Open Science

Peter Murray-Rust,

***ContentMine.org**, and University of Cambridge*

Opencon2015, Bologna, IT 2015-11-18

What is “Open”?

Why is it essential?

Open Data

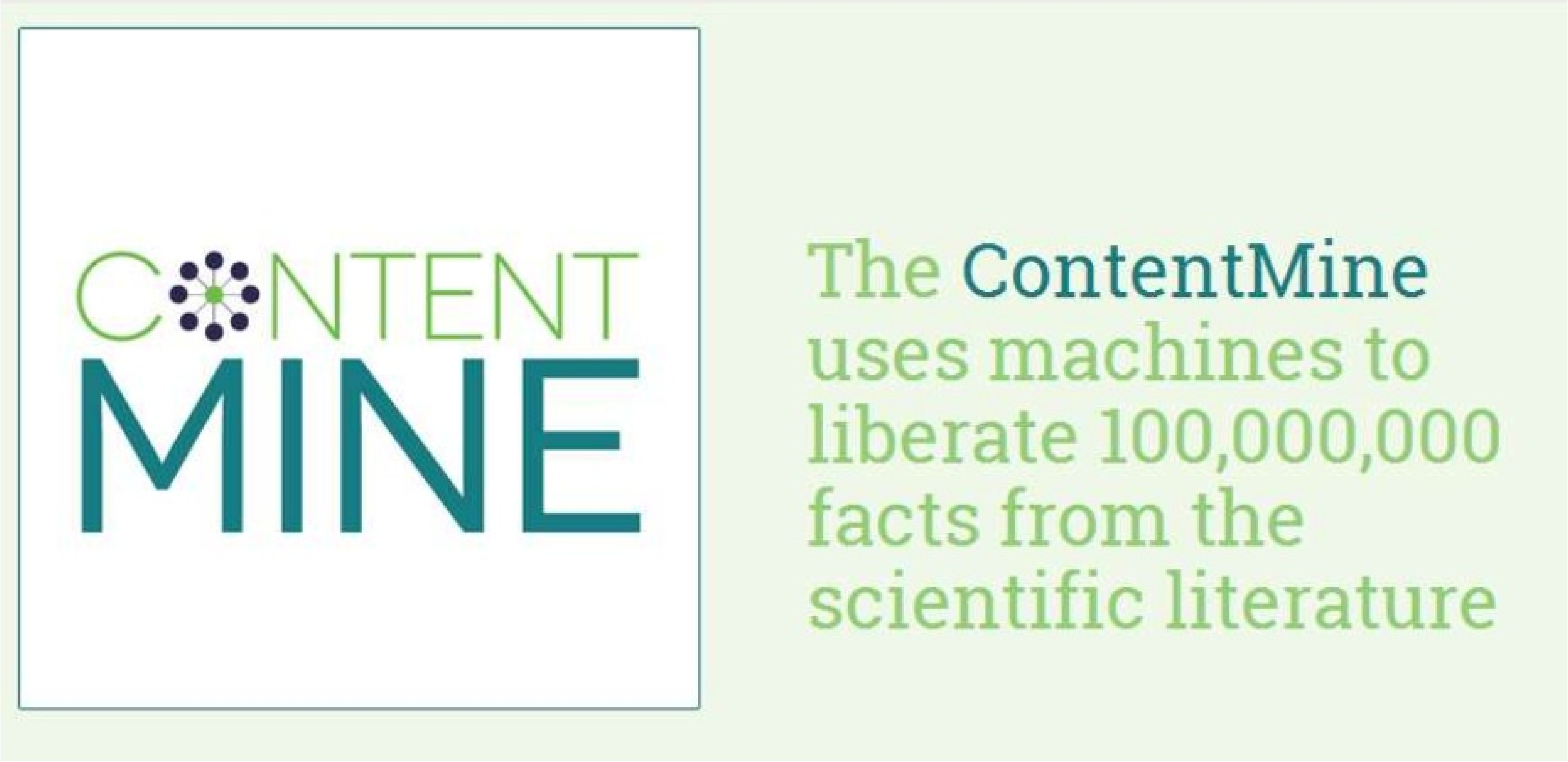
Content Mining – a battle we must win

Young researchers are the present (*Mike Eisen*)

UNIVERSITY OF CAMBRIDGE

Open Knowledge Foundation

SHUTTLEWORTH FUNDED



The ContentMine uses machines to liberate 100,000,000 facts from the scientific literature

The Right to Read is the Right to Mine*

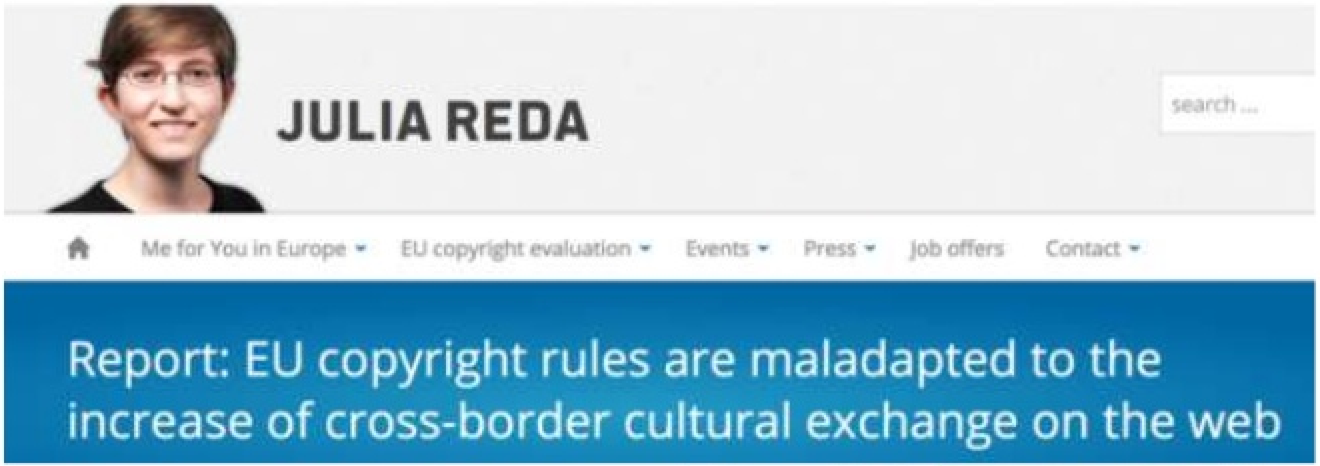
*PeterMurray-Rust, 2011

http://contentmine.org

SHUTTLEWORTH
FELLOW

My European Heroes

NEELIE KROES

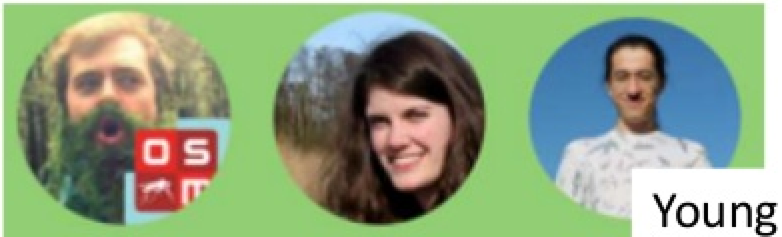


JULIA REDA

search ...

Me for You in Europe ▾ | EU copyright evaluation ▾ | Events ▾ | Press ▾ | Job offers | Contact ▾

Report: EU copyright rules are maladapted to the increase of cross-border cultural exchange on the web



Young People(*ContentMine*)

Messages

- The system is completely broken
- We are at war with major publishers
- Students have the power to change the world
- Universities need help from students
- Open is a state of mind
- The opposite of Open is broken [1]
- Friction destroys Open
- Don’t buy it, build it …
- … TOGETHER

[1] (John Wilbanks)

Breaking news: Elsevier stopped me doing my research *Chris Hartgerink*

@Senficon (Julia Reda) :Text & Data mining in times of **#copyright** maximalism:

"Elsevier stopped me doing my research" http://onsnetwork.org/chartgerink/2015/11/16/elsevier-stopped-me-doing-my-research/ … **#opencon** **#TDM**

Chris Hartgerink's blog post

I am a statistician interested in detecting potentially problematic research such as data fabrication, which results in unreliable findings and can harm policy-making, confound funding decisions, and hampers research progress.

To this end, I am content mining results reported in the psychology literature. Content mining the literature is a valuable avenue of investigating research questions with innovative methods. For example, our research group has written an automated program to mine research papers for errors in the reported results and found that 1/8 papers (of 30,000) contains at least one result that could directly influence the substantive conclusion [1].

In new research, I am trying to extract test results, figures, tables, and other information reported in papers throughout the majority of the psychology literature. As such, I need the research papers published in psychology that I can mine for these data. To this end, I started 'bulk' downloading research papers from, for instance, Sciencedirect. I was doing this for scholarly purposes and took into account potential server load by limiting the amount of papers I downloaded per minute to 9. I had no intention to redistribute the downloaded materials, had legal access to them because my university pays a subscription, and I only wanted to extract facts from these papers.

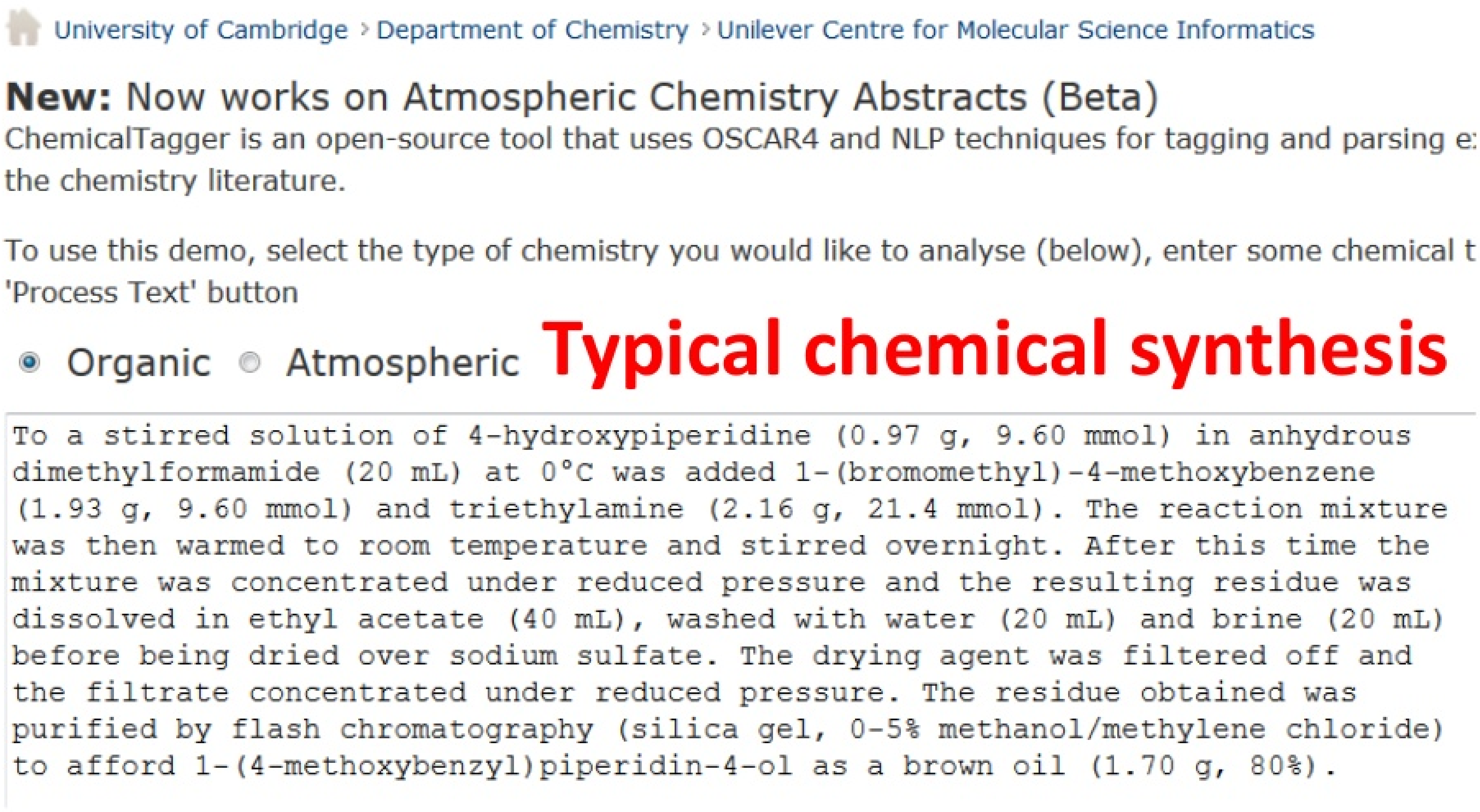
Full disclosure, I downloaded approximately 30GB of data from Sciencedirect in approximately 10 days. This boils down to a server load of 0.0021GB/[min], 0.125GB/h, 3GB/day.

Approximately two weeks after I started downloading psychology research papers, Elsevier notified my university that this was a violation of the access contract, that this could be considered stealing of content, and that they wanted it to stop. My librarian explicitly instructed me to stop downloading (which I did immediately), otherwise Elsevier would cut all access to Sciencedirect for my university.

I am now not able to mine a substantial part of the literature, and because of this Elsevier is directly hampering me in my research.

[1] Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). Behavior Research Methods, 1–22. doi: 10.3758/s13428-015-0664-2

http://chemicaltagger.ch.cam.ac.uk/



University of Cambridge › Department of Chemistry › Unilever Centre for Molecular Science Informatics

New: Now works on Atmospheric Chemistry Abstracts (Beta)

ChemicalTagger is an open-source tool that uses OSCAR4 and NLP techniques for tagging and parsing e: the chemistry literature.

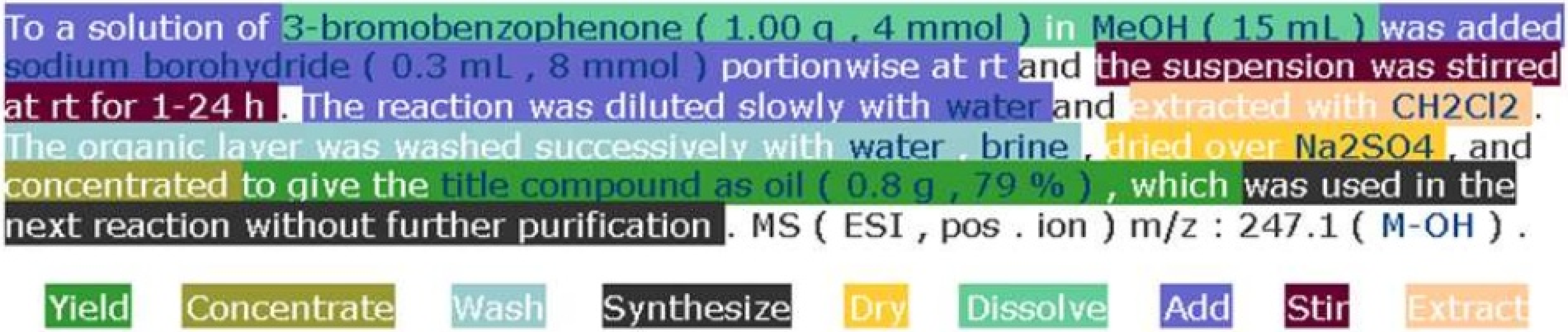
To use this demo, select the type of chemistry you would like to analyse (below), enter some chemical t 'Process Text' button

● Organic ○ Atmospheric

Typical chemical synthesis

To a stirred solution of 4-hydroxypiperidine (0.97 g, 9.60 mmol) in anhydrous dimethylformamide (20 mL) at 0°C was added 1-(bromomethyl)-4-methoxybenzene (1.93 g, 9.60 mmol) and triethylamine (2.16 g, 21.4 mmol). The reaction mixture was then warmed to room temperature and stirred overnight. After this time the mixture was concentrated under reduced pressure and the resulting residue was dissolved in ethyl acetate (40 mL), washed with water (20 mL) and brine (20 mL) before being dried over sodium sulfate. The drying agent was filtered off and the filtrate concentrated under reduced pressure. The residue obtained was purified by flash chromatography (silica gel, 0-5% methanol/methylene chloride) to afford 1-(4-methoxybenzyl)piperidin-4-ol as a brown oil (1.70 g, 80%).

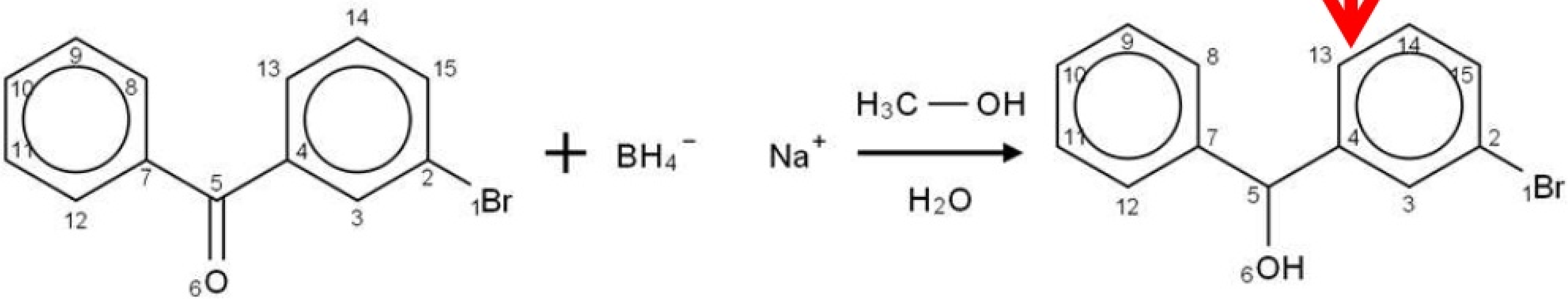
Open Content Mining of FACTs



To a solution of 3-bromobenzophenone (1.00 g , 4 mmol) in MeOH (15 mL) was added sodium borohydride (0.3 mL , 8 mmol) portionwise at rt and the suspension was stirred at rt for 1-24 h . The reaction was diluted slowly with water and extracted with CH2Cl2 . The organic layer was washed successively with water , brine , dried over Na2SO4 , and concentrated to give the title compound as oil (0.8 g , 79 %) , which was used in the next reaction without further purification . MS (ESI , pos . ion) m/z : 247.1 (M-OH) .

Yield Concentrate Wash Synthesize Dry Dissolve Add Stir Extract

Machines can interpret chemical reactions

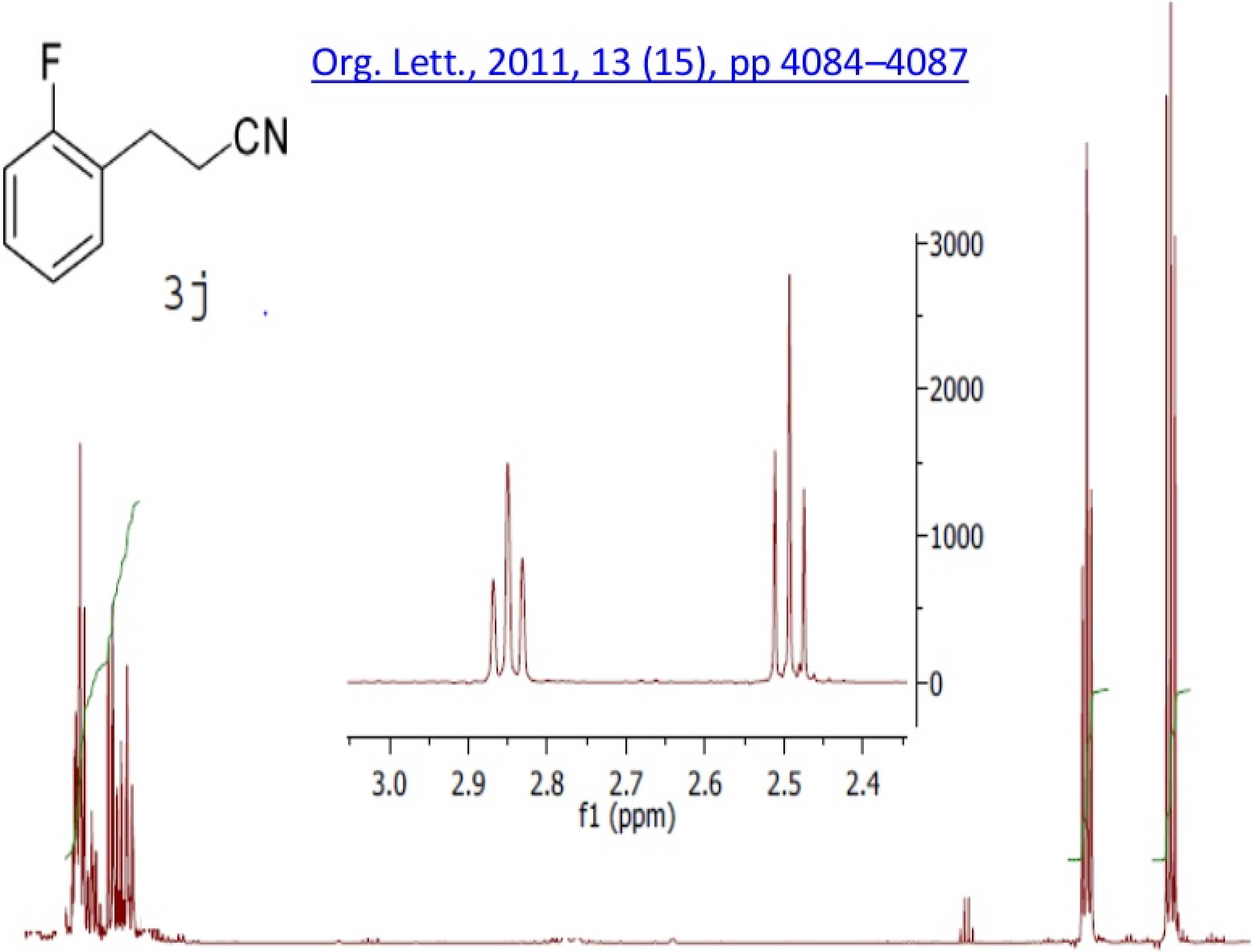


We have done 500,000 patents. There are > 3,000,000 reactions/year. Added value > 1B Eur.

C) What's the problem with this spectrum?

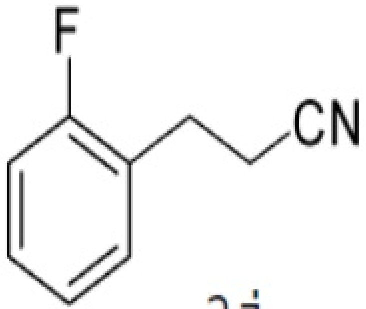
Original thanks to ChemBark

Org. Lett., 2011, 13 (15), pp 4084–4087

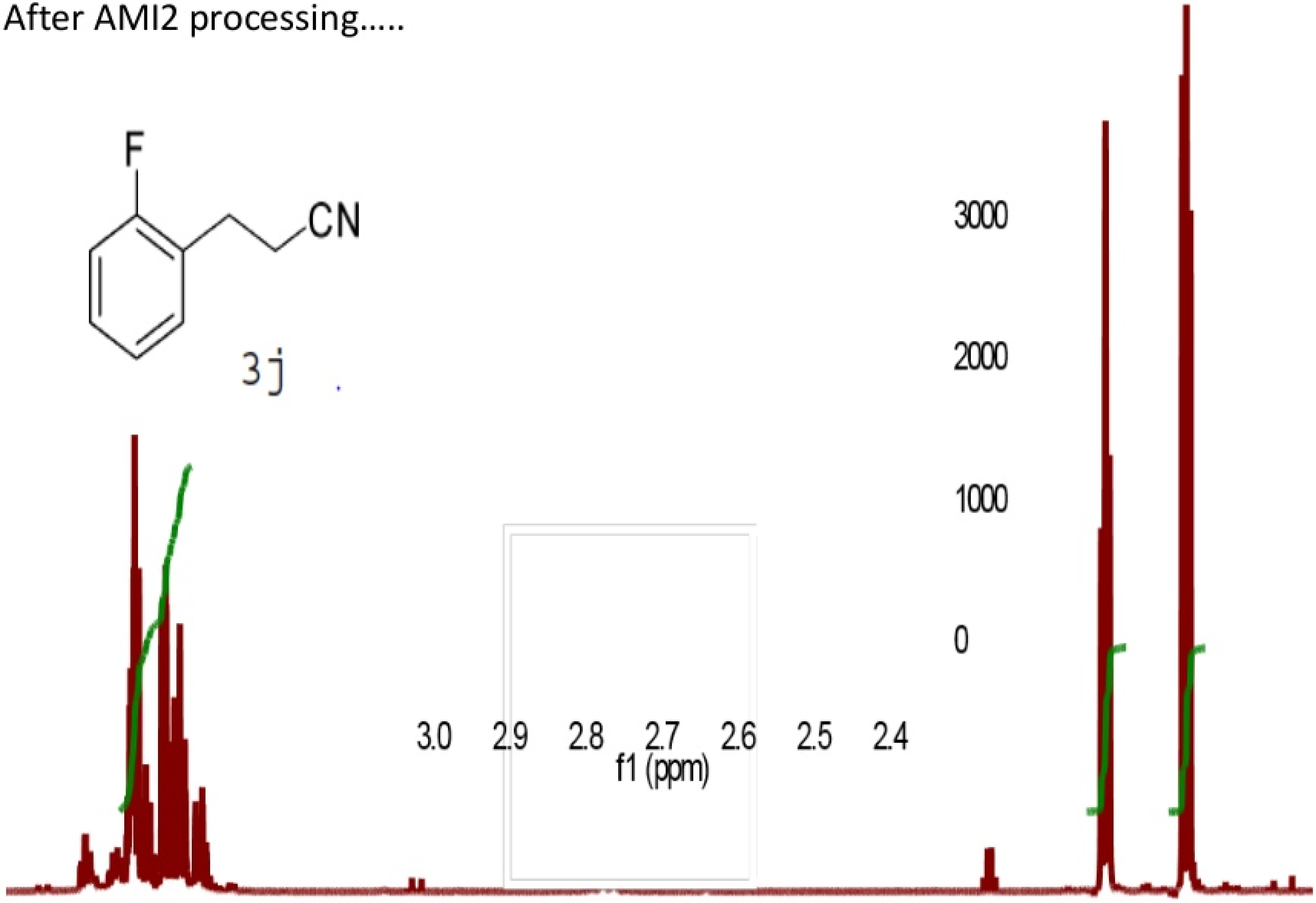


3j

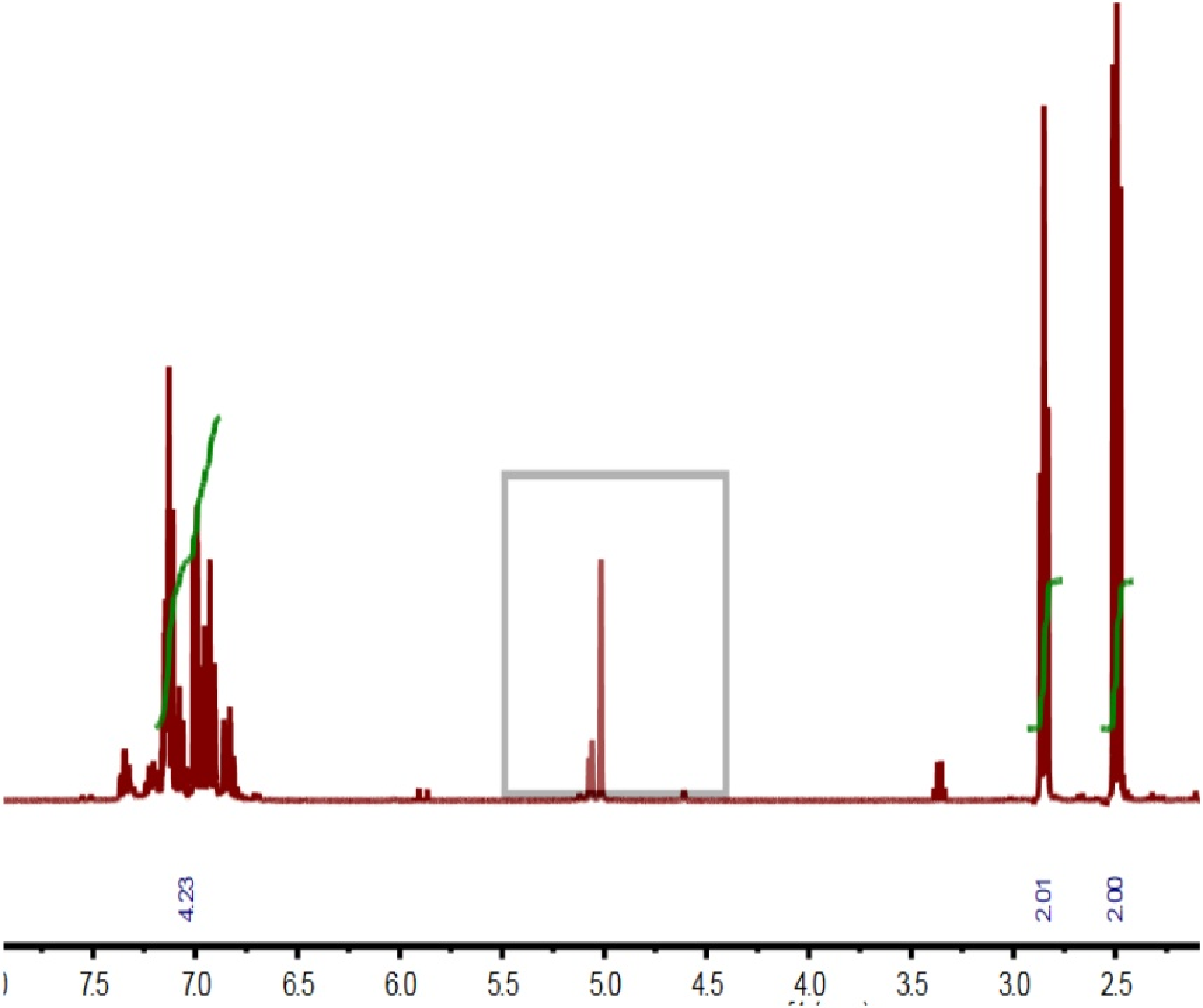
After AMI2 processing.....



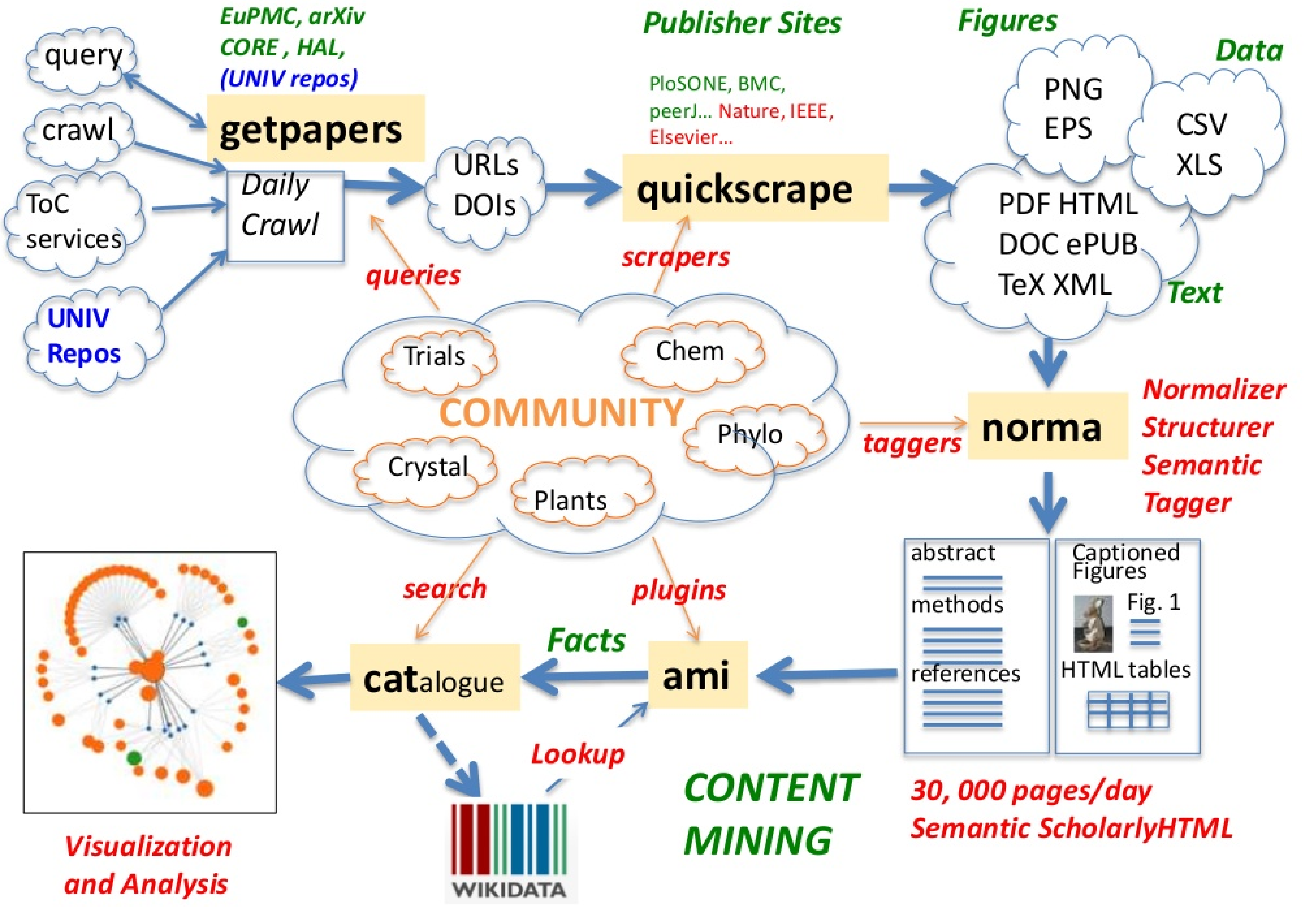
3j



... AMI2 has detected a square


4.23
2.01
2.00
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5

CONTENTMINE Complete OPEN Platform for Mining Scientific Literature



query

crawl

ToC services

UNIV Repos

EuPMC, arXiv
CORE , HAL,
(UNIV repos)

getpapers

Daily Crawl

URLs DOIs

Publisher Sites

PloSONE, BMC,
peerJ... Nature, IEEE,
Elsevier...

quickscrape

Figures

PNG EPS

Data

CSV XLS

PDF HTML
DOC ePUB
TeX XML

Text

queries

scrapers

COMMUNITY

Trials

Chem

Phylo

Crystal

Plants

taggers

norma

Normalizer
Structurer
Semantic
Tagger

abstract

methods

references

Captioned Figures

Fig. 1

HTML tables

30, 000 pages/day
Semantic ScholarlyHTML

search

plugins

Facts

ami

catalogue

Lookup

WIKIDATA

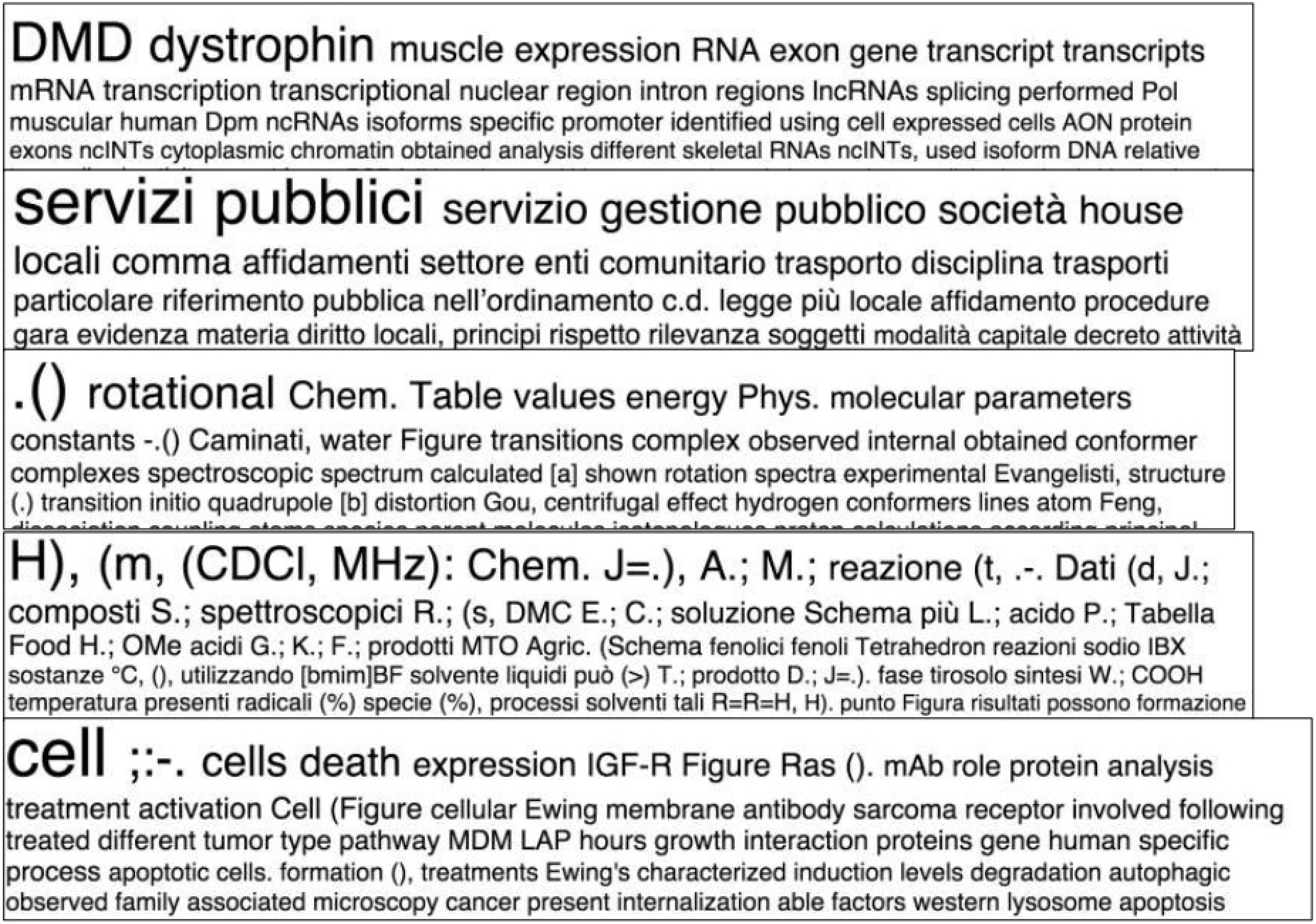
CONTENT MINING

Visualization and Analysis

Stand back! I am about to do ContentMining!

- Erriquez Daniela, Esame finale: Bologna, Aprile 2014
- Dott.ssa Elena Fiorentini, n. 0000274966, TESI DI DOTTORATO, Bologna
- Qian Gou, Esame finale: Bologna, finale 2014
- Maurizio BARONTINI, UNIVERSITÀ DEGLI STUDI DELLA TUSCIA DI VITERBO
- Terracciano Mario, Esame finale anno 2014

BagOfWords for Italian Theses



DMD dystrophin muscle expression RNA exon gene transcript transcripts mRNA transcription transcriptional nuclear region intron regions lncRNAs splicing performed Pol muscular human Dpm ncRNAs isoforms specific promoter identified using cell expressed cells AON protein exons ncINTs cytoplasmic chromatin obtained analysis different skeletal RNAs ncINTs, used isoform DNA relative

servizi pubblici servizio gestione pubblico società house locali comma affidamenti settore enti comunitario trasporto disciplina trasporti particolare riferimento pubblica nell’ordinamento c.d. legge più locale affidamento procedure gara evidenza materia diritto locali, principi rispetto rilevanza soggetti modalità capitale decreto attività

.() rotational Chem. Table values energy Phys. molecular parameters constants -.() Caminati, water Figure transitions complex observed internal obtained conformer complexes spectroscopic spectrum calculated [a] shown rotation spectra experimental Evangelisti, structure (.) transition initio quadrupole [b] distortion Gou, centrifugal effect hydrogen conformers lines atom Feng, dissociation coupling atoms species parent molecules isotopologues proton calculations according principal

H), (m, (CDCl, MHz): Chem. J=.), A.; M.; reazione (t, .-. Dati (d, J.; composti S.; spettroscopici R.; (s, DMC E.; C.; soluzione Schema più L.; acido P.; Tabella Food H.; OMe acidi G.; K.; F.; prodotti MTO Agric. (Schema fenolici fenoli Tetrahedron reazioni sodio IBX sostanze °C, (), utilizzando [bmim]BF solvente liquidi può (>) T.; prodotto D.; J=.). fase tirosolo sintesi W.; COOH temperatura presenti radicali (%) specie (%), processi solventi tali R=R=H, H). punto Figura risultati possono formazione

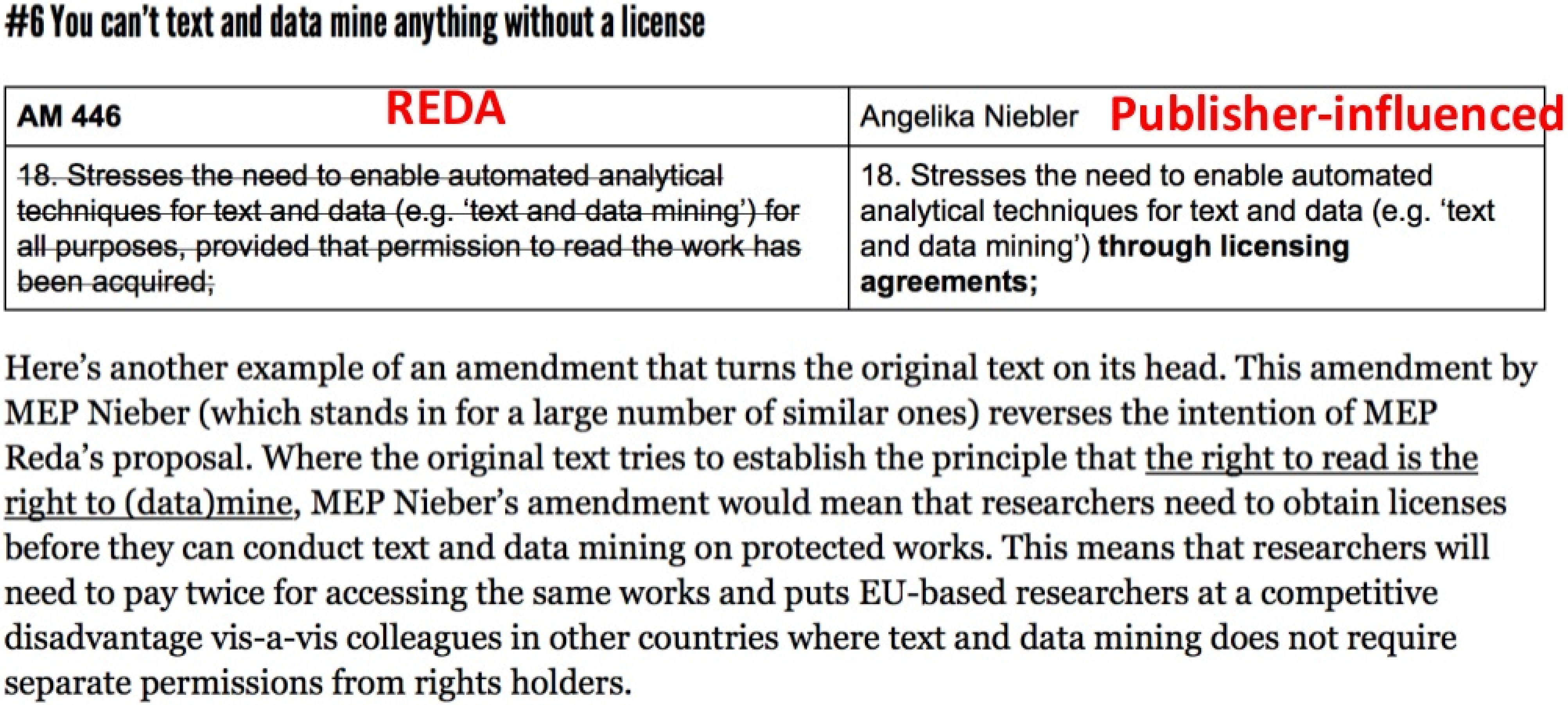
cell ;:-. cells death expression IGF-R Figure Ras (). mAb role protein analysis treatment activation Cell (Figure cellular Ewing membrane antibody sarcoma receptor involved following treated different tumor type pathway MDM LAP hours growth interaction proteins gene human specific process apoptotic cells. formation (), treatments Ewing's characterized induction levels degradation autophagic observed family associated microscopy cancer present internalization able factors western lysosome apoptosis

Refs: Erriquez_Daniela_tesi, Fiorentina_Elena_tesi, Gou_Qian_Tesi, mbarontini_tesid, terracciano_maria_tesi

Copyright and Mining

- UK (“Hargreaves”) 2014 legislation:
  - “personal” **“non-commercial*”** “research” “data analytics”
  - legitimizes copying (?to disk), but not publishing
- PMR-premise: You cannot do reproducible scientific mining and avoid violating copyright.

Massive political activity in Europe

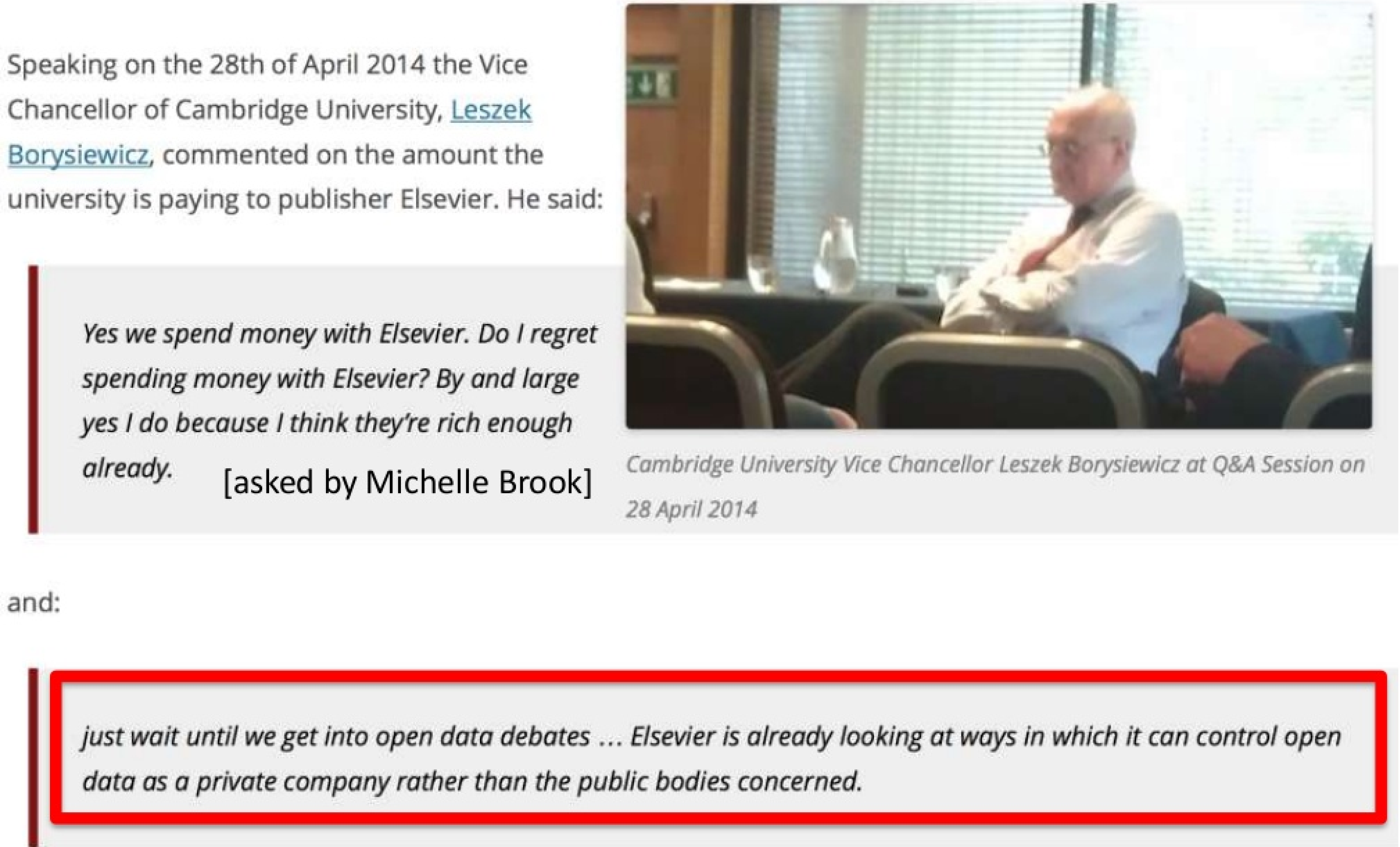


#6 You can't text and data mine anything without a license

AM 446 REDA	Angelika Niebler Publisher-influenced
~~18. Stresses the need to enable automated analytical techniques for text and data (e.g. 'text and data mining') for all purposes, provided that permission to read the work has been acquired;~~	18. Stresses the need to enable automated analytical techniques for text and data (e.g. 'text and data mining') **through licensing agreements;**

Here's another example of an amendment that turns the original text on its head. This amendment by MEP Nieber (which stands in for a large number of similar ones) reverses the intention of MEP Reda's proposal. Where the original text tries to establish the principle that the right to read is the right to (data)mine, MEP Nieber's amendment would mean that researchers need to obtain licenses before they can conduct text and data mining on protected works. This means that researchers will need to pay twice for accessing the same works and puts EU-based researchers at a competitive disadvantage vis-a-vis colleagues in other countries where text and data mining does not require separate permissions from rights holders.

Elsevier wants to control Open Data



Speaking on the 28th of April 2014 the Vice Chancellor of Cambridge University, Leszek Borysiewicz, commented on the amount the university is paying to publisher Elsevier. He said:

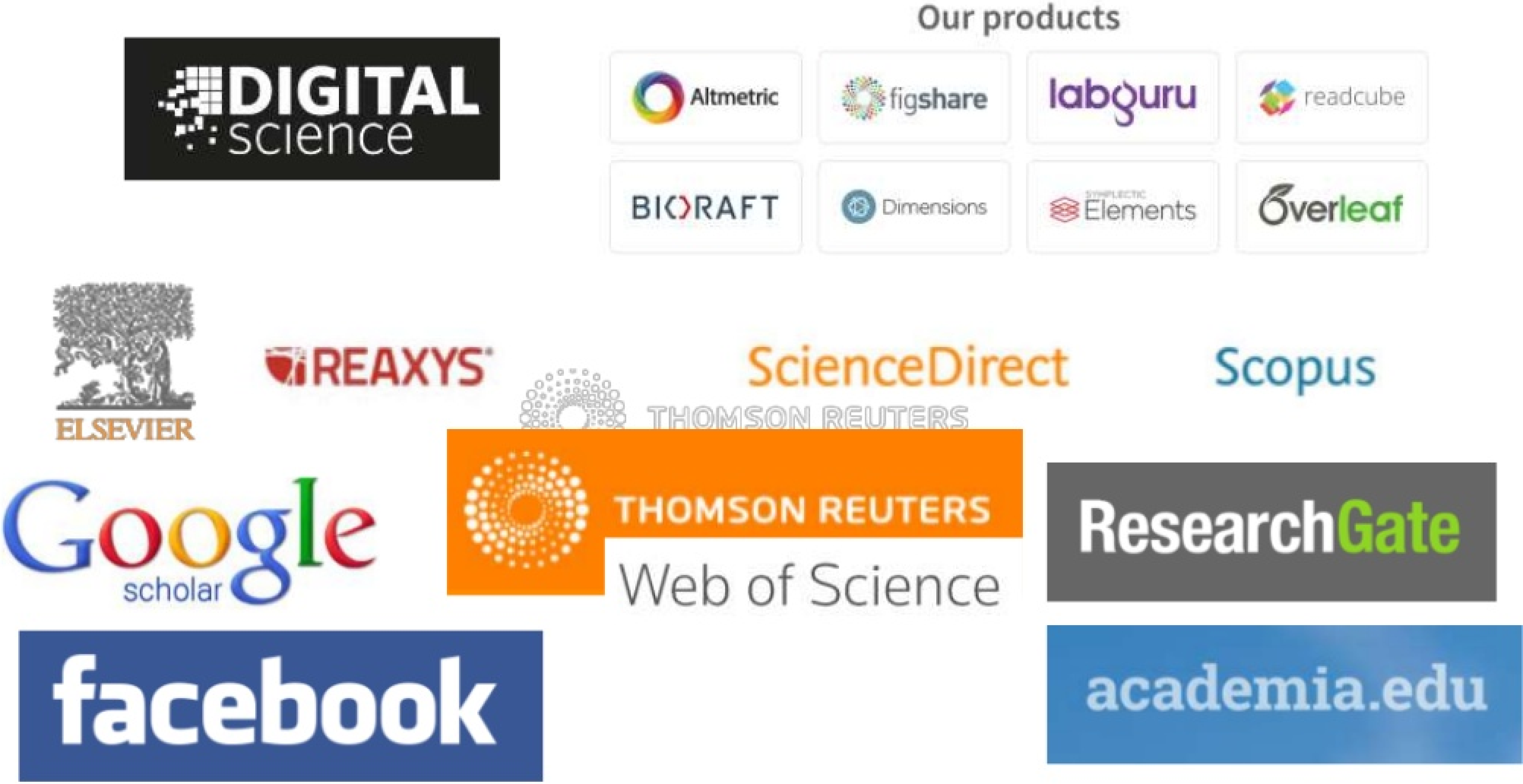
> *Yes we spend money with Elsevier. Do I regret spending money with Elsevier? By and large yes I do because I think they're rich enough already.* [asked by Michelle Brook]

Cambridge University Vice Chancellor Leszek Borysiewicz at Q&A Session on 28 April 2014

and:

> *just wait until we get into open data debates … Elsevier is already looking at ways in which it can control open data as a private company rather than the public bodies concerned.*

Scholarly infrastructure becomes closed



No accountability for monitoring and control

A System Failure of Scholarly Publishing

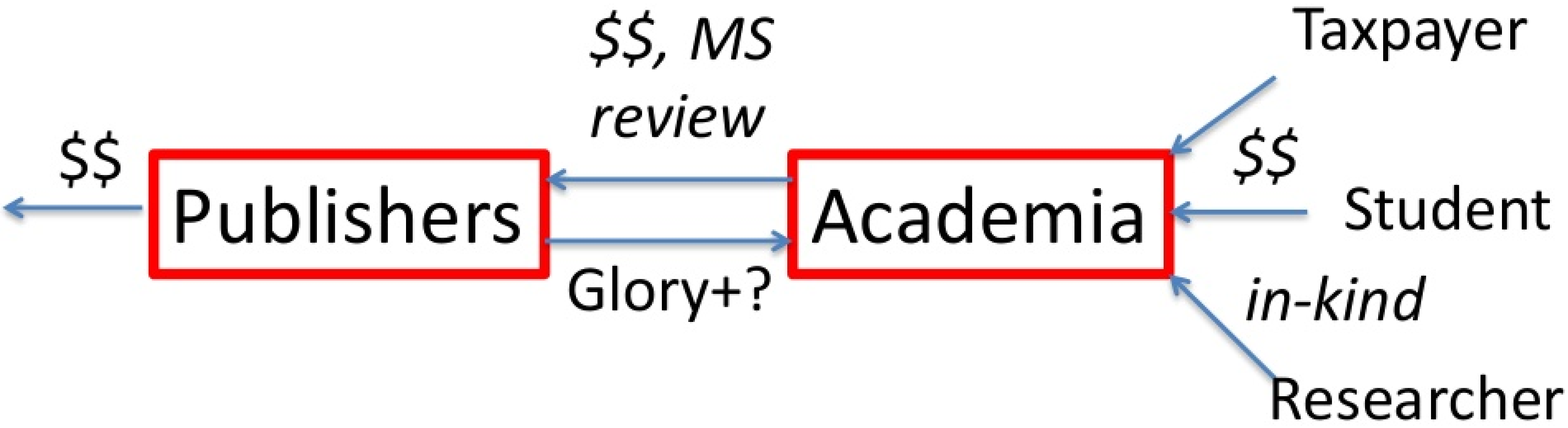
http://www.nytimes.com/2015/04/08/opinion/yes-we-were-warned-about-ebola.html

We were stunned recently when we stumbled across an article by European researchers in **Annals of Virology [1982**]: “The results seem to indicate that **Liberia has to be included in the Ebola virus endemic zone.**” In the future, the authors asserted, “medical personnel in Liberian health centers should be aware of the possibility that they may come across active cases and thus be prepared to avoid nosocomial epidemics,” referring to hospital-acquired infection.

Adage in public health: “**The road to inaction is paved with research papers**.”

Bernice Dahn (chief medical officer of Liberia’s Ministry of Health)
Vera Mussah (director of county health services)
Cameron Nutt (Ebola response adviser to Partners in Health)

The Publisher-Academic complex[1]



[1] The Military-Industrial-Academic complex (1961)

(Dwight D Eisenhower, US President)