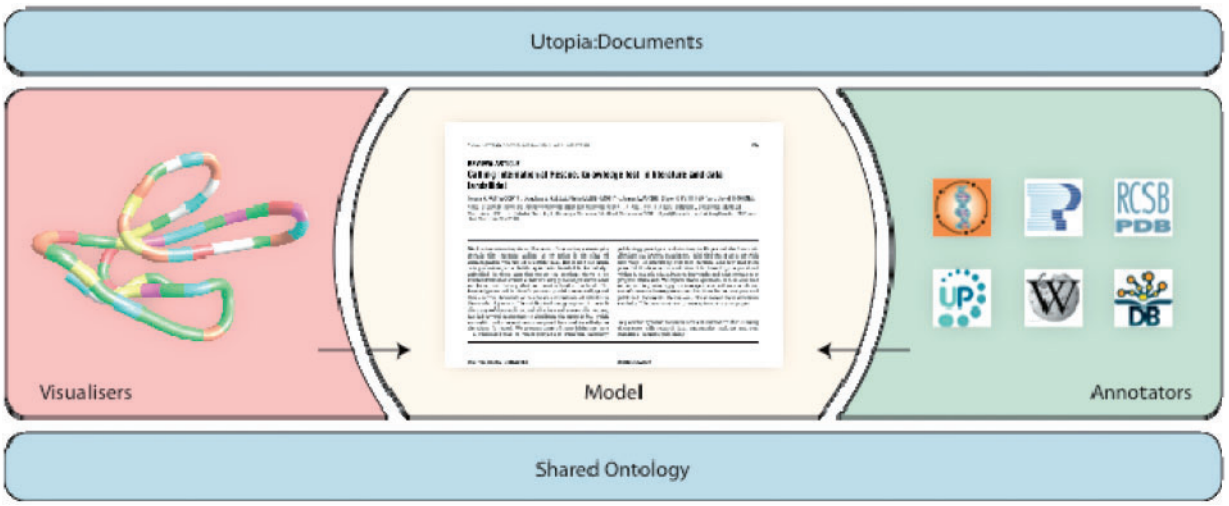

Fig. 2. The architecture of Utopia Documents, showing the relationship between the GUI (top), plugins (middle) and ontology (bottom).

omitted. Utopia Documents allows editors and authors to annotate terms with definitions from online resources (Wikipedia, UniProtKB, PDB, etc.), and permits readers to easily find definitions for themselves.

2.2 Interactive content and auxiliary data

Figures and tables in printed form are typically static snapshots of richer data (which are nowadays often available elsewhere online). For example, a table may represent the salient fragment of a much larger experimental dataset, or an image of a protein structure might highlight one specific feature of that molecule. Utopia Documents is able to transform such static tables and figures, *in situ,* into dynamic, interactive objects, providing richer access to the underlying data.

2.3 Linking references to source articles

Most articles published today are made available in electronic form, and substantial efforts are also made by publishers to make their back-catalogues electronically accessible. Navigating the multitude of online repositories and bibliographic tools, however, is complex. Utopia Documents simplifies the process of finding related articles by automatically linking references to their digital online versions.

3 IMPLEMENTATION

The software architecture comprises three main components, as shown in Figure 2: 'the core', providing generic mechanisms for displaying and manipulating articles, both programmatically and interactively; 'the plugins', which analyze, annotate and visualize document features, either automatically or under the guidance of a user; and 'the ontology', which is used to semantically integrate the other components.

3.1 The core

Optimized for interactivity, the multi-threaded core of Utopia Documents is written in C++ using Trolltech/Nokia's Qt toolkit. The core serves two purposes: (i) it performs the relatively mundane tasks necessary to generate and manage the interactive Graphical User Interface (GUI) and to co-ordinate the behavior of the plugins, which are loaded on demand at run-time; (ii) it carries out the low-level analysis of PDF documents, including reading their file format and converting them into both a visual representation to be displayed on-screen and a hierarchical semantic model for later higher-level analysis and annotation by the plugins. The analysis performed by the core is generic in nature, and is restricted at this stage to identifying typographical and layout-based features common to scholarly papers from any discipline. Once this raw structure has been generated, using various heuristics, the system then identifies higher-level typographical constructs (titles, sections, headings, figures, tables, references, etc.). Annotations identifying these features are assembled, and added to the raw hierarchy to form a semantic model that is then shared with, and further annotated by, the plugins. From these 'structural semantics', a 'fingerprint' is created that uniquely identifies the article being read, and allows the annotations to be associated with it.

3.2 The plugins

Two broad classes of plugin are defined. 'Annotators' inspect a document's content and semantic structure, then either apply local algorithms or communicate with external services in order to create annotations containing additional content (e.g. definitions of terms, user comments, links to other resources). Annotator plugins may be configured to execute automatically when a document is loaded, typically performing document-wide tasks, such as identifying terms of biological or chemical interest; alternatively, they may be invoked manually via the GUI—in these cases, the plugins have access to the GUI's state, and can generate context-specific annotations (e.g. associating a highlighted region of text with a specific comment made by a user, or finding the definition of a highlighted concept in an online database.) 'Visualizers' provide various mechanisms for displaying and interacting with annotations: e.g. an annotation containing static images, links and 'rich text' may be displayed using a browser-like visualizer, whereas one containing the structure of a molecule from the PDB might be displayed as an interactive 3D object. Both types of plugin may be written in C++ or Python, and are executed in their own asynchronous environment, marshalled by the core.

3.3 The ontology

Rather than create 'hard-wired' relationships between the system's components, a simple ontology (in its current form, a hierarchical taxonomy) connects the plugins to the core and to one another. This form of semantic integration allows the components to cooperate flexibly in the analysis of document content and structure, and allows plugins to be developed independently of one another, with sensible relationships and behavior being inferred at run-time rather than being pre-determined: e.g. an annotator plugin may mark content as containing 'protein structure'; a visualizer plugin, encountering this annotation at a later stage, can then decide whether to display this as a 2D static image, an interactive 3D model or as a 1D amino acid sequence.

3.4 Access to remote resources

Via its plugins, Utopia Documents has access to a wealth of bioinformatics data. Each plugin can use whatever client libraries are appropriate to access web-service endpoints (both SOAP- and REST-style), as well as other remotely accessible resources, such as relational databases and RDF stores. Of particular note here are two substantial 'linked data' initiatives that have proven to be of enormous value to our work. The first of these, the Bio2RDF project (Belleau *et al*., 2008), combines the content of many of the major life-science databases as a federated linked-data network accessible via SPARQL and REST interfaces. This both offers a single mechanism via which Utopia Documents can search multiple primary databases, and enforces a consistent naming scheme between sources, allowing results to be interrelated. The second (and more general), DBPedia, is a machine-readable RDF-based conversion of the popular human-readable Wikipedia (Auer *et al*., 2007). Although containing much information that is irrelevant to the life sciences, Wikipedia (and thus DBPedia) has evolved to represent a significant and mostly authoritative corpus of scientific knowledge—a study performed by the journal *Nature* concluded that its entries were as accurate (or indeed, as error prone) as those published in *Encylopaedia Britannica* (Giles, 2005, 2006). The combined application of ontologies and RDF in DBPedia allows queries performed by Utopia Documents to traverse only the portions of the DBPedia network that are semantically related to the life sciences. Thus, in the context of a paper on enzymatic substrate cleavage, a search initiated via Utopia Documents for the term 'cleavage' returns far more appropriate definitions than would the same search in a more generic context.

Utopia Documents is freely available via the project web site for Mac OS X (10.4 and later), Microsoft Windows XP and Vista and Ubuntu Linux. We welcome any feedback on the software.

4 RESULTS AND DISCUSSION

Utopia Documents was developed in response to the need to achieve tighter coupling between published articles and their underlying data, ultimately to facilitate knowledge discovery. The tool was designed with two classes of user in mind; the reader, as consumer of published material; and the journal editor, as curator. To this end, the software was piloted with Portland Press Limited (PPL) with the goal of rendering the content of *BJ* electronic publications and supplemental data richer and more accessible.

To achieve this, an 'editor's version' of Utopia Documents, with customized plugins, was integrated with PPL's editorial and document-management workflows, allowing *BJ* editors to mark up article content prior to publication. In terms of functionality, the editor's version of the software behaves much the same as the reader's, with the additional feature that relationships between concepts in a document and online definitions/records can be made permanent in order to be shared with readers (Fig. 3g). The role of the editors was therefore to explore each pre-publication PDF, annotating terms and figures with definitions and interactive content and then validating them with a 'stamp of approval' (i.e. the *BJ* icon).

With the customized software in-house, article annotation was fairly swift, individual papers taking 10–30 min, depending on their suitability for mark-up. The launch issue of the Semantic *BJ* (December 2009; http://www.biochemj.org/bj/424/3/) was primarily handled by two editors; since then, the whole editorial team has been involved in the successful mark-up of eight further issues. Entities relating to protein sequences and structures have been, of necessity, the main targets for mark-up, because this was the functionality built into the original Utopia toolkit. The kinds of additional mark-up provided by the software include links from the text to external web sites, term definitions from ontologies and controlled vocabularies, embedded data and materials (images, videos, etc.) and links to interactive tools for sequence alignment and 3D molecular visualization.

To allow readers to benefit from these semantic enhancements, a reader's version of the software was made freely available (http://getutopia.com). The tool installs easily on the desktop as an alternative PDF viewer. Once opened, it displays a window consisting of three regions (Fig. 3): the main reading pane displays the article itself and supports the pagination, searching, zooming and scrolling features typical of PDF readers. Below this, thumbnail images give an overview of the document and allow rapid navigation through it. The sidebar on the right displays the contents of annotations, providing term definitions and access to auxiliary data as the article is explored. When no specific terms are selected, the sidebar defaults to displaying document-wide metadata [including the title, authors, keywords, abbreviations, etc. (3d)], in addition to the cited references (3e)—these are linked, where available, via open-access publishing agreements or institutional or individual subscriptions, to the online versions of the original articles. Where the PDF version is not available to the reader, clicking on the reference currently launches a Google Scholar search instead.

To avoid cluttering the text with 'highlighter pen'-type marks, the presence of annotations, or availability of auxiliary data, is indicated by discreet colored glyphs in the margin. Similar marks are added to the corner of the corresponding thumbnail in the pager, to indicate that additional information exists somewhere on that page. Mousing-over a glyph highlights the nearby terms, or document regions, that contain annotations; selecting these areas causes the associated data to be displayed—this may involve populating the sidebar with definitions, or may activate an embedded interactive visualization. Highlighting any word or phrase in the paper (3a) initiates a context-sensitive search of the online resources to which Utopia Documents is connected, all results again appearing in the sidebar. At the bottom of the sidebar (3b), a 'lookup' feature allows searches for terms not explicitly mentioned in the paper.

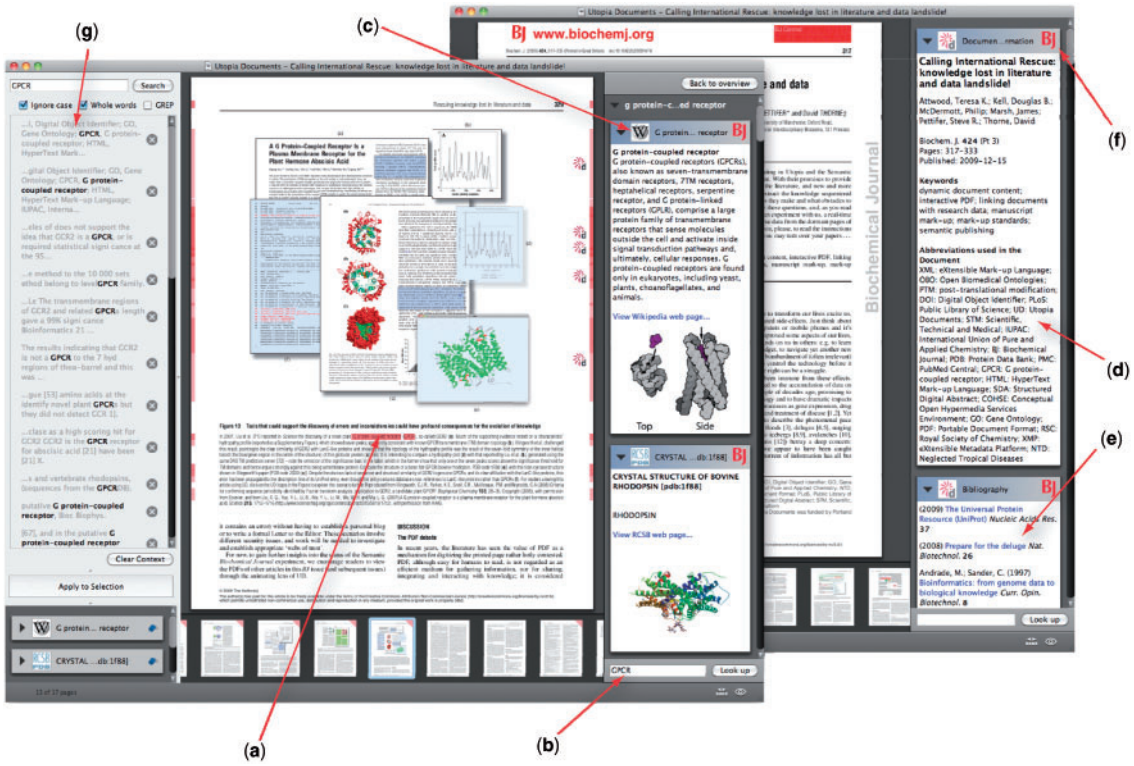

Fig. 3. Utopia Documents' user interface showing: (**a**) a selected term in the article; (**b**) manual term lookup; (**c**) resulting definitions of that term retrieved from Wikipedia (via DBPedia) and the PDB; (**d**) metadata relating to the whole document (shown when no specific term definition is selected); (**e**) live links to articles in the article's bibliography; (**f**) an icon indicating the 'authority' for a particular annotation (here, the *BJ*) and (**g**) the panel used by *BJ* editorial staff to associate terms with annotations (note that this is only available in the 'editor's version' of Utopia Documents).

4.1 Annotations

An annotated term or region in a document may be associated with definitions and/or database records from a variety of sources. Selecting a term invokes the display of all possible definitions, allowing the reader (or editor) to select for themselves the most appropriate version. The provenance of these definitions is indicated in their headers, as illustrated in Figure 3: the icon on the left (3c) represents the item's origin [e.g., UniprotKB, Wikipedia, KEGG (Kanehisa *et al.*, 2010)], while the presence of an icon on the right-hand side of the header (3f) indicates the person, group or organization who made, and endorsed, the association between a term and this specific definition (here, publisher-validated annotations carry the *BJ* logo).

4.2 Interactive content

The current version of Utopia Documents supports three forms of embedded interactive content; as with term definitions, these are indicated by red glyphs in the margins. Selecting these causes a 'media player'-like panel to appear, which the reader can use to control the behavior of the interactive content. Activating the triangular 'play' button replaces the static content, *in situ*, with its interactive version; the neighboring 'pop-up' button opens a new window leaving the static page unchanged. Each type of interactive content has its own functionality: 3D molecules (Fig. 4), for example, can be rotated, zoomed and rendered in a variety of styles (e.g. space-fill, backbone or cartoon); sequences and their associated features can be inspected individually, or edited as multiple alignments; and tables of data can be manipulated or converted automatically into scatter-plots or histograms. Figure 4 illustrates the simple transformation from static images of tables and figures into semantically annotated, interactive objects.

Utopia Documents provides new ways of reading, of interacting with and ultimately of assimilating the knowledge embodied within research articles. The approach taken here departs from many initiatives in scholarly publishing in that the focus for enrichment is the hitherto-largely-neglected static PDF file, rather than HTML- or XML-based files. The subject of 'static PDF' versus 'dynamic online' articles has been hotly contested in the literature, the general consensus being that PDF is semantically limited by comparison with other online formats and is thus antithetical to the spirit of web publishing (Lynch, 2007; Renear and Palmer, 2009; Shotton *et al.*, 2009; Wilbanks, 2007). We argue that PDFs are merely a mechanism for rendering words and figures, and are thus no more or less 'semantic' than the HTML used to generate web pages. Utopia Documents is hence an attempt to provide a semantic bridge that connects the benefits of both the static and the dynamic online incarnations of published texts. Inevitably, those who prefer to read articles online in a web browser will view the need to download a new, desktop-based PDF reader as a weakness. Our view is, rather, that Utopia Documents complements browser-based tools, providing a novel mechanism for unleashing knowledge that is

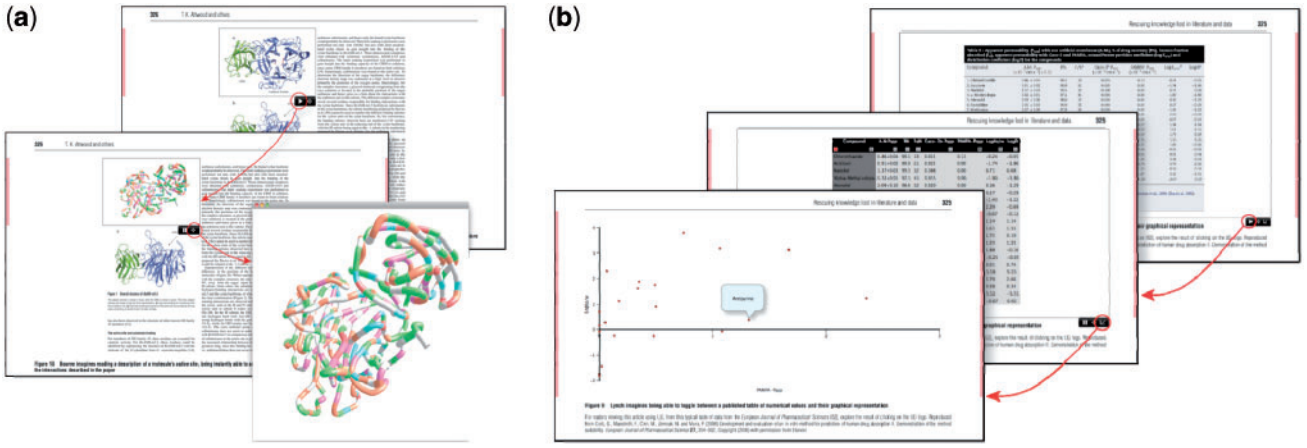

Fig. 4. Image sequences showing the transformation of a 2D image (left-hand panel) and of a static table of figures (right-hand panel) into interactive objects: i.e. a manipulable 3D model (coordinates extracted from the PDB) and a set of 'live' figures and a customizable semantic graph.

otherwise locked in personal, publisher and/or institutional PDF-file archives.

In contrast with approaches for creating dynamic (as opposed to 'semantic') life-science PDF articles (Kumar *et al.*, 2008; Ruthensteiner and Hess, 2008) that use Adobe Acrobat's support for 'Universal 3D Data' (U3D), Utopia Documents does not insert its augmented content into the PDF file itself, but instead blends additional visual material into the display process at the final stages of rendering. This mechanism presents a number of benefits over the generic U3D approach: (i) the underlying PDF file remains small and compact, and does not become bloated by the large polygonal meshes necessary for rendering 3D molecules; (ii) rather than the 'one size fits all' U3D approach, Utopia Documents is able to select appropriate rendering and interaction algorithms for different types of artifact; (iii) Utopia Documents is able to maintain a semantic relationship between the underlying scholarly article and the object being rendered; and importantly; (iv) the original PDF, as an 'object of record', remains unadulterated and its integrity can be verified by examining it with a conventional PDF viewer.

The philosophy embodied in Utopia Documents is to hide as much of the underlying complexity as possible, to avoid requiring users (whether editors, authors or readers) to change their existing document-reading behaviors, and to present no significant extra hurdles to publication. Like the initiatives in semantic publishing outlined earlier, the Semantic *BJ*, powered by Utopia Documents, is a pilot, the success of which will depend on various factors, including whether the barriers to adoption are sufficiently low, and whether the approach is considered to add sufficient value. Although it is too early to assess the impact of the pilot on readers of the Semantic *BJ*, the take-up of the software by the *BJ*'s full editorial team, and it use to mark up every issue since the launch, is a testament to the software's ease-of-use. Of course, as the project with PPL develops, we will gather relevant usage and usability data in order to provide a more meaningful evaluation.

Many of the projects discussed in this article have exploited fairly traditional text-mining methods, in conjunction with controlled vocabularies and ontologies, to facilitate the launch of relevant external web pages from marked-up entities in documents. As such, they come with all the limitations in precision of current text-mining tools; this brings a significant overhead to readers in terms of having to identify errors. Of course, the difficulty for non-experts in any given field is to be able to recognize when particular annotations really are errors, and failure to identify them as such leads to the danger of error propagation. In light of these issues, we took a slightly different approach to entity mark-up in this first incarnation of Utopia Documents, taking advantage of linked-data initiatives to facilitate mark-up and add value to published texts. However, because the functionality of the system is easily customizable via its flexible plugin architecture, any text-mining tool or database that is accessible via web services can be trivially added to the suite. As a demonstration of the potential of this architecture, in collaboration with their developers, three prototype plugins that link Utopia to other systems have been implemented:

Reflect: as mentioned earlier, the Reflect system is primarily used as a means of augmenting HTML content online, either by accessing a web page via the project's portal, or by installing a browser plugin (http://reflect.ws/). Its entity-recognition engine, however, may also be accessed programmatically via a web service, which, given a section of text, identifies objects of biological interest and returns links to the summary pop-ups. Integration of Reflect's functionality with Utopia Documents is therefore a comparatively straightforward task: as a user reads a PDF document, its textual content is extracted and sent to the Reflect web service; the resulting entities are then highlighted in the PDF article, and linked to the appropriate pop-up, which is displayed when a highlighted term is selected. A particular advantage of this integration is that it provides the reader with a light-weight mechanism for verifying or cross-checking results returned from multiple sources (e.g. Reflect, Bio2RDF, DBpedia/Wikipedia).

GPCRDB: this is a specialist database describing sequences, ligand-binding constants and mutations relating to G protein-coupled receptors (http://www.gpcr.org/). Its recently developed web-service interface provides programmatic access to much of its content, enabling Utopia Documents to identify and highlight receptors and their associated mutants when encountered in PDFs. Thus, the presence of a GPCR in an article triggers the creation of a link to a description of that receptor in the database, which is displayed in the sidebar. The article is then scanned for mutants, which in turn are linked to the relevant mutant records in

GPCRDB. Having identified an appropriate receptor, the software then automatically translates between the sequence co-ordinates, allowing 'equivalent' residues to be readily mapped between them.

ACKnowledge Enhancer and the Concept Wiki: the Concept Wiki is a repository of community-editable concepts, currently relating to people and proteins, stored as RDF triples and fronted by a wiki-like interface (http://www.conceptwiki.org). Its associated ACKnowledge Enhancer is an analysis tool that links HTML content to relevant objects in the Concept Wiki and other online sources, exposing these to the user as selectable HTML highlights that, when activated, generate dynamic pop-ups. As with the Reflect plugin, integration with these systems via their web services provides a straightforward way of migrating functionality previously only available for HTML content to scientific PDF articles.

Videos showing these plugins in use are available at http://getutopia.com.

Utopia Documents is at an early stage of development and there is more work to be done. In the future, as well as opening its APIs to other developers, we plan to extend its scope 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 (Smith *et al*., 2007). Furthermore, the growing impetus of 'institutional repositories' as vehicles for collecting and sharing scholarly publications and data, and an increase in the acceptance of open access publishing, together present many interesting possibilities that we are keen to explore.

Another planned extension is to allow readers to append annotations and notes/comments to articles. There are various 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; (iii) a reader might wish to append notes to be shared with all subsequent readers of the article (e.g., because the paper describes an exciting breakthrough or because it contains an error)—these scenarios involve different security issues, and hence we will need to investigate how to establish appropriate 'webs of trust'. Ultimately, allowing users to append their own annotations (in addition to those endorsed by publishers) should help to involve authors in the manuscript mark-up process.

Utopia Documents brings us a step closer to integrated scholarly literature and research data. The software is poised to make contributions in a number of areas: for publishers, it offers a mechanism for adding value to oft-neglected PDF archives; for scientists whose routine work involves having to attach meaning to raw data from high-throughput biology experiments (database curators, bench biologists, researchers in pharmaceutical companies, etc.), it provides seamless links between facts published in articles, information deposited in databases and the requisite interactive tools to analyze and verify them; for readers in general, it provides both an enhanced reading experience and exciting new opportunities for knowledge discovery and 'community peer review'.

ACKNOWLEDGEMENTS

We thank all Portland Press staff for helping to realize the Semantic *BJ*, and, in particular, Rhonda Oliver and Audrey McCulloch for their courage, patience and positive collaboration. For their help and guidance in developing interfaces and plugins to their software, we also thank: Gert Vriend and Bas Vroling (GPCRDB); Barend Mons, Jan Velterop, Hailiang Mei (Concept Wiki); and Lars Juhl Jensen and Sean O'Donoghue (Reflect).

Funding: Portland Press Limited (The Semantic *Biochemical Journal* project) (Utopia Documents); European Union (EMBRACE, grant LHSG-CT-2004-512092); Biotechnology and Biological Sciences Research Council (Target practice, grant BBE0160651); Engineering and Physical Sciences Research Council (Doctoral Training Account).

Conflict of Interest: none declared.

REFERENCES

Attwood,T.K. *et al.* (2009) Calling international rescue – knowledge lost in literature and data landslide! *Biochem. J.*, **424**, 317–333.

Auer,S. *et al*. (2007) DBpedia : a nucleus for a web of open data. In Aberer,K. *et al*. (eds) *The Semantic Web.* Springer, Berlin/Heidelberg, pp.722–735.

Bairoch,A. (2009) The future of annotation/biocuration. *Nature Precedings* [Epub ahead of print, doi:10.1038/npre.2009.3092.1].

Belleau,F. *et al*. (2008) Bio2RDF: a semantic web atlas of post genomic knowledge about human and mouse. In Istrail,S. *et al*. (eds) *Data Integration in the Life Sciences.* Springer, Berlin/Heidelberg, pp. 153–160.

Boeckmann,B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Bourne,P. (2005) Will a Biological Database Be Different from a Biological Journal? *PLoS Comput. Biol.*, **1**, e34.

Chatr-aryamontri,A. *et al*. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res*., **35**, D572–D574.

Ceol,A. *et al*. (2008) Linking entries in protein interaction database to structured text: The FEBS Letters experiment. *FEBS Letters*, **582**, 1171–1177.

Editorial (2007) ALPSP/Charlesworth Awards 2007. *Learn. Pub.*, **20**, 317–318.

Fink,J.L. *et al*. (2008) BioLit: integrating biological literature with databases. *Nucleic Acids Res.*, **36**, W385–W389.

Giles,J., (2005) Internet encyclopaedias go head to head. *Nature*, **438**, 900–901.

Giles,J., (2006) Statistical flaw trips up study of bad stats. *Nature*, **443**, 379.

Kanehisa,M. *et al*. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*., **38**, D355–D360.

Kouranov,A. *et al.* (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res*., **34**, D302–D305.

Kumar,P. *et al*. (2008) Grasping molecular structures through publication-integrated 3D models. *Trends Biochem. Sci*., **33**, 408–412.

Lee,J.H. *et al*. (2008) Interaction between fortilin and transforming growth factor-beta stimulated clone-22 (TSC-22) prevents apoptosis via the destabilization of TSC-22. *FEBS Letters*, **582**, 1210–1218.

Lynch,C. (2007) The shape of the scientific article in developing cyberinfrastructure. *CTWatch Q.*, **3**, 5–10.

Pafilis,E. *et al*. (2009) Reflect: augmented browsing for the life scientist. *Nat. Biotechnol*., **27**, 508–510.

Pettifer,S.R. *et al*. (2004) UTOPIA - User-friendly Tools for OPerating Informatics Applications. *Comp. Funct. Genom.*, **5**, CFG359.

Pettifer,S., *et al*. (2009) Visualising biological data: a semantic approach to tool and database integration. *BMC Bioinformatics*, **10**, S19.

Renear,A.H. and Palmer,C.L. (2009) Strategic reading, ontologies, and the future of scientific publishing. *Science*, **325**, 828–832.

Ruthensteiner,B. and Hess,M. (2008) Embedding 3D models of biological specimens in PDF publications. *Microsc Res Tech*., **71**, 778–786.

Seringhaus,M.R. and Gerstein,M.B. (2007) Publishing perishing? Towards tomorrow's information architecture. *BMC Bioinformatics*, **8**,17.

Shotton,D. *et al*. (2009) Adventures in semantic publishing: exemplar semantic enhancements of a research article. *PLoS Comput. Biol.*, **5**, e1000361.

Smith,B. *et al*. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.

The UniProt Consortium (2009) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **37**, D169–D174.

Wilbanks,J. (2007) Cyberinfrastructure for knowledge sharing. *CTWatch Q.*, **3**, 58–66.