Mining chemistry from journals and theses

Richard Kidd
Publisher
kiddr@rsc.org
@rkiddr

ROYAL SOCIETY OF CHEMISTRY

Mining is a good description

Just like mining

Why

What you get out

Effort and quality

Automation

Why

You have a mountain of stuff which contains valuable nuggets

You (more or less) know what you're looking for

You know what you're going to do with it once you have it

What you get out

You get lots of stuff out

It requires sifting and grading

It’s a triumph if you manage to extract 80-90% of what is there

You will go back to the heap and redo it

Effort and quality

That which is easy to get out - is well known and unlikely to be novel

The novel and interesting is likely to be rare and not easily defined

Automation

Do the initial investigations by hand

Send in the machines later

Still needs some humans tweaking the valves

So…

TDM can help define the composition of the mountain (or bulk data)

Error likely to be large on the bucket (or article) level

How important is TDM in terms of publishing and especially at RSC?

Occasional requests

TDM added to licences on request

Part of the conversation

But it is occasional

pharma | text experts | researchers

Why do we take our approach?

Predates the UK copyright exception

Allow corporates and academia

Exception could have been clearer

Our community TDM projects

SciBorg | OSCAR | ChETA

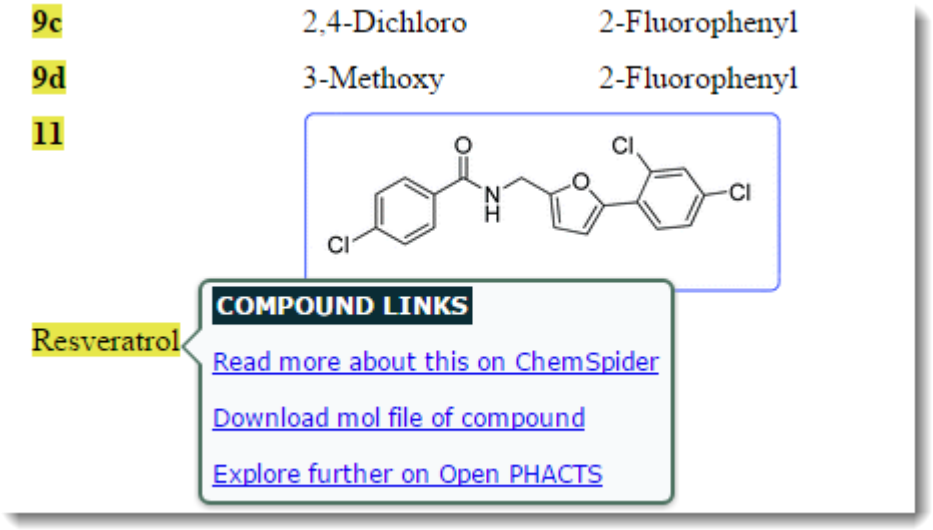
Pistoia SESL | TREC Chemistry

How do we benefit?

Enhanced articles

Start of community data standards

View of the future



Controlling TDM?

Issues around load and approach

Possible issues about quality

4. Possibility of obtaining appreciable yields in methane homologation through a two-step reaction at 250 °C on a platinum catalyst

Annie Amariglio, Pierre Par°ja, Mohammed Belgued and Henri Amariglio

J. CHEM. SOC., CHEM. COMMUN., 1994 561 Possibility of obtaining Appreciable Yields in **Methane**... 239, 54506 Vandaeuvre les Nancy, France **Methane** is converted (>40% yield) to higher alkanes at 250?C... **methane**. 1-7 In the first step, **methane** is chemisorbed133-4 or decomposed2.5-7 on the metal

J. Chem. Soc., Chem. Commun., 994, °561 - 562

Citation and credit

Perceived issue about derivative works

P-D-R and TDM Discussions

Clear that standard access is important, and people write their own queries for their own expertise and use cases

See Martin Romacker (Roche) for a great presentation indicating variability in content and structure across publishers

The future?

Standard portals
(e.g. CrossRef Text & Data Mining, CCC)
aggregated solutions seem sensible

Still experimental outside pharma

Expect interesting commercial derivatives
to appear

A different (non-article) view

National Compound Collection pilot

15 UK institutions | 9 pharma and academic groups | British Library

Explore availability of the dissertation resource end-to-end

http://rsc.li/1E7ct56

National Compound Collection

Manual extraction, but report generated by BL covered IP issues around copyright and ownership – both actual and perceived

Copyright of theses?

Availability?

When does the activity become commercial?

Copyright issues

Dissertation copyright varies

Institutional agreements

Author copyright

Published or not?

Pilot objectives

700 theses – 45,000 compounds

Screening for interesting drug candidates
Mapping the chain
Reward at each stage

Funders encouraging submission ?
Mining of old collections ?
Prove and extend ?

Relevance to TDM?

Still uncertainty over UK copyright exception

Researcher and institution differences in practices and assumptions

Summary

Our approach is to be permissive

Personal view: real functional uses now, especially in biomed, rare and experimental elsewhere

Just a better search engine? Or a stepping stone to community data standards and structured data


DANKE!