Utopia Documents: transforming how industrial scientists interact with the scientific literature

Steve Pettifer, Terri Attwood, James Marsh and Dave Thorne

Abstract: The number of articles published in the scientific literature continues to grow at an alarming rate, making it increasingly difficult for researchers to find key facts. Although electronic publications and libraries make it relatively easy to access articles, working out which ones are important, how their content relates to the growing number of online databases and resources, and making this information easily available to colleagues remains challenging. We present Utopia Documents, a visualization tool that provides a fresh perspective on the scientific literature.

Key words: semantic publishing; visualization; PDF; collaborative systems.

It is a curious fact that, throughout history, man's capacity to create knowledge has far outstripped his ability to disseminate, assimilate, and exploit that knowledge effectively. Until recently, the bottleneck was very much a physical one, with hand-written scrolls, and then printed manuscripts and books, being deposited in libraries to which access was limited, either by politics, social factors, or simply the need to physically travel to a particular location in order to access the content. Digital technologies have not solved the problem – and, it could be argued, have in some ways made matters worse.

Published by Woodhead Publishing Limited, 2012

Although it is no longer necessary to visit a library in person to read a scholarly work, so much material now exists that making sense of what is relevant is becoming increasingly difficult. In the life sciences alone, a new article is published every two minutes [1], and the number of articles has been growing exponentially at around 4% per year for the past 20 years [2]. In 2011, for example, over 14 000 papers were published relating in some way to Human Immunodeficiency Virus (admittedly, this is a much broader topic than any one researcher would care to explore in detail; nevertheless, it gives some indication of the scale of the problem) [3]. If we add to the peer-reviewed literature all the information that is deposited in databases, or written informally on personal blogs, matters get more complex still. The problem has been characterized, at a broad level, as being one of 'filter failure' [4], that is it's not that there is too much information per se – as, surely, lots of information has to be a good thing, and there has always been more available than we are able to consume – but, rather, that we are lacking mechanisms to convert information into meaningful knowledge, to find what is relevant to our particular needs, and to make connections between the disjoint data.

The problem is particularly acute in the life sciences. As pharmaceutical companies race to find new drugs, it has become clear that mastery of the literature is at least as important as mastery of the test tube or the sequencing machine. It has been documented that some failures of pharmaceutical company projects – which are costly on a mind-boggling scale, and typically result in failure – could have been avoided if scientists within the company had simply been able to find the relevant scientific article beforehand [5]. This inability to find the 'right article' is not for the want of trying: pharmaceutical companies employ large teams of scientists to scour the literature for just such information; it is simply that, even with the best tools and the most proficient scientists, there is more information 'out there' than can currently be processed in a timely way.

The danger of characterizing this as simple 'filter failure' is that it makes the solution sound trivial: get a better 'filter', and the problem goes away. The difficulty, of course, is that no filter yet exists that is sophisticated enough to cope with the complexities of life science data and the way in which it is described in 'human-readable' form. The simple key word searches or faceted browsing techniques that work so well for buying consumer goods or for booking a holiday online, break down quickly when faced with complex, context-sensitive, inter-related, incomplete and often ambiguous scientific data and articles. In many cases information gleaned from articles is already known in some part of an organization, but its relevance to some other new project is not discovered until after the fact. For organizations such as pharmaceutical companies, as the volume of information grows, the problem of 'knowing what we already know' is set to become worse [6].

At the heart of this problem lies a simple truth. We are now completely reliant on machines to help process the knowledge we generate; but humans and machines speak radically different languages. To perform efficiently, machines require formal, unambiguous and rigorous descriptions of concepts, captured in machine-readable form. Humans, on the other hand, need the freedom to write 'natural language', with all its associated ambiguities, nuances and subtleties, in order to set down and disseminate ideas. The trouble is, other than trained 'ontologists', mathematicians, and logicians, very few scientists can turn complex thoughts into the kind of formal, mathematical representations that can be manipulated computationally; and very few ontologists, mathematicians, and logicians understand biology and chemistry in enough detail to capture leading-edge thinking in the pharmaceutical or life sciences. For the time being, at least, we are faced with the challenge of bridging this divide.

Text- and data-mining techniques have made some progress here, attempting to extract meaningful assertions from written prose, and to capture these in machine-readable form (e.g. [7]). But this is far from being a solved problem – the vagaries of natural-language processing, coupled with the complex relationships between the terminology used in the life sciences and records in biomedical databases make this a non-trivial task.

In recent years, in what we consider to be a frustrated and frustrating distraction from the real issues, many have pointed the finger of blame at the file format used by publishers to distribute scientific articles: Adobe's PDF. It has variously been described as 'an insult to science', 'antithetical to the spirit of the web', and like 'building a telephone and then using it to transmit Morse Code', as though somehow the format itself is responsible for preventing machines from accessing its content [8]. Although it is true that extracting text from PDFs is a slightly more unwieldy process than it is from some other formats, such as XHTML, to accuse the PDF as the culprit is to entirely miss the point: the real problem arises from the gulf between human- and machine-readable language, not merely from the file-format in which this natural language is being stored. It is worth setting the record straight on two PDF-related facts.

1. Although originating in a commercial environment, PDF is now – and has been for some considerable time – an open format. The specifications of the PDF have been publically available for many years, and have been formally 'public domain' since 2008, when Adobe released the format as ISO 32000-1 and officially waived any patents relating to creating or consuming PDF content [9]. It is, however, undoubtedly a more verbose, and arguably less-elegant format than more recent offerings.
2. The PDF has the potential to be as 'semantically rich' as any of the other alternatives (e.g. eBook, XHTML), and, since version 1.4 in 2001, has had the facility to include arbitrarily sophisticated meta-data. The fact that virtually no publishers use this feature is not a fault of the PDF, and such semantic richness is just as conspicuously absent from the other formats used to publish scientific material. Speculation as to which link in the chain between author, publisher, and consumer causes this omission, we leave entirely to the reader.

Regardless of whether the PDF is a suitable vehicle for disseminating scientific knowledge, there are two irrefutable facts: first, it is by far the most popular medium by which scientific articles are consumed (accounting for over 80% of downloaded articles); and second, even if publishers move to a more 'web friendly' format at some point in the future, vast numbers of legacy articles will remain only as PDFs (many of which are digital scans of back-catalogues that pre-date even this ancient format, some of which have had text extracted via optical character recognition, while many exist only as images) – at the very least, these would require converting into something more 'semantic'. In spite of significant effort by publishers to enhance their online/HTML content with additional links, meta-data, data and multimedia extensions, the PDF stubbornly remains the format of choice for most readers. The recent growth in modern tools for managing collections of PDFs (e.g. Mendeley [10] and Papers [11]) gives additional backing to the view that the PDF is not in any danger of disappearing yet.

In an attempt to bridge the divide between what scientists actually read, and what computers need to support the process of taming the scientific literature, we embarked on a project to build a new tool for linking the content of articles with research data. The resulting software – Utopia Documents [12, 13] – combines all the advantages of the PDF with the interactivity and interconnectedness of a blog or web page. Acting as an alternative desktop application to Adobe Acrobat and other

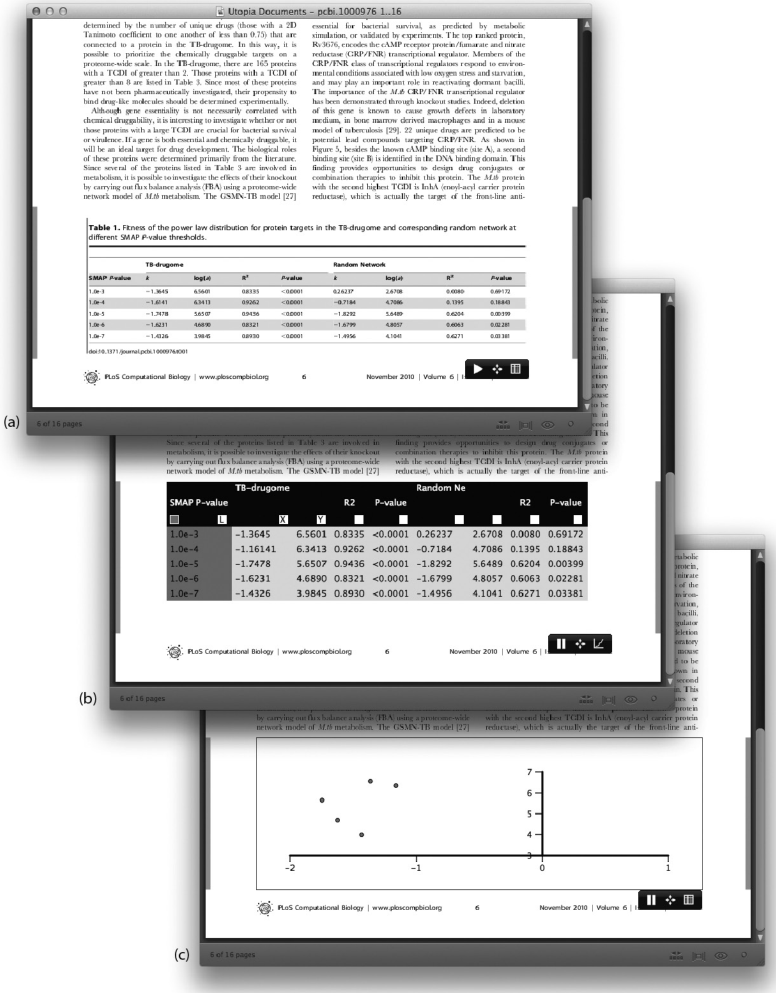
PDF viewers, Utopia Documents forms two-way links between the traditional human-readable PDF and any available semantic models produced by publishers or the community, and public or in-house data sources. The act of reading an article triggers heuristics that attempt to automatically identify features 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, key words, sections, figures, etc.) by their typographical layout, and ignoring less-important artifacts and non-document content (such as watermarks and additional information in the header or margins). Having gleaned this information from the PDF, Utopia Documents is then able to automatically link citations and references to online repositories.

As well as creating inbound and outbound links, Utopia Documents 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 (Figure 15.1); it can generate charts – already linked to the source data – on the fly; it can turn objects and images (e.g. protein sequences and 3D 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 to 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. Although it is possible for journal publishers to enhance PDFs with content that Utopia can recognize (noted below), it is important to state that the majority of Utopia's features work on all modern scientific PDFs, and are not limited to only a few specialized documents.

15.1 Utopia Documents in industry

In a commercial context, Utopia Documents aims to provide a new way to enable 'joined-up thinking' within institutions, reducing the cost of relating knowledge from within scientific articles to existing



determined by the number of unique drugs (those with a 2D Tanimoto coefficient to one another of less than 0.75) that are connected to a protein in the TB-drugome. In this way, it is possible to prioritize the chemically druggable targets on a proteome-wide scale. In the TB-drugome, there are 165 proteins with a TCDI of greater than 2. Those proteins with a TCDI of greater than 8 are listed in Table 3. Since most of these proteins have not been pharmaceutically investigated, their propensity to bind drug-like molecules should be determined experimentally.

Although gene essentiality is not necessarily correlated with chemical druggability, it is interesting to investigate whether or not those proteins with a large TCDI are crucial for bacterial survival or virulence. If a gene is both essential and chemically druggable, it will be an ideal target for drug development. The biological roles of these proteins were determined primarily from the literature. Since several of the proteins listed in Table 3 are involved in metabolism, it is possible to investigate the effects of their knockout by carrying out flux balance analysis (FBA) using a proteome-wide network model of *M.tb* metabolism. The GSMN-TB model [27] essential for bacterial survival, as predicted by metabolic simulation, or validated by experiments. The top ranked protein, Rv3676, encodes the cAMP receptor protein/fumarate and nitrate reductase (CRP/FNR) transcriptional regulator. Members of the CRP/FNR class of transcriptional regulators respond to environmental conditions associated with low oxygen stress and starvation, and may play an important role in reactivating dormant bacilli. The importance of the *M.tb* CRP/FNR transcriptional regulator has been demonstrated through knockout studies. Indeed, deletion of this gene is known to cause growth defects in laboratory medium, in bone marrow derived macrophages and in a mouse model of tuberculosis [29]. 22 unique drugs are predicted to be potential lead compounds targeting CRP/FNR. As shown in Figure 5, besides the known cAMP binding site (site A), a second binding site (site B) is identified in the DNA binding domain. This finding provides opportunities to design drug conjugates or combination therapies to inhibit this protein. The *M.tb* protein with the second highest TCDI is InhA (enoyl-acyl carrier protein reductase), which is actually the target of the front-line anti-

Table 1. Fitness of the power law distribution for protein targets in the TB-drugome and corresponding random network at different SMAP *P*-value thresholds.

	TB-drugome				Random Network			
SMAP *P*-value	*k*	log(*a*)	R^2	*P*-value	*k*	log(*a*)	R^2	*P*-value
1.0e-3	−1.3645	6.5601	0.8335	<0.0001	0.26237	2.6708	0.0080	0.69172
1.0e-4	−1.6141	6.3413	0.9262	<0.0001	−0.7184	4.7086	0.1395	0.18843
1.0e-5	−1.7478	5.6507	0.9436	<0.0001	−1.8292	5.6489	0.6204	0.00399
1.0e-6	−1.6231	4.6890	0.8321	<0.0001	−1.6799	4.8057	0.6063	0.02281
1.0e-7	−1.4326	3.9845	0.8930	<0.0001	−1.4956	4.1041	0.6271	0.03381

doi:10.1371/journal.pcbi.1000976.t001

PLoS Computational Biology | www.ploscompbiol.org 6 November 2010 | Volume 6 | I

	TB–drugome				Random Ne			
SMAP P–value			R2	P–value			R2	P–value
☐ L	X	Y	☐	☐	☐	☐	☐	☐
1.0e-3	–1.3645	6.5601	0.8335	<0.0001	0.26237	2.6708	0.0080	0.69172
1.0e-4	–1.16141	6.3413	0.9262	<0.0001	–0.7184	4.7086	0.1395	0.18843
1.0e-5	–1.7478	5.6507	0.9436	<0.0001	–1.8292	5.6489	0.6204	0.00399
1.0e-6	–1.6231	4.6890	0.8321	<0.0001	–1.6799	4.8057	0.6063	0.02281
1.0e-7	–1.4326	3.9845	0.8930	<0.0001	–1.4956	4.1041	0.6271	0.03381

Figure 15.1 A static PDF table (a) is converted first into an editable ‘spreadsheet’ (b), and then into a dynamic scatterplot (c)

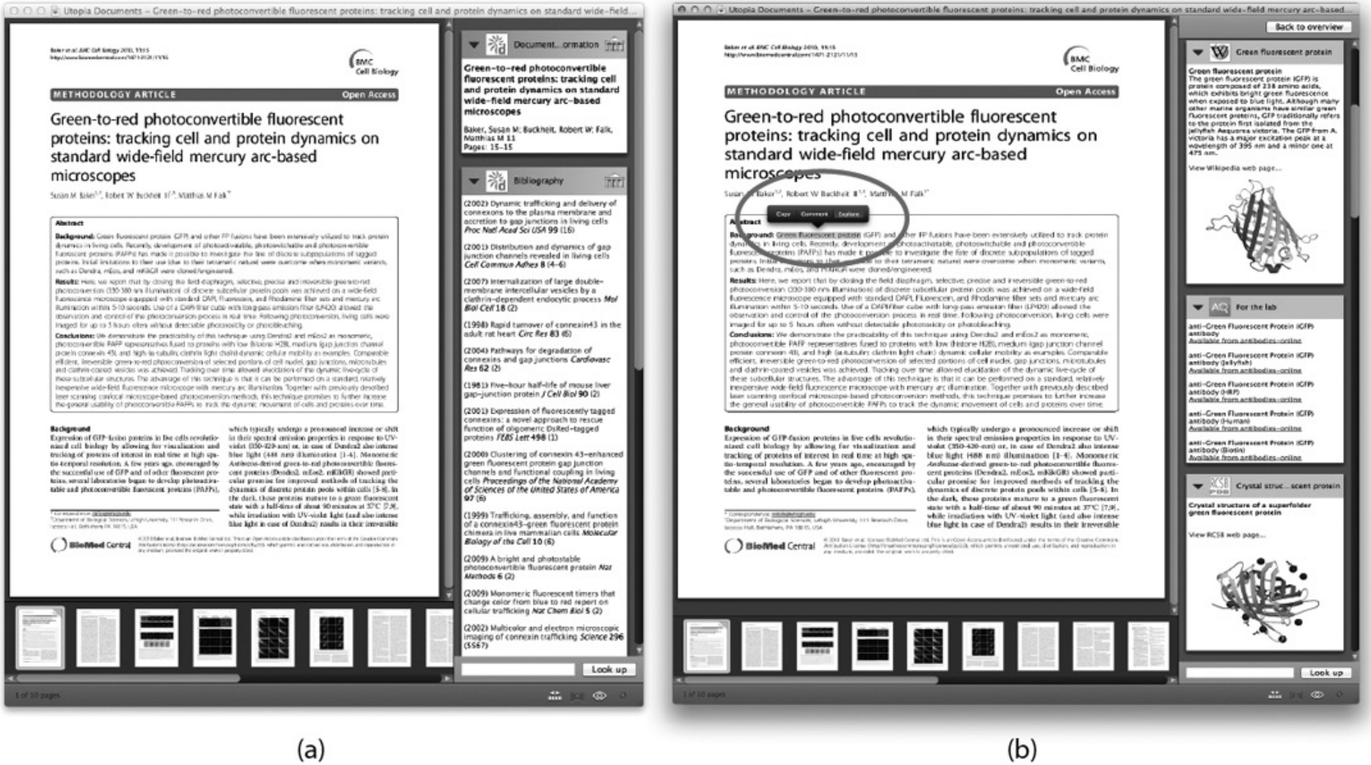
in-house and public data, and at the same time supporting auditable interaction with the literature. This results in simplifying the ‘due diligence’ processes essential in modern drug discovery. Other ‘knowledge management’ tools typically provide web-based interfaces that allow scientists to explore their content, but treat the scientific article itself as a distinct and separate entity, somewhat divorced from the discovery process. The consequence of this approach is that scientists are required to repeatedly switch from one application to another while pursuing a line of thinking that requires access both to the literature and to data. Often the situation is worse, and they are forced to print documents out, to scribble notes in the margin and thereby remove the potential for interconnected data completely. Not only does the act of consciously switching require a certain amount of mental inertia (leading inevitably to, 'Oh, I'll look that up next time when I'm using Application X'), it interrupts the workflow and thought processes associated with reading the literature. In turn, this makes documenting associations between concepts found in articles and data recorded in other systems difficult and fragmented.

Consider a typical industry use-case: a scientist is reading a paper describing a number of proteins and/or small molecules relevant to a disease or phenotype. Some of these might be new to the reader, who may wish to know a little more about them in order to better understand the findings in the paper. Traditionally, manual copy/pasting of protein and drug names into numerous search engines is the only solution. For compounds, this can also require tedious manual sketching of the molecule into a structure editor. The difficulty in performing what should be very simple tasks often results in either lost time or simply abandoning the enquiry.

Now, consider these same scenarios using Utopia Documents. Simply selecting the protein name using the cursor brings up a rich information panel about that entity (Figure 15.2). Alternatively, text-mining services can be invoked directly from within the Utopia environment to automatically identify interesting entities and relationships, and to link these to in-house data. The result is a highly visual and organized index of the genes, compounds, and diseases mentioned in the text. For compounds, Utopia goes even further, as Figure 15.3 shows. Here, the software has identified a number of small molecules mentioned only by name in the text of the paper. Using (in this example) the ChemSpider [14] API, the actual structures have been retrieved, along with other information, providing a completely new chemical index browser to the paper. Functionality developed as part of the EU Open PHACTS project [15] also allows the retrieval of activity data from public and corporate structure–activity relationship databases.

Finally, because of Utopia's unique software architecture, many other types of plug-in can be developed, including the ability to interact with tables and to embed applications, such as 3D protein structure viewers, directly in the document. Utopia provides a two-way API, allowing web services to retrieve structured content from the PDF ('give me all tables

Published by Woodhead Publishing Limited, 2012



Utopia Documents – Green-to-red photoconvertible fluorescent proteins: tracking cell and protein dynamics on standard wide-field mercury arc-based...

Baker et al. BMC Cell Biology 2010, 11:15
http://www.biomedcentral.com/1471-2121/11/15

BMC Cell Biology

METHODOLOGY ARTICLE Open Access

Green-to-red photoconvertible fluorescent proteins: tracking cell and protein dynamics on standard wide-field mercury arc-based microscopes

Susan M Baker, Robert W Buckheit III, Matthias M Falk

Abstract

Background: Green fluorescent protein (GFP) and other FP fusions have been extensively utilized to track protein dynamics in living cells. Recently, development of photoactivatable, photoswitchable and photoconvertible fluorescent proteins (PAFPs) has made it possible to investigate the fate of discrete subpopulations of tagged proteins. Initial limitations to their use (due to their tetrameric nature) were overcome when monomeric variants, such as Dendra, mEos, and mKikGR were cloned/engineered.

Results: Here, we report that by closing the field diaphragm, selective, precise and irreversible green-to-red photoconversion (330-380 nm illumination) of discrete subcellular protein pools was achieved on a wide-field fluorescence microscope equipped with standard DAPI, Fluorescein, and Rhodamine filter sets and mercury arc illumination within 5-10 seconds. Use of a DAPI-filter cube with long-pass emission filter (LP420) allowed the observation and control of the photoconversion process in real time. Following photoconversion, living cells were imaged for up to 5 hours often without detectable phototoxicity or photobleaching.

Conclusions: We demonstrate the practicability of this technique using Dendra2 and mEos2 as monomeric, photoconvertible PAFP representatives fused to proteins with low (histone H2B), medium (gap junction channel protein connexin 43), and high (α-tubulin; clathrin light chain) dynamic cellular mobility as examples. Comparable efficient, irreversible green-to-red photoconversion of selected portions of cell nuclei, gap junctions, microtubules and clathrin-coated vesicles was achieved. Tracking over time allowed elucidation of the dynamic live-cycle of these subcellular structures. The advantage of this technique is that it can be performed on a standard, relatively inexpensive wide-field fluorescence microscope with mercury arc illumination. Together with previously described laser scanning confocal microscope-based photoconversion methods, this technique promises to further increase the general usability of photoconvertible PAFPs to track the dynamic movement of cells and proteins over time.

Background

Expression of GFP-fusion proteins in live cells revolutionized cell biology by allowing for visualization and tracking of proteins of interest in real time at high spatio-temporal resolution. A few years ago, encouraged by the successful use of GFP and of other fluorescent proteins, several laboratories began to develop photoactivatable and photoconvertible fluorescent proteins (PAFPs), which typically undergo a pronounced increase or shift in their spectral emission properties in response to UV-violet (350-420-nm) or, in case of Dendra2 also intense blue light (488 nm) illumination [1-4]. Monomeric *Anthozoa*-derived green-to-red photoconvertible fluorescent proteins (Dendra2, mEos2, mKikGR) showed particular promise for improved methods of tracking the dynamics of discrete protein pools within cells [5-8]. In the dark, these proteins mature to a green fluorescent state with a half-time of about 90 minutes at 37°C [7,9], while irradiation with UV-violet light (and also intense blue light in case of Dendra2) results in their irreversible

BioMed Central

Document...ormation

Green-to-red photoconvertible fluorescent proteins: tracking cell and protein dynamics on standard wide-field mercury arc-based microscopes

Baker, Susan M; Buckheit, Robert W; Falk, Matthias M 11
Pages: 15–15

Bibliography

(2002) Dynamic trafficking and delivery of connexons to the plasma membrane and accretion to gap junctions in living cells *Proc Natl Acad Sci USA* **99** (16)

(2001) Distribution and dynamics of gap junction channels revealed in living cells *Cell Commun Adhes* **8** (4-6)

(2007) Internalization of large double-membrane intercellular vesicles by a clathrin-dependent endocytic process *Mol Biol Cell* **18** (2)

(1998) Rapid turnover of connexin43 in the adult rat heart *Circ Res* **83** (6)

(2004) Pathways for degradation of connexins and gap junctions *Cardiovasc Res* **62** (2)

(1981) Five-hour half-life of mouse liver gap-junction protein *J Cell Biol* **90** (2)

(2001) Expression of fluorescently tagged connexins: a novel approach to rescue function of oligomeric DsRed-tagged proteins *FEBS Lett* **498** (1)

(2000) Clustering of connexin 43–enhanced green fluorescent protein gap junction channels and functional coupling in living cells *Proceedings of the National Academy of Sciences of the United States of America* **97** (6)

(1999) Trafficking, assembly, and function of a connexin43–green fluorescent protein chimera in live mammalian cells *Molecular Biology of the Cell* **10** (6)

(2009) A bright and photostable photoconvertible fluorescent protein *Nat Methods* **6** (2)

(2009) Monomeric fluorescent timers that change color from blue to red report on cellular trafficking *Nat Chem Biol* **5** (2)

(2002) Multicolor and electron microscopic imaging of connexin trafficking *Science* **296** (5567)

Green fluorescent protein

Green fluorescent protein
The green fluorescent protein (GFP) is protein composed of 238 amino acids, which exhibits bright green fluorescence when exposed to blue light. Although many other marine organisms have similar green fluorescent proteins, GFP traditionally refers to the protein first isolated from the jellyfish Aequorea victoria. The GFP from A. victoria has a major excitation peak at a wavelength of 395 nm and a minor one at 475 nm.

View Wikipedia web page...

For the lab

anti-Green Fluorescent Protein (GFP) antibody
Available from antibodies-online

anti-Green Fluorescent Protein (GFP) antibody (Jellyfish)
Available from antibodies-online

anti-Green Fluorescent Protein (GFP) antibody (HRP)
Available from antibodies-online

anti-Green Fluorescent Protein (GFP) antibody (Human)
Available from antibodies-online

anti-Green Fluorescent Protein (GFP) antibody (Biotin)
Available from antibodies-online

Crystal struc...scent protein

Crystal structure of a superfolder green fluorescent protein

View RCSB web page...

Figure 15.2 In (a), Utopia Documents shows meta-data relating to the article currently being read; in (b), details of a specific term are displayed, harvested in real time from online data sources

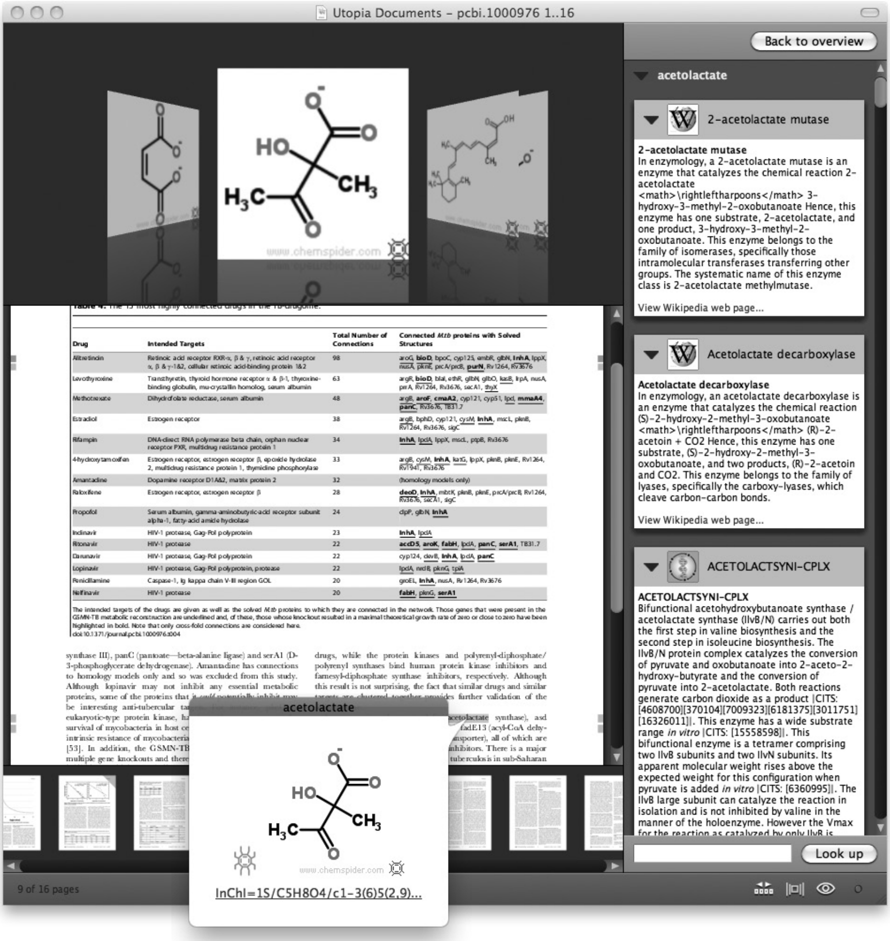


Table 4. The 15 most highly connected drugs in the TB-drugome.

Drug	Intended Targets	Total Number of Connections	Connected *Mtb* proteins with Solved Structures
Alitretinoin	Retinoic acid receptor RXR-α, β & γ, retinoic acid receptor α, β & γ-1&2, cellular retinoic acid-binding protein 1&2	98	aroG, **bioD**, bpoC, cyp125, embR, glbN, **InhA**, lppX, nusA, pknE, prcA/prcB, **purN**, Rv1264, Rv3676
Levothyroxine	Transthyretin, thyroid hormone receptor α & β-1, thyroxine-binding globulin, mu-crystallin homolog, serum albumin	63	argR, **bioD**, blaI, ethR, glbN, glbO, kasB, lrpA, nusA, perA, Rv1264, Rv3676, secA1, thyX
Methotrexate	Dihydrofolate reductase, serum albumin	48	argB, **aroF**, **cmaA2**, cyp121, cyp51, lpd, **mmaA4**, **panC**, Rv3676, TB31.7
Estradiol	Estrogen receptor	38	argB, bphD, cyp121, cysM, **InhA**, mscL, pknB, Rv1264, Rv3676, sigC
Rifampin	DNA-direct RNA polymerase beta chain, orphan nuclear receptor PXR, multidrug resistance protein 1	34	**InhA**, lpdA, lppX, mscL, ptpB, Rv3676
4-hydroxytamoxifen	Estrogen receptor, estrogen receptor β, epoxide hydrolase 2, multidrug resistance protein 1, thymidine phosphorylase	33	argB, cysM, **InhA**, katG, lppX, pknB, pknE, Rv1264, Rv1941, Rv3676
Amantadine	Dopamine receptor D1A&2, matrix protein 2	32	(homology models only)
Raloxifene	Estrogen receptor, estrogen receptor β	28	**deoD**, **InhA**, mbtK, pknB, pknE, prcA/prcB, Rv1264, Rv3676, secA1, sigC
Propofol	Serum albumin, gamma-aminobutyric-acid receptor subunit alpha-1, fatty-acid amide hydrolase	24	dipP, glbN, **InhA**
Indinavir	HIV-1 protease, Gag-Pol polyprotein	23	**InhA**, lpdA
Ritonavir	HIV-1 protease	22	**accD5**, **aroK**, **fabH**, lpdA, **panC**, **serA1**, TB31.7
Darunavir	HIV-1 protease, Gag-Pol polyprotein	22	cyp124, devB, **InhA**, lpdA, **panC**
Lopinavir	HIV-1 protease, Gag-Pol polyprotein, protease	22	lpdA, nrdB, pknG, tpiA
Penicillamine	Caspase-1, Ig kappa chain V-III region GOL	20	groEL, **InhA**, nusA, Rv1264, Rv3676
Nelfinavir	HIV-1 protease	20	**fabH**, pknG, **serA1**

The intended targets of the drugs are given as well as the solved *Mtb* proteins to which they are connected in the network. Those genes that were present in the GSMN-TB metabolic reconstruction are underlined and, of these, those whose knockout resulted in a maximal theoretical growth rate of zero or close to zero have been highlighted in bold. Note that only cross-fold connections are considered here.
doi:10.1371/journal.pcbi.1000976.t004

synthase III), panC (pantoate—beta-alanine ligase) and serA1 (D-3-phosphoglycerate dehydrogenase). Amantadine has connections to homology models only and so was excluded from this study. Although lopinavir may not inhibit any essential metabolic proteins, some of the proteins that it *could* potentially inhibit may be interesting anti-tubercular ta
eukaryotic-type protein kinase, ha
survival of mycobacteria in host ce
intrinsic resistance of mycobacteria
[53]. In addition, the GSMN-TB
multiple gene knockouts and there

drugs, while the protein kinases and polyrenyl-diphosphate/polyrenyl synthases bind human protein kinase inhibitors and farnesyl-diphosphate synthase inhibitors, respectively. Although this result is not surprising, the fact that similar drugs and similar targets are clustered together provides further validation of the
acetolactate synthase), asd
fadE13 (acyl-CoA dehy-
ansporter), all of which are
inhibitors. There is a major
tuberculosis in sub-Saharan

acetolactate

2-acetolactate mutase

2-acetolactate mutase
In enzymology, a 2-acetolactate mutase is an enzyme that catalyzes the chemical reaction 2-acetolactate <math>\rightleftharpoons</math> 3-hydroxy-3-methyl-2-oxobutanoate Hence, this enzyme has one substrate, 2-acetolactate, and one product, 3-hydroxy-3-methyl-2-oxobutanoate. This enzyme belongs to the family of isomerases, specifically those intramolecular transferases transferring other groups. The systematic name of this enzyme class is 2-acetolactate methylmutase.

View Wikipedia web page...

Acetolactate decarboxylase

Acetolactate decarboxylase
In enzymology, an acetolactate decarboxylase is an enzyme that catalyzes the chemical reaction (S)-2-hydroxy-2-methyl-3-oxobutanoate <math>\rightleftharpoons</math> (R)-2-acetoin + CO2 Hence, this enzyme has one substrate, (S)-2-hydroxy-2-methyl-3-oxobutanoate, and two products, (R)-2-acetoin and CO2. This enzyme belongs to the family of lyases, specifically the carboxy-lyases, which cleave carbon-carbon bonds.

View Wikipedia web page...

ACETOLACTSYNI-CPLX

ACETOLACTSYNI-CPLX
Bifunctional acetohydroxybutanoate synthase / acetolactate synthase (IlvB/N) carries out both the first step in valine biosynthesis and the second step in isoleucine biosynthesis. The IlvB/N protein complex catalyzes the conversion of pyruvate and oxobutanoate into 2-aceto-2-hydroxy-butyrate and the conversion of pyruvate into 2-acetolactate. Both reactions generate carbon dioxide as a product |CITS: [4608700][370104][7009323][6181375][3011751][16326011]|. This enzyme has a wide substrate range *in vitro* |CITS: [15558598]|. This bifunctional enzyme is a tetramer comprising two IlvB subunits and two IlvN subunits. Its apparent molecular weight rises above the expected weight for this configuration when pyruvate is added *in vitro* |CITS: [6360995]|. The IlvB large subunit can catalyze the reaction in isolation and is not inhibited by valine in the manner of the holoenzyme. However the Vmax for the reaction as catalyzed by only IlvB is

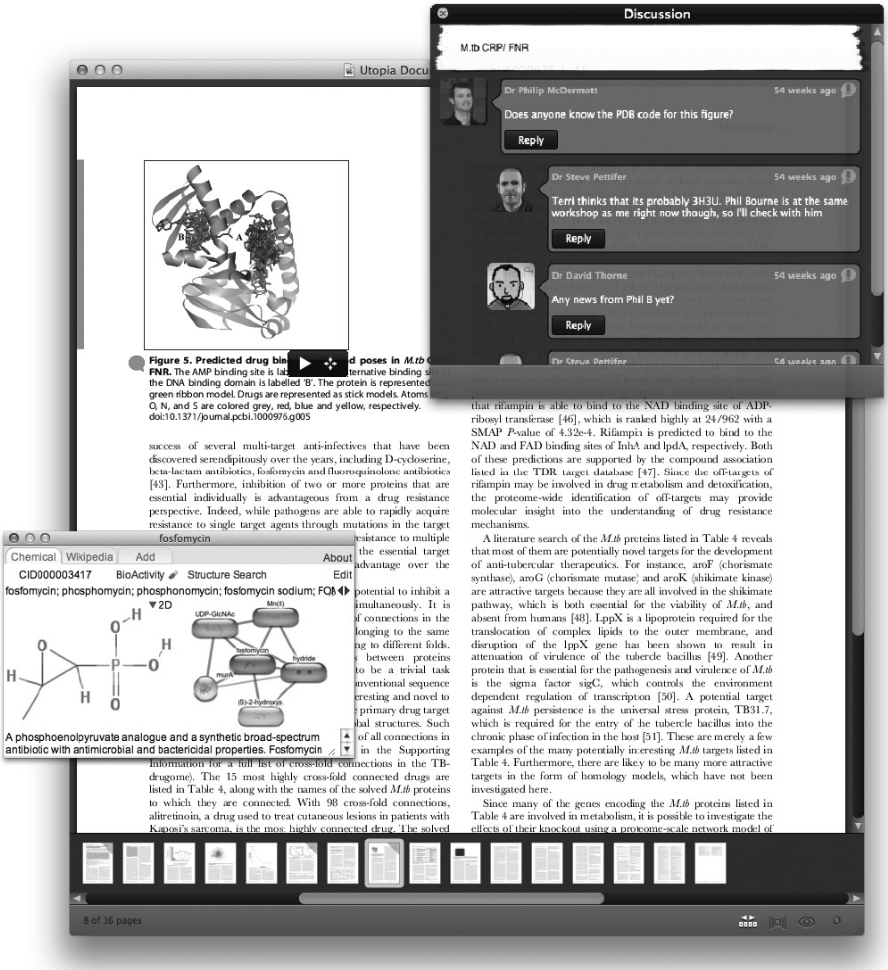
Figure 15.3 A text-mining algorithm has identified chemical entities in the article being read, details of which are displayed in the sidebar and top ‘flow browser’

from the document’) and inject content into the PDF (‘show these data when the user highlights a compound’). This provides many opportunities to interconnect with internal systems within a company. For instance, highlighting a protein might call for a look-up of the internal project-management system to list which groups are actively researching this target for relevant indications. Similarly, highlighting a compound might invoke a structure-search of the corporate small-molecule database, identifying structurally similar compounds in the internal file. By these means, Utopia connects the content of the PDF into a company's systems, finally joining one of the most important pieces of the scientific process to the existing information infrastructure.

15.2 Enabling collaboration

Utopia stands apart from other PDF-management tools in the way in which it promotes collaboration and sharing of knowledge within an institution. Most life science industry projects are run by dynamic, ever-changing teams of scientists, with new members being introduced as specialist expertise is required, and scientists typically contributing to many projects at any given time. A researcher faces a constant battle to stay on top of the literature in each project, determining which papers should be read and in what order, and exchanging thoughts with colleagues on the good and bad of what they have digested. Often the solution to both these needs is ad hoc email, reading papers suggested by colleagues and replying with relevant critique. Of course, this relies on the researcher being copied into the email thread and, assuming that this has happened, on finding the email in an ever-growing inbox (in all probability, emailing a published article to a colleague, even within the same organization, is likely to be a technical violation of the license under which the original reader obtained the article).

Utopia changes this by providing a team-based approach to literature management. Unlike other document-management tools, this enables collaboration not just at the level of a paper, but rather, at the level of a paper's content. Individuals can manage their own paper collections in a familiar and intuitive interface; however, by simply creating team folders and sharing these with colleagues, interesting papers and their related data and annotations can be circulated within relevant groups, without the need to remember to send that email and get everyone's name right. Simple icons inform team members of new papers in the collection (Figures 15.4 and 15.5). As well as comments, annotations on articles can include data-files (e.g. where the group has re-interpreted a large data set), or links to data held in knowledge-management systems. Thus, anyone reading the paper will see that there is additional information from inside the company that might make a real difference to their interpretation.



Discussion

M.tb CRP/ FNR

Dr Philip McDermott — 54 weeks ago

Does anyone know the PDB code for this figure?

Reply

Dr Steve Pettifer — 54 weeks ago

Terri thinks that its probably 3H3U. Phil Bourne is at the same workshop as me right now though, so I'll check with him

Reply

Dr David Thorne — 54 weeks ago

Any news from Phil B yet?

Reply

Dr Steve Pettifer — 54 weeks ago

Figure 5. Predicted drug bin… …d poses in *M.tb* C… FNR. The AMP binding site is lab… …ternative binding si… the DNA binding domain is labelled 'B'. The protein is represented… green ribbon model. Drugs are represented as stick models. Atoms… O, N, and S are colored grey, red, blue and yellow, respectively. doi:10.1371/journal.pcbi.1000976.g005

success of several multi-target anti-infectives that have been discovered serendipitously over the years, including D-cycloserine, beta-lactam antibiotics, fosfomycin and fluoroquinolone antibiotics [43]. Furthermore, inhibition of two or more proteins that are essential individually is advantageous from a drug resistance perspective. Indeed, while pathogens are able to rapidly acquire resistance to single target agents through mutations in the target

esistance to multiple
the essential target
advantage over the

potential to inhibit a
imultaneously. It is
f connections in the
longing to the same
ng to different folds.
s between proteins
to be a trivial task
nventional sequence
eresting and novel to
primary drug target
bal structures. Such
of all connections in
in the Supporting

Information for a full list of cross-fold connections in the TB-drugome). The 15 most highly cross-fold connected drugs are listed in Table 4, along with the names of the solved *M.tb* proteins to which they are connected. With 98 cross-fold connections, alitretinoin, a drug used to treat cutaneous lesions in patients with Kaposi's sarcoma, is the mos: highly connected drug. The solved

that rifampin is able to bind to the NAD binding site of ADP-ribosyl transferase [46], which is ranked highly at 24/962 with a SMAP *P*-value of 4.32e-4. Rifampin is predicted to bind to the NAD and FAD binding sites of InhA and lpdA, respectively. Both of these predictions are supported by the compound association listed in the TDR target database [47]. Since the off-targets of rifampin may be involved in drug metabolism and detoxification, the proteome-wide identification of off-targets may provide molecular insight into the understanding of drug resistance mechanisms.

A literature search of the *M.tb* proteins listed in Table 4 reveals that most of them are potentially novel targets for the development of anti-tubercular therapeutics. For instance, aroF (chorismate synthase), aroG (chorismate mutase) and aroK (shikimate kinase) are attractive targets because they are all involved in the shikimate pathway, which is both essential for the viability of *M.tb*, and absent from humans [48]. LppX is a lipoprotein required for the translocation of complex lipids to the outer membrane, and disruption of the lppX gene has been shown to result in attenuation of virulence of the tubercle bacillus [49]. Another protein that is essential for the pathogenesis and virulence of *M.tb* is the sigma factor sigC, which controls the environment dependent regulation of transcription [50]. A potential target against *M.tb* persistence is the universal stress protein, TB31.7, which is required for the entry of :he tubercle bacillus into the chronic phase of infection in the host [51]. These are merely a few examples of the many potentially in:eresting *M.tb* targets listed in Table 4. Furthermore, there are likey to be many more attractive targets in the form of homology models, which have not been investigated here.

Since many of the genes encoding the *M.tb* proteins listed in Table 4 are involved in metabolism, it is possible to investigate the effects of their knockout using a proteome-scale network model of

fosfomycin

CID000003417

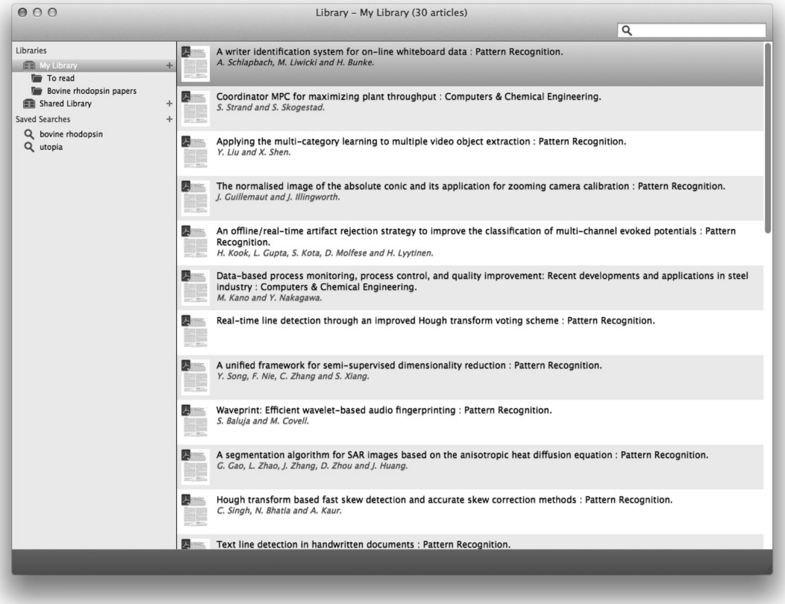
fosfomycin; phosphomycin; phosphonomycin; fosfomycin sodium; FO

A phosphoenolpyruvate analogue and a synthetic broad-spectrum antibiotic with antimicrobial and bactericidal properties. Fosfomycin

Figure 15.4 Comments added to an article can be shared with other users, without the need to share a specific copy of the PDF

15.3 Sharing, while playing by the rules

Unlike other tools, Utopia does not rely on users exchanging specific, annotated copies of PDF articles, or on the storage of PDFs in bespoke institutional document repositories (all of which raise legal questions around the storage and distribution of copyrighted PDFs and the breach



A writer identification system for on-line whiteboard data : Pattern Recognition.
A. Schlapbach, M. Liwicki and H. Bunke.

Coordinator MPC for maximizing plant throughput : Computers & Chemical Engineering.
S. Strand and S. Skogestad.

Applying the multi-category learning to multiple video object extraction : Pattern Recognition.
Y. Liu and X. Shen.

The normalised image of the absolute conic and its application for zooming camera calibration : Pattern Recognition.
J. Guillemaut and J. Illingworth.

An offline/real-time artifact rejection strategy to improve the classification of multi-channel evoked potentials : Pattern Recognition.
H. Kook, L. Gupta, S. Kota, D. Molfese and H. Lyytinen.

Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry : Computers & Chemical Engineering.
M. Kano and Y. Nakagawa.

Real-time line detection through an improved Hough transform voting scheme : Pattern Recognition.

A unified framework for semi-supervised dimensionality reduction : Pattern Recognition.
Y. Song, F. Nie, C. Zhang and S. Xiang.

Waveprint: Efficient wavelet-based audio fingerprinting : Pattern Recognition.
S. Baluja and M. Covell.

A segmentation algorithm for SAR images based on the anisotropic heat diffusion equation : Pattern Recognition.
G. Gao, L. Zhao, J. Zhang, D. Zhou and J. Huang.

Hough transform based fast skew detection and accurate skew correction methods : Pattern Recognition.
C. Singh, N. Bhatia and A. Kaur.

Text line detection in handwritten documents : Pattern Recognition.

Figure 15.5 Utopia Library provides a mechanism for managing collections of articles

of publishers' licenses). When a user opens a PDF file in Utopia, a small semantic fingerprint is created that uniquely identifies that paper to the system. Any annotations or attachments are associated with this fingerprint and not with the PDF file itself. This has significant consequences for sharing and copyright within corporate environments.

- If a scientist on the other side of the world goes to the journal web site and downloads a completely new PDF of a specific article, it has the same fingerprint as any other copy of that article (even if the PDF is watermarked by the digital delivery system to include provenance information). When opened in Utopia, this fingerprint is recognized, and all enhancements and company annotations are retrieved and overlaid on the article in real time. This means that colleagues who may not even know that this paper was of interest to someone in the company are able to exploit the tacit knowledge of the organization that would otherwise be lost.
- A corporate PDF repository is not required. When user A posts a paper to the shared folder, user B sees the title, citation and so on, but does not get a physical copy of the paper until they decide to read that

article. Then, a new copy of the PDF is automatically downloaded via the company's digital delivery service, allowing copyright to be respected and usage/access metrics to be accurately maintained.

15.4 History and future of Utopia Documents

Utopia Documents was originally developed in collaboration with Portland Press Ltd, who have been using the software since December 2009 to enhance the content of their flagship publication, the *Biochemical Journal* [16]. A publishers' version of the software allows 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 markup (typically, 10–15 minutes per article, depending on its suitability for annotation) as part of the normal publishing process. To date, around 1000 articles have been annotated in this way as part of the *Semantic Biochemical Journal* (*http://www.biochemj.org*). More recently, the publications of the Royal Society of Chemistry [17] have been augmented by automatically generated annotations on the chemical compounds mentioned in articles, and a plug-in to highlight these in Utopia Documents is included in the latest releases (covering around 50 000 articles, at the time of writing). Other extensions include linking articles automatically to their Altmetric [18] statistics (giving an indication of the popularity of an article based on its citation in blogs, tweets, and other social media), and to online data repositories such as Dryad [19].

Recently, we have engaged in projects with pharmaceutical companies to install the system within an enterprise environment. Utopia consists of two main components, the PDF viewer installed on any user's PC, and a server, which holds the annotations, fingerprints, and other essential meta-data. Installation of this server behind the corporate firewall provides the security and privacy often required by commercial organizations. In addition, the system will then have access to secure in-house text-mining and data services. Ultimately, this allows commercial entities to use the software without fear of their activities being tracked by competitors, or their sensitive in-house knowledge 'leaking' into the public domain. We have explored two major approaches to this configuration: the traditional single installation on a company server, and a cloud-based approach. For the latter, a service provider hosts a secure, tailored instance of the server to which only users from a specific company have access. All of the advantages of an internal installation are preserved, but with much less overhead in maintaining the system. Our experience of the use of Utopia in industry has suggested that the team-based sharing is of greatest initial interest. Industry scientists urgently need tools to help them manage information more effectively, and to manage it as a team.

Deciding on an appropriate license for Utopia Documents has been an interesting and complex endeavor. With the vast majority of its funding so far having come from commercial bodies (primarily publishers and the pharmaceutical industry), rather than from research grants or the public purse, we do not feel the same ethical imperative to make the software open source as we might do had the development been funded through grants derived from public funds. Yet, as academics, we nevertheless find openness extremely attractive. The difficulty arises when the ideals of openness meet the reality of finding sustainable income streams that will allow us to continue the research and development necessary to maintain Utopia Documents as a useful and powerful tool for scientists, whether in commercial or academic environments. What has become clear is that very few of our end-users care about the detailed licensing of the software, as long as the tool itself is robust, reliable, freely available, and has some plausible longevity.

Although it is notionally possible to sustain software development and research on the back of 'consultancy' associated with open-source software, there appear to be relatively few concrete examples of where this has worked as a long-term solution for keeping a research team together. Our current model, therefore, is one in which the core software is – and will remain – free to use without restriction (in both academic and commercial settings), while we retain the rights to customize the tool for commercial environments, or to provide specific services to publishers. As academics, finding a model that both sustains our research and provides a freely available resource for the community and 'the good of science' has been challenging on many levels; we are nevertheless optimistic that our current approach provides a reasonable balance between long-term sustainability and free availability, and are excited about the potential of Utopia Documents as a means of making more of the scientific literature.

15.5 References

[1] Attwood TK, Kell DB, Mcdermott P, Marsh J, Pettifer SR and Thorne D. Calling international rescue: knowledge lost in literature and data landslide! *Biochemical Journal*, Dec 2009.

[2] Lu Z. '*PubMed and beyond: a survey of web tools for searching bio-medical literature*.' Database 2011: baq036.

[3] Statistics generated Jan 2012 by searching *http://www.ncbi.nlm.nih.gov/pubmed/* for articles mentioning HIV, published in 2011.

[4] *http://blip.tv/web2expo/web-2-0-expo-ny-clay-shirky-shirky-com-it-s-not-information-overload-it-s-filter-failure-1283699*

[5] Korstanje C. Integrated assessment of preclinical data: shifting high attrition rates to earlier phase drug development. *Current Opinion in Investigational Drugs* 2003;4(5):519–21.

[6] Superti-Furga G, Wieland F and Cesareni G. 'Finally: The digital, democratic age of scientific abstracts.' *FEBS Letters* 2008;582(8):1169.

[7] Ananiadou S and McNaught J (Eds) *Text Mining for Biology and Biomedicine*. Artech House Books. ISBN 978-1-58053-984-5, 2006

[8] Pettifer S, McDermott P, Marsh J, Thorne D, Villéger A and Attwood TK. Ceci n'est pas un hamburger: modelling and representing the scholarly article. *Learned Publishing*, Jul 2011.

[9] International Organization for Standardization *Document management — Portable document format — Part 1*: PDF 1.7, ISO3200-1:2008.

[10] *www.mendeley.com*

[11] *http://www.mekentosj.com/papers*

[12] *www.utopiadocs.com*

[13] Attwood TK, Kell DB, Mcdermott P, Marsh J, Pettifer SR and Thorne D. Utopia Documents: linking scholarly literature with research data. *Bioinformatics* 2010;26:i540–6.

[14] *www.chemspider.com*

[15] *www.openphacts.org*

[16] *www.biochemj.org*

[17] *www.rsc.org*

[18] *www.altmetric.com*

[19] *www.datadryad.org*