The (young) Researcher POV

~~Text & Data Mining~~
Content Mining

Ross Mounce
University of Bath, PhD Candidate
Open Knowledge Foundation Panton Fellow

Open Knowledge Foundation

UNIVERSITY OF BATH

#Licences4Europe @RMounce

What is content mining?

- Text
- Pictures
- Videos
- Audio
- Metadata
- ...not just text or data!

The importance of scale

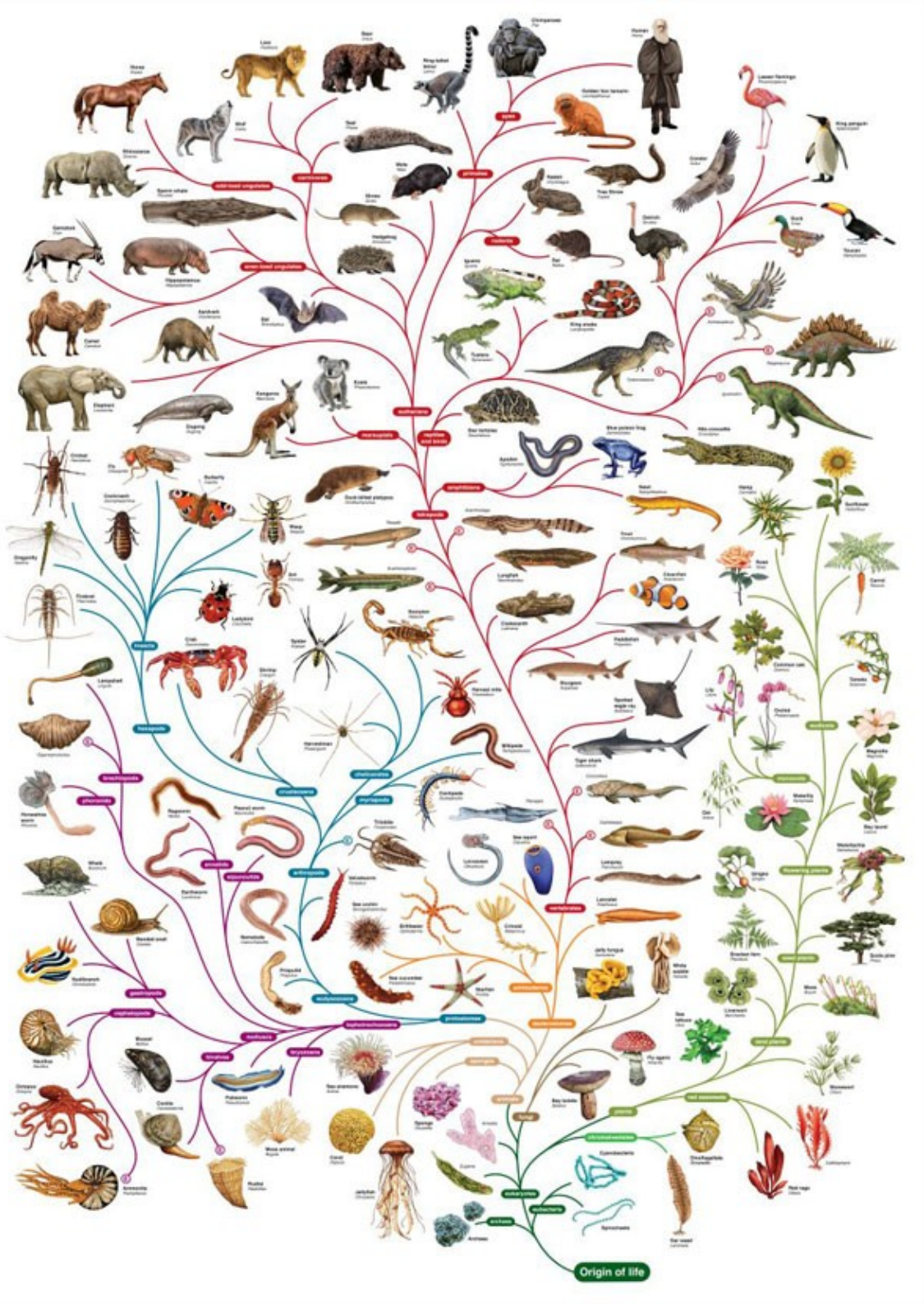
- The indexed WWW has >13 billion pages
- >630,000,000 sites (Netcraft, Feb 2013)
- 2 million scholarly articles published per year c. 4% growth rate each year
- >50,000,000 scholarly articles so far (Jinha, 2010)
- ~72,000 PDFs from PloS are just ~15GB
- >4,000,000 English-language Wikipedia pages downloaded as text are just ~9GB

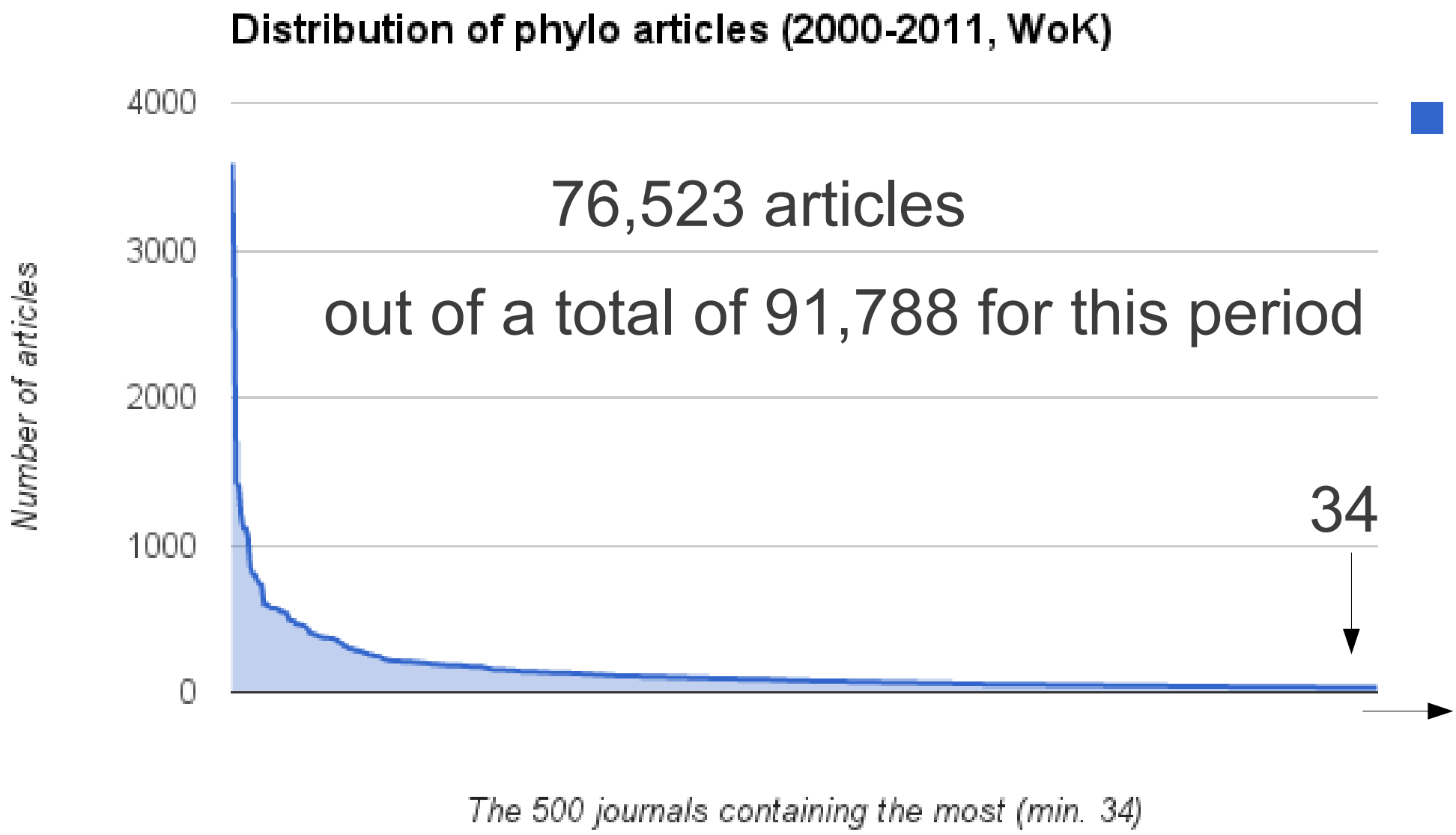
My interest in Content Mining

- Facts cannot be copyrighted
- Billions of facts are unfortunately trapped only in copyright-protected scholarly research articles
- My collaborators & I simply want to liberate these facts to make them openly available for everyone with no restrictions on access or use
- Within this, my particular interest in evolutionary history; e.g. phylogenetic tree data

I want to help reconstruct the 'Tree of Life'. Thousands of other scientists also want to do this.

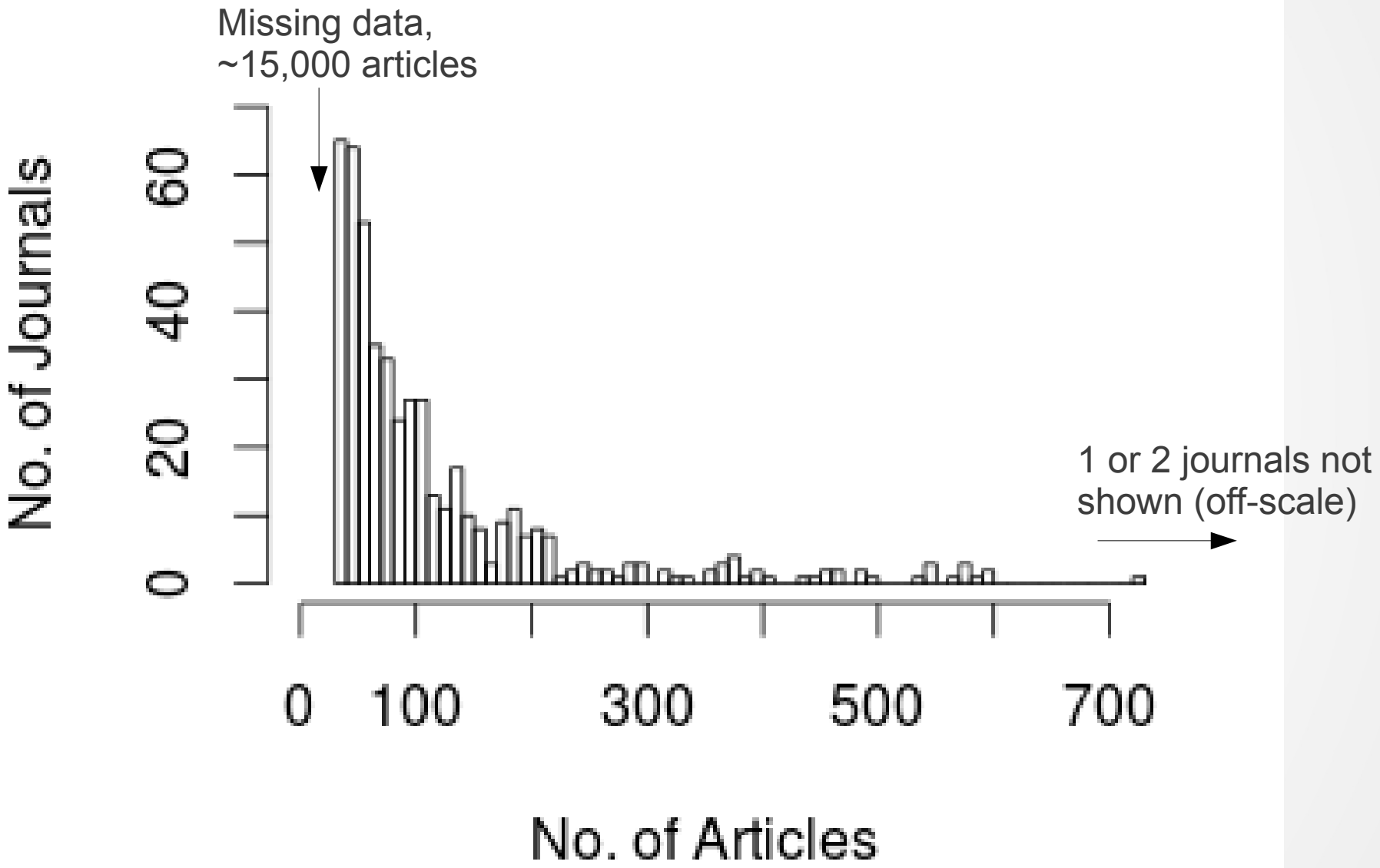
Phylogenetic Research gets published piecemeal across >100,000 papers. Hard to synthesise, much of it not in PMC





My content of interest is scattered across 1000+ journals

Histogram


Missing data,
~15,000 articles
No. of Journals
0
20
40
60
1 or 2 journals not
shown (off-scale)
0
100
300
500
700
No. of Articles

Lots of journals, lots of publishers

Just those top 500 most phylogeny 'rich' journals are published by ~120 different publishers

~ 17%	Elsevier (subscription journals)
~ 13%	Springer (subscription journals)
~ 15%	Wiley (subscription journals)

The biggest 3 publishers *combined* still only have less than 50% of what I want to mine

The long tail of content (15,000 articles) in journals outside the 'top 500' is likely to come from many more different publishers.

Standard license agreements

- Many explicitly do not allow mining

e.g. Nature, JSTOR, AIP, ACS, Elsevier, Taylor & Francis...

“This licence does not include any derivative use of the Site or the Materials, any collection and use of any product listings, descriptions, or prices; any downloading or copying of account information for the benefit of another merchant; **or any use of data mining, robots or similar data gathering** and extraction tools...”

InformaWorld

Source

See also: http://www.cdlib.org/services/collections/redactions/
http://www.mpdl.mpg.de/services/ezb-readme_en.htm

Asking permission doesn't scale

- There are 90,000 different publishers in 215 different countries listed in Ulrich's Periodicals Directory & >336,000 periodicals.

“I had a phone call on Friday with my university librarian and six (!) Elsevier employees.”

Heather Piwowar

5 March → 16 April *just* to get permission/access to start work on just one publisher's content

Could have done all of the analysis in time period.
Hugely intimidating & patronizing process, an utter waste of time

Why aren't more people mining?

A) Many disciplines not computationally-savvy frankly

B) Lack of awareness that this technology & capability exist & what it can do

C) Lack of teaching/training about content mining

D) Futility of even *trying* – how can one get permission from 1000's of publishers in the context of short research funding time-scales. Choose another research project – it's less risky.

E) Live in the wrong jurisdiction (e.g. Europe) and you're at a massive disadvantage relative to S. Korea, Japan, Israel, US... due to local laws.
(Probably more reasons too, this list is not exhaustive)

The Right to Read is The Right to Mine

- Through my institutional affiliation (U. of Bath) I have legitimate paid access to many thousands of paywalled journals
- If I visit the BL or NHM library I have legitimate access to even more
- Our POV: “The Right to Read is The Right to Mine”
- We do not see any distinction between computer-assisted 'human' reading aka 'ocular access' (e.g. viewing a PDF), and computer-assisted computer reading – aka content mining.

  The only difference is the latter is clearly more time & resource efficient – it's much quicker and it scales well.

Blocking & Criminalizing Research

- I have had my access to at least one publisher (BioOne) cut-off before. My 'crime' – downloading more than 25 PDFs in 5mins.
- Elsevier once blocked access to the entire U. of Cambridge campus for a week(!) because Peter Murray-Rust through legitimate access downloaded 'too many' PDFs in the course of his research
- Countless other cases, large majority not widely reported
- ...Aaron Swartz – need I say more?

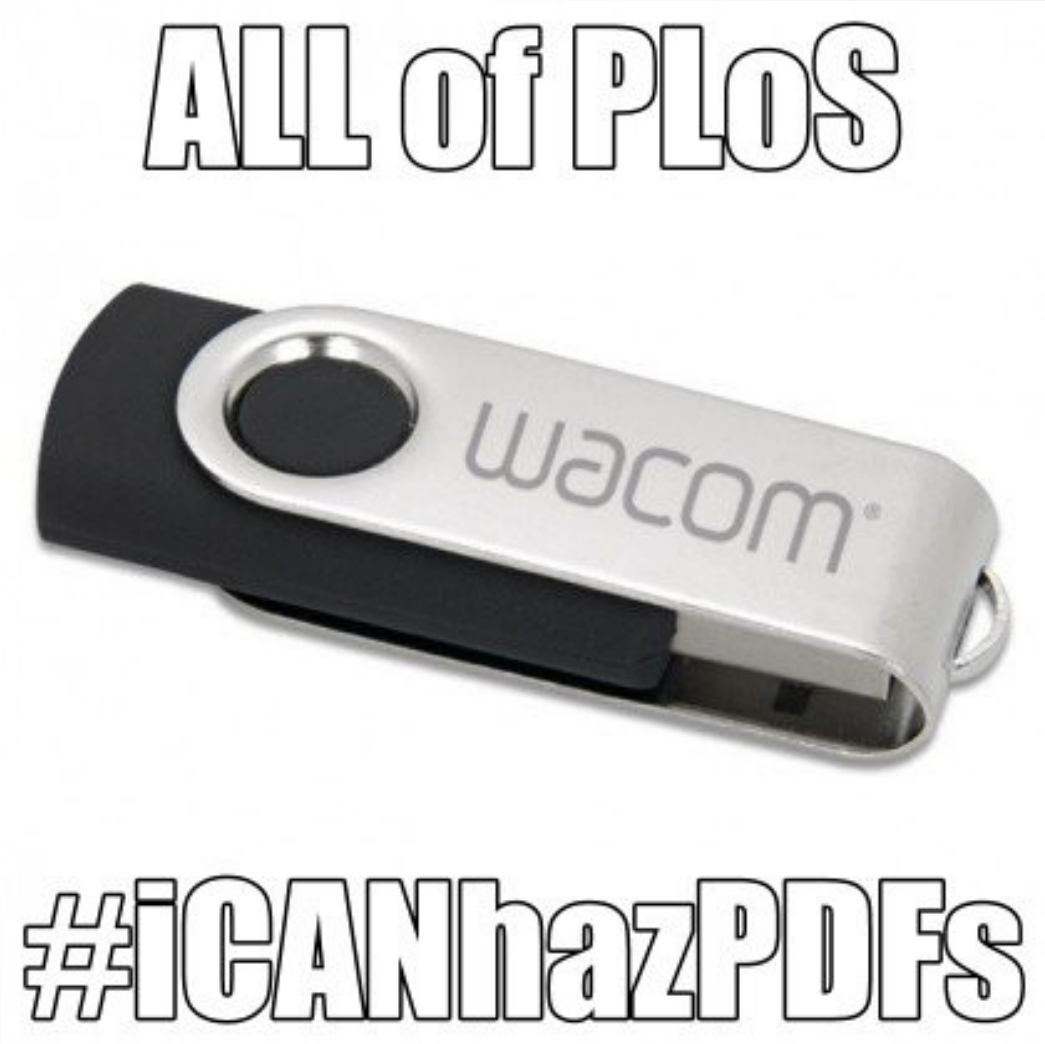
We have legitimate access, we do not seek to redistribute content wholesale. What's the problem? Why are we being criminalized?

OA publishers make it easy

One can easily download the entire content of many OA publishers e.g. PloS, BMC, Hindawi...

They *actively* facilitate & encourage corpus downloads

All of PLoS as PDF up to mid-2010 is just 15GB (~72,000 articles)



It's largely **only** OA publishers allowing content mining to occur

“When permission is requested [by researchers], 35% of publisher respondents allow mining in the majority or all of cases”

[tellingly, 85% of those “35% of publishers” that allow it are Open Access publishers, **not** subscription-access publishers]

from the Publishing Research Consortium's own report

(Smit & van der Graaf, 2011)

This is also the reason why only 13% of PMC is 'safe' to mine, despite 100% of PMC being 'free to read' (ocular access only)

For-profit publishers have incentives to actively block content mining

“53 % of publisher respondents will **decline mining requests** if the results can replace or compete with their own products and services”

from the Publishing Research Consortium's own report

(Smit & van der Graaf, 2011)

My POV: some publishers are clearly blocking the liberation of uncopyrightable **facts** from their content so they can continue making money from access to, or services around them.

N.B. >80% of research is public sector funded

The Hargreaves Report (UK)

Happily for me, the legal situation is about to change in the UK

(from October, 2013)

The UK government has accepted and is enforcing many of the changes recommended in:

'Digital Opportunity: A Review of Intellectual Property and Growth'

An independent report by Prof. Ian Hargreaves (2011)

Limitations and exceptions for non-commercial content mining research usage of copyright material will soon be legal.

I, the Open Knowledge Foundation, and many other groups think this is clearly the optimal, well-tested solution to the mining access/permission problem

Thanks

“But keep your minds open: maybe in some cases licensing won’t be the solution”

Neelie Kroes, Vice-President of the European Commission responsible for the Digital Agenda, 4th Feb 2013

Content mining for research is clearly one of those cases

Sincere thanks to the Open Knowledge Foundation for travel funding without which I wouldn't be here.