Institution: University of Manchester

Unit of Assessment: UoA5

Title of case study: The Utopia Suite: realising semantic knowledge discovery and data linkage in the publishing and pharmaceutical industries

1. Summary of the impact

The need to manage, analyse and interpret the volumes of data and literature generated by modern high-throughput biology has become a major barrier to progress. Research at the University of Manchester on interoperability and advanced interfaces has resulted in innovative software (Utopia Documents) that links biomedical data with scientific literature. The software has been adopted by international publishing houses (Portland Press, Elsevier, Springer, *etc*.), allowing them to explore new business models, and by pharmaceutical companies (*e.g.* AstraZeneca, Roche), providing new opportunities to explore more efficient, cost-effective methods for exploiting and sharing in-house data and knowledge. The research also led to a spin-out company, Lost Island Labs, in 2012, which expects a profit [text removed for publication] in its first year.

2. Underpinning research

The underpinning research took place at the University of Manchester (UoM) from 2001 to date. The key researchers were:

Professor Teresa Attwood (2001 to date)
Dr Steve Pettifer (2001 to date)
Mr James Sinnott (2002-2007, Research Assistant)
Dr David Thorne (2009 to date, Post-Doctoral Research Associate; 2004-2009, PhD student; 2001-2004, BSc student)
Dr James Marsh (2008 to date, Post-Doctoral Research Associate)
Dr Phil McDermott (2011-2012, Post-Doctoral Research Associate; 2006-2011, PhD student)

The aim of the research was to develop easy-to-use software tools to facilitate protein sequence analysis. UoM researchers introduced the idea of tool and data integration, using novel semantic technologies; their subsequent innovation was to use these semantic approaches to enable tighter, bi-directional links between research data and scientific literature. At one end of the spectrum, important targets were genome/proteome annotation (especially functional characterisation of proteins of pharmaceutical interest); at the other, PDF articles, as vehicles for providing interactive access to biomedical data.

The key steps in the research were as follows:

1. From 2002, the researchers showed how semantic approaches could be used to integrate conventional bioinformatics tools. The first prototype, combining the functionality of a multiple protein sequence alignment editor with a 3D molecular viewer, was realised in the public release of the Utopia sequence analysis suite (2004) [1].
2. The research team went on to demonstrate the feasibility of using Web services for semantic integration of disparate tools and data-sets within the Utopia framework [2].
3. In 2007, Utopia came to the attention of the managing director of Portland Press Limited (PPL), who needed to add value to their static online content. With support from PPL, the researchers demonstrated the feasibility of integrating Utopia's interactive tools directly with PDF articles, thereby creating Utopia Documents [3,4].
4. The researchers went on to demonstrate the feasibility of deploying the Utopia Suite behind security-controlled company firewalls.
5. A subsequent collaboration with Bio-Prodict, a Dutch computational biotech company, saw the release of bespoke, pharmaceutically-relevant database-plugins for Utopia Documents [5,6].

This work remains highly productive, with numerous on-going collaborations with the publishing and pharmaceutical industries, as described in Section 4.

3. References to the research

The research was published in leading bioinformatics journals (*Bioinformatics, BMC Bioinformatics*), including the top journal in the field for database publications (*Nucleic Acids Research)*. Utopia Documents launch article in the first issue of the Semantic *Biochemical Journal (BJ)* [3] received, in its first 4 months from publication, 700 PDF downloads and 658 full-text page-views (with 2,808 downloads, it was the third most downloaded *BJ* paper a year after publication; now, with >8,000 downloads, it is the 11th most downloaded paper since November 2009); the article was highlighted in Garten and Altman's *Future Medicine* Editorial as "*a paper of considerable interest*"; and was featured in *The Biochemist* (December 2009, pp.23-38) and a Science & Society Feature in *EMBO Reports* (May 2010, Vol. 11, No. 5, pp 345-349).

1. **Pettifer, S.R., Sinnott, J.R., Attwood, T.K**. (2004) UTOPIA - User-friendly tools for operating informatics applications. *Comp. Funct. Genom.* 5 (1). 56-60. DOI: 10.1002/cfg.359
2. **Pettifer, S., Thorne, D., McDermott, P., Marsh, J.,** Villeger, A., Kell, D.B., **Attwood, T.K.** (2009) Visualising biological data: a semantic approach to tool and database integration. *BMC Bioinformatics.* 10(6). S19. DOI: 10.1186/1471-2105-10-S6-S19
3. **Attwood, T.K.,** Kell, D.B., **McDermott, P., Marsh, J., Pettifer, S.R., Thorne, D.** (2009) Calling International Rescue: knowledge lost in literature and data landslide! *Biochem. J.* 424 (3). 317-333. DOI: 10.1042/BJ20091474
4. **Attwood, T.K.,** Kell, D.B., **McDermott, P., Marsh, J., Pettifer, S.R., Thorne, D.** (2010) Utopia Documents: linking scholarly literature with research data. *Bioinformatics.* 26(18). i568-i574. DOI: 10.1093/bioinformatics/btq383
5. Vroling, B., **Thorne, D., McDermott, P., Attwood, T.K.,** Vriend, G., **Pettifer, S.R.** (2011) Integrating GPCR-specific information with full text articles. *BMC Bioinformatics.* 12 (1). 362. DOI: 10.1186/1471-2105-12-362
6. Vroling, B., **Thorne, D., McDermott, P.,** Joosten, H.J., **Attwood, T.K., Pettifer, S.,** Vriend, G. (2012) NucleaRDB: information system for nuclear receptors. *Nucleic Acids Res.* 40. D377-D380. DOI: 10.1093/nar/gkr960

4. Details of the impact

Context

The life sciences are fast becoming data-rich but knowledge-poor. High-throughput biology is generating data and articles at a rate that makes it virtually impossible for individuals to keep up and remain expert in their fields. This problem has become particularly acute in large organisations such as pharmaceutical companies. Top pharma companies have noted that significant numbers (possibly as many as ~25%) of late-stage failures could be eliminated years earlier by making all internal information in documents more widely available. The ability to recover in-house data efficiently therefore provides an opportunity to reduce the risk of late-stage failures and the costs of drug discovery and development.

Pathways to impact

The research was presented to major EU consortia, leading bioinformatics institutes (EBI, SIB), conferences, EMBO- and FEBS-funded training schools, and workshops organised by the publishing and drug-discovery industries (SGUK12, ALPSP, EuroQSAR, APE2012).

The Utopia software has also been featured and promoted in a variety of online formats: these have included an 'Elevator Pitch' in *The Guardian* (6/10/10), and blogs by David Worlock (12/10/11), Jodi Schneider (14/11/10), Duncan Hull (11/12/09), Richard Kid (21/06/11) and others; videos about Utopia have been viewed >5,000 times on SkipTV (uploaded 20/12/09), YouTube (16/12/09) and SciVee (27/04/10). The basic PDF reader is free to download for academic and commercial users under a software licence agreement.

Reach and significance of the impact

Utopia Documents makes a unique contribution to scholarly publishing on a global scale. Combining advantages of PDF files (*i.e.*, readability, portability, archivability) with web-style

interactivity, the software allows readers to interact with and annotate article PDFs, and to share their commentaries and annotations with others.

- [Text removed for publication] [A].

The software offers publishers a way of enriching their vast, inert PDF back-catalogues, and driving more traffic to their online content. Utopia thus provides new commercial perspectives to publishers who are revisiting business models in view of open access initiatives.

Utopia Documents enables pharmaceutical companies to recover and exploit in-house knowledge that is otherwise lost during the drug-discovery process.

- [Text removed for publication] [B].

Providing efficiency and cost saving in the global drug discovery industry:
The software is in the process of being rolled-out across AstraZeneca (AZ) and Roche's global infrastructures, and discussions are ongoing with British Telecom to provide Utopia Documents as a cloud-based service for the wider pharmaceutical industry.

- [Text removed for publication] [C].

The software is now also fully integrated with the commercial 3D-molecule databases marketed by Bio-Prodict [D].

Changing the publishing platforms of UK-based publishers:

- Portland Press Ltd (PPL): A collaborative project resulted in the launch of the Semantic *Biochemical Journal (BJ)* in 2009, a 'publishing first' [E]. This has transformed the publishing pipeline at PPL, and all *BJ* content since 2009 (>1,600 publications) has been integrated with Utopia Documents [F]. [Text removed for publication].
    - During contract negotiations with Elsevier, Portland Press's decision to fund the Semantic *BJ* project was described by Elsevier as "*heroic and visionary"* for a small publisher.
    - The Semantic *BJ* (powered by Utopia Documents) was a finalist for the ALPSP Charlesworth prize for publishing innovation in 2010 [G].
- Royal Society of Chemistry: the Utopia team provided PDF access to ChemSpider (the 2010 ALPSP Charlesworth award winner for Publishing Innovation) via Utopia Documents, rolling this new functionality out across the Royal Society of Chemistry's entire publishing platform.

Leading a change in business strategy of international publishing houses:
Version 2.2 of the Utopia Documents Web-enabled PDF-reader for scientific content is now integrated with a variety of publishing platforms including Springer, learned societies (the Royal Society of Chemistry, American Society of Plant Biologists), and open access formats (eLife, PLoS, PeerJ, BioMedCentral, and open access articles in PubMed Central).

Other on-going projects include:

- Elsevier/SciVerse-Science Direct, AQnowledge, SiBite Ltd.: Technology from Utopia (licenced via AQnowledge, and in collaboration with SciBite Ltd.) is being used to provide Elsevier with 'business intelligence', linking scientific articles to commercially available lab supplies. A successful 50-journal pilot on Science Direct averaged ~100 reader-supplier interactions per day – subsequent projections across the full coverage of life science journals in Science Direct suggest ~150 million displays per year of the Utopia-driven AQnowledge link 'box'. To date, the software has been used to analyse >300,000 full-text articles from Elsevier's catalogue. It is also being implemented at CiteULike (a scientific reference managing and discovery site) [B].
- CrossRef: In collaboration with CrossRef (a not-for-profit organisation representing over 4,500 scholarly publishers), Attwood's research group is providing software for converting legacy PDF articles into semantically rich, computer-readable documents and improving their citation analysis. [Text removed for publication] [H].

- Wiley-Blackwell: A project is ongoing (2010-2014) to create the first 'Utopian' text-book.

Formation of a spin-out company, Lost Island Labs:

In 2012, Utopia IP was transferred into a start-up company (Lost Island Labs.), to manage the interface between Utopia and its industrial partners, to create a more sustainable environment for its development, and to share revenues with UoM. Within its first year of business, Lost Island Labs expects a profit [text removed for publication].

A revenue-sharing agreement between Lost Island Labs, AQnowledge and various scientific, medicine and technology publishers (including Elsevier, BioMed Central, Cold Spring Harbor and ~25 lab-product suppliers) has been set up.

5. Sources to corroborate the impact

A. Letter from Chief Executive, Associated of Learned and Professional Society Publishers (ALPSP), *corroborating the value of Semantic BJ to ALPSP.*

B. Letter from CEO, AQnowledge, *corroborating value of Utopia to aqnowledge.com.*

C. Letter from Director of Informatics, AstraZeneca, *corroborating the value of Utopia to the pharmaceutical industry*.

D. BioProdict's technology and product range: www.bio-prodict.nl/#technolog

E. The Semantic Biochemical Journal: www.biochemj.org/bj/semantic_faq.htm

F. Letter from Head of Editorial Department, Portland Press Limited, *corroborating value of Utopia in article mark-up.*

G. *Semantic BJ* shortlisting for ALPSP Award for Publishing Innovation 2010: http://www.alpsp.org/Ebusiness/AboutALPSP/ALPSPStatements/Statementdetails.aspx?ID=77

H. Letter from Executive Director, CrossRef, *corroborating value of Utopia to CrossRef & small publishers.*