A day in the life of *a content miner and team* Peter Murray-Rust

Miners, pirates, chemists and the odd Cambridge pub mingle to make sure life is never dull for Peter and his inspired team.

Insights follows them on a day in their busy but fun and exciting lives in the world of text and data mining.

It's tough for Peter getting out of bed today – yesterday he travelled to Brussels and back, fighting for 'The Right to Read is the Right to Mine' (R2RR2M). Content mining – also known as text and data mining (TDM) – is a hot topic in Europe. It's got huge promise, with two million scholarly publications a year and so much data that we can't take it all in – 5,000 papers a day (and grey literature, and theses, and…). So we must have machines to help.

But there's a snag. Mining involves copying, and copying *could* violate copyright – the law is so complex and fuzzy that no one is absolutely sure. 'We need clarity', say European Commissioners, and the European Parliament (EP) agrees. The law needs changing. But the 'rights owners' – particularly the publishers – feel threatened and have fought against reform for several years, including massive lobbying in Brussels.

Meet Julia Reda, MEP, Pirate Party, one of our Euro-heroes, who produced a far-reaching document on Copyright Reform for the EP. Balanced (we think!) but hugely controversial, with 500+ amendments, many proposing even more reform and as many clawing back. Julia got 80 invites to dinner from lobbyists in the first week! But she wants to find out what it's all about, so Tom – a chemist from Imperial – prepares a distribution system. Peter arranges a hack day in Brussels and promises Julia she can learn how to use ContentMine software. Running the Ubuntu Linux system, she installed and ran a query in 15 minutes.

> 'Mining involves copying, and copying *could* violate copyright'

> 'Julia got 80 invites to dinner from lobbyists in the first week!'

Julia Reda MEP running ContentMine software.

Now for the details. What's ContentMine? Talk alone does not make change, so we have to build the tools. So, while we were fighting for the UK Right to Mine (The 'Hargreaves' reform of UK copyright, enacted in 2014) we were also building the software to do it. And two years ago the Shuttleworth Foundation – which funds people to 'change the world' – funded Peter to make ContentMine actually happen. We gathered an amazing group of people, mainly in Cambridge, who share this passion and together we have created everything we need: the design, the software, the practice, workshops. We're a non-profit: 'contentmine.org'. And it's for you – librarians, publishers, managers, scholars, students and citizens.

'Talk alone does not make change, so we have to build the tools.'

We're now set up to mine the complete scholarly literature every day!

Everything's free, open source, open data, open everything. We're working with several parts of Cambridge University and particularly members of the library community at Cambridge. Mining can be daunting, but Yvonne Nobis and Danny Kingsley have jumped right in. We're hoping to go fully live in May 2016 – certainly by the time you are reading this in July. And we're running the actual kit in chemistry, which has a wonderful group of computer officers who Peter has worked with for 15 years.

'Everything's free, open source, open data, open everything.'

Some of the Cambridge team. From left: Jenny Molloy, Tom Arrow, Yvonne Nobis and Danny Kingsley.

Jenny – an Oxbridge mosquito and plant scientist – has been with ContentMine since the start, managing our collaboration and funding. Today we're testing the system and planning the future. We can work anywhere there's a university WiFi so we're in the University GradPad, overlooking the river. We can also work in Makespace, cafés, colleges, departments and the famous Panton Arms pub. Classic Cambridge.

Here, Jenny's trying out a query for 'Zika' (the virus that's infecting Africa and South America, with worries about further spread). ContentMine is so simple to use. You just type the search term ('Zika') and it automatically gives you all the papers and instantly searches and indexes them and links the results to Wikipedia. It's almost as if the literature is alive – without any prompting it finds 150 papers, highlights the most important concepts (words, diseases, species). It tells Jenny that Zika is spread by Aedes mosquitoes, that there are related diseases such as West Nile and Yellow Fever, that deltamethrin is used as an insecticide. It even summarizes the institutions and funders who are most involved. All in two minutes.

'It's almost as if the literature is alive'

Drs Jenny Molloy and Peter Murray-Rust with a backdrop of the River Cam, colleges, pubs and the Mathematical Bridge.

We're looking for collaborators – scientists and librarians – and want to bring them together. We love hack days; ContentMine is a great tool for this – all you need is a warm room, WiFi, and pizza/buns/drinks. Anything from two hours upwards. We've run one for the Wellcome Trust, for European Bioinformatics Institute (EBI), for agricultural scientists on the Norwich campus (TGAC, John Innes, Sainsbury). We've developed lots of dictionaries to help the search – diseases, agricultural practice, drugs, plant chemistry. This is a huge opportunity for University libraries – it's fun and a great way of making contact.

'all you need is a warm room, WiFi, and pizza/buns/drinks'

And tomorrow our great collaborator Gita arrives from India for a two-year Cambridge lectureship on plant chemistry. Have to make sure we have a thick coat and umbrella for her as the weather's not wonderful. She's building a world resource – a database of essential oils (plant chemicals with aromatic, flavouring, medicinal and other properties). We've done preliminary searches with ContentMine and she went, 'Wow! What a treasure chest of information!' If we can do that in ten minutes, what can we do in two years ☺? Tom and Peter are off to EBI tomorrow to meet Magnus Manske (Sanger Centre), who's a wonderful Wikipedian, and then to meet Chris Steinbeck (ChEBI) to integrate chemical search.

Footnote

Text and data mining is sometimes mistakenly thought to be piracy, or illegal. Everything we do is ethical and legal within the UK, which specifically legitimized this two years ago. We take great care not to break copyright and we can use everything from open access papers, but from closed ones we publish only uncopyrightable facts and snippet context.

'Everything we do is ethical and legal within the UK'

Of course, Cambridge is a great place to find resources – it subscribes to 'most' academic journals and has a huge resource of other digitized material: theses, books, manuscripts. ContentMine technology is easily installed and run. It's an excellent way for libraries to offer new tools to scholars of all disciplines. For example we are working with scientists who do systematic literature reviews of reports on clinical trials and suchlike. They may have to 'read' 10,000 papers in a few days and our software can filter out most of these in seconds.

Oh – and we are organizing a public hack day in June for Open Technology Week in Cambridge, where anyone can try this. And another in the autumn on Systematic Reviews. Do come ...

Abbreviations and Acronyms

A list of the abbreviations and acronyms used in this feature and other *Insights* articles can be accessed here – click on the URL below and then select the 'Abbreviations and Acronyms' link at the top of the page it directs you to: **http://www.uksg.org/publications#aa**

Article copyright: © 2016 Tom Arrow, Jenny Molloy and Peter Murray-Rust. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use and distribution provided the original author and source are credited.

Corresponding author:
Tom Arrow
ContentMine, UK
E-mail: thomasarrow@gmail.com

Co-authors:
Peter Murray-Rust
Reader Emeritus in Molecular Informatics
Department of Chemistry, University of Cambridge, UK
E-mail: pm286@cam.ac.uk

ORCID iD: http://orcid.org/0000-0003-3386-3972

Jenny Molloy
Department of Plant Sciences, University of Cambridge, UK
E-mail: jenny@contentmine.org

To cite this article:

Arrow, T, Molloy, J C and Murray-Rust, P, A day in the life of *a content miner and team*, *Insights*, 2016, 29(2), 208–211; DOI: http://dx.doi.org/10.1629/uksg.310

Published by UKSG in association with Ubiquity Press on 05 July 2016