Why aggregating content is important for TDM, and what are the challenges in doing it right?

Petr Knoth
CORE (COnnecting REpositories)
Knowledge Media institute, The Open University

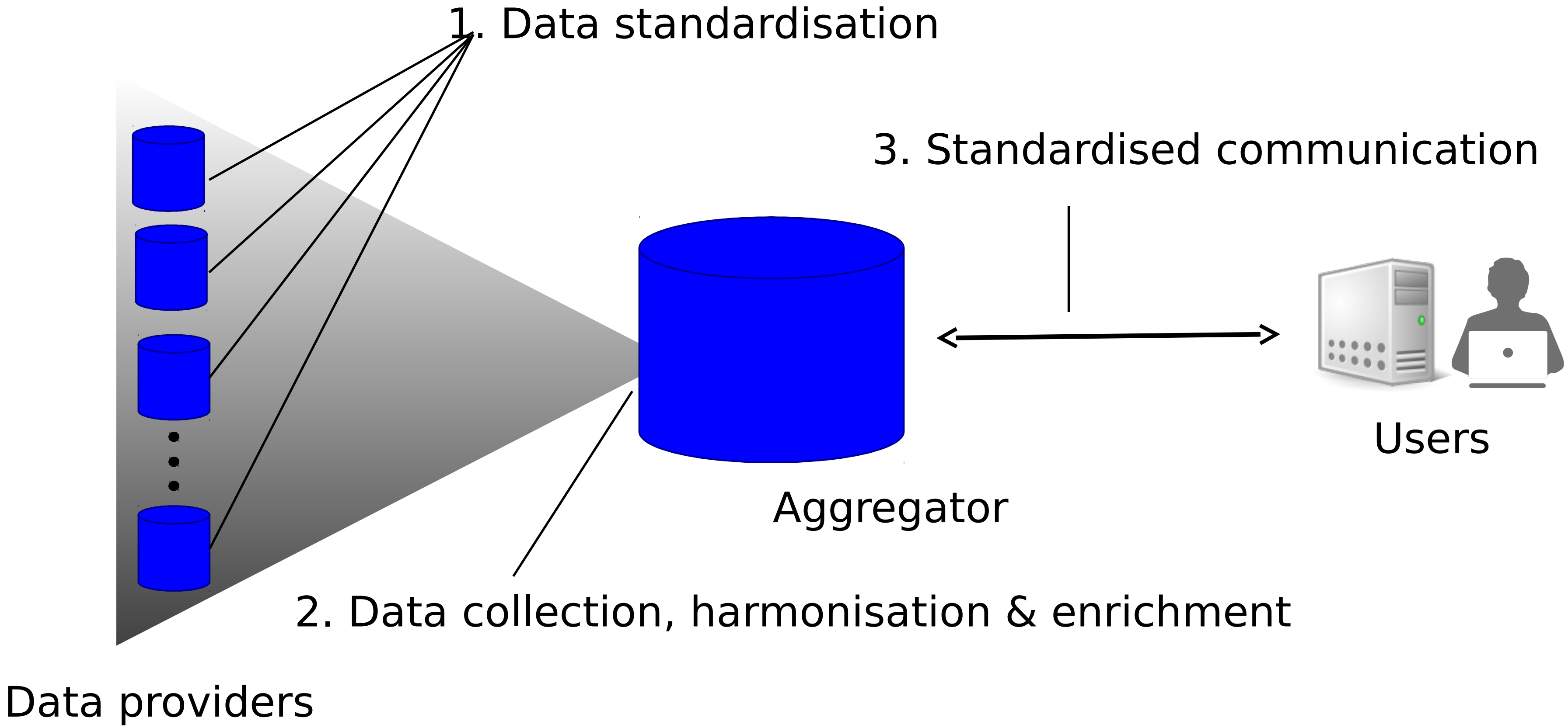
Thanks for allowing me to use your slides Petr! I have made some edits...

What is an aggregation?

Aggregators are intermediaries between providers and users that collect resources from many sources and add value by improving access to them.

Examples in the physical world: libraries, book stores, museums, art galleries, supermarkets
Examples in the digital world: digital libraries (e.g. PubMed) & collections (e.g. The European Library, Europeana), search engines (e.g. Google + Google Cache), newspaper aggregators (e.g. Google News), online retailers (e.g. Amazon), travel aggregators (e.g. Kayak), insurance aggregators (GoCompare), aggregators of research papers (e.g. CORE).

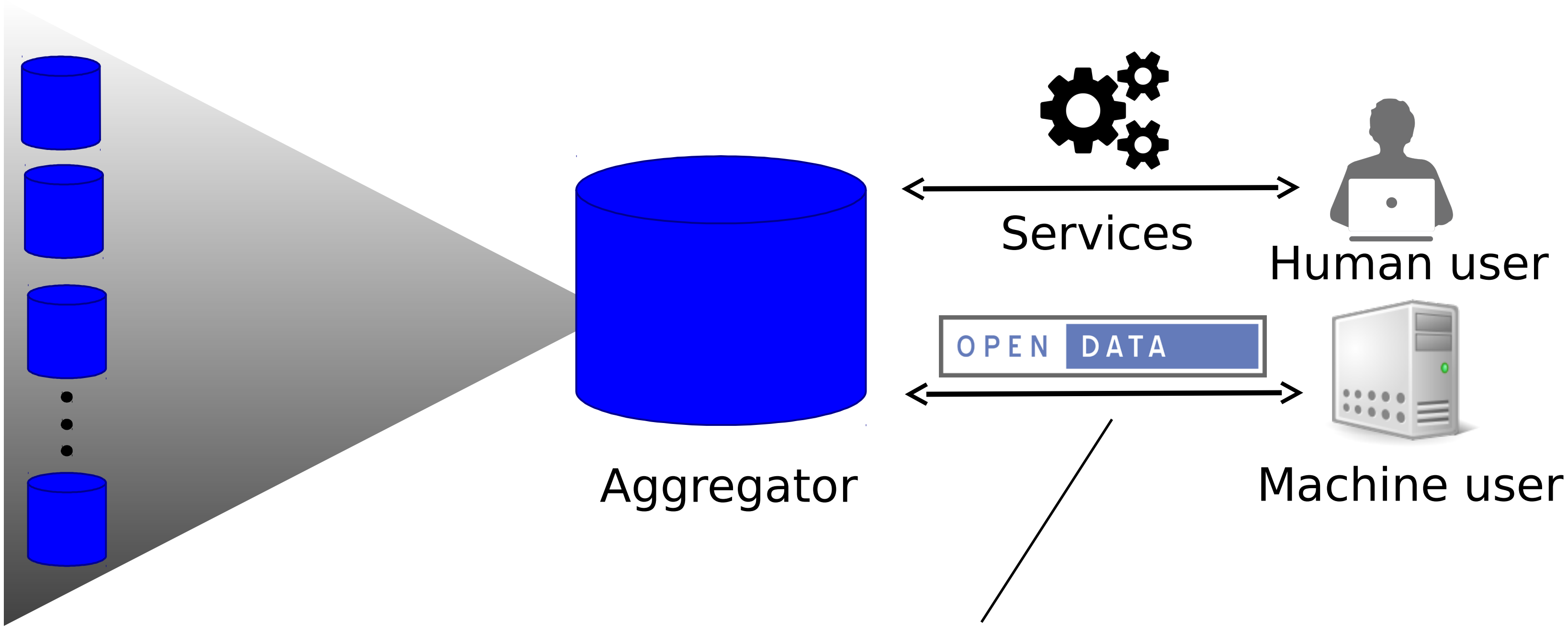
What is an aggregation?


1. Data standardisation
3. Standardised communication
Users
Aggregator
2. Data collection, harmonisation & enrichment
Data providers

Successful online aggregators add value

- Reduce the time to access information
- Standardise and harmonise content from many providers
- Enrich content with new information
- Provide harmonised access to users
- Enable the discovery of new information

Few aggregators provide unrestricted access to data



Services

Human user

Aggregator

Machine user

Data providers

Typically little or no support

Only few provide access to (open) data

- Enable the analysis of the content to generate new knowledge (TDM, business intelligence, data science)

- Enable the creation of new services and tools on top of aggregators (including TDM services).

Outline

1. What is aggregating?
2. **Why is it important for TDM**
3. Example: aggregating Open Access research papers
4. What are the challenges?

Why are aggregations important for TDM

- Up to 90% of a typical text-miner’s time is spent on gathering and harmonising the data. This is wasteful!
- Large scale aggregations are expensive, so can hardly be maintained professionally by individual text-miners => professional maintenance and wide sharing of the aggregated data needed.
- Reduce costs and enable the quick start of TDM projects.
- Enable the deployment of TDM solutions as services in real-world applications.

Outline

1. What is aggregating?
2. Why is it important for TDM
3. **Example: aggregating Open Access research papers**
4. What are the challenges?

What is Open Access exactly?

By “open access” to [peer-reviewed research literature], we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself.

[BOAI, 2002]

TDM for new scientific discoveries

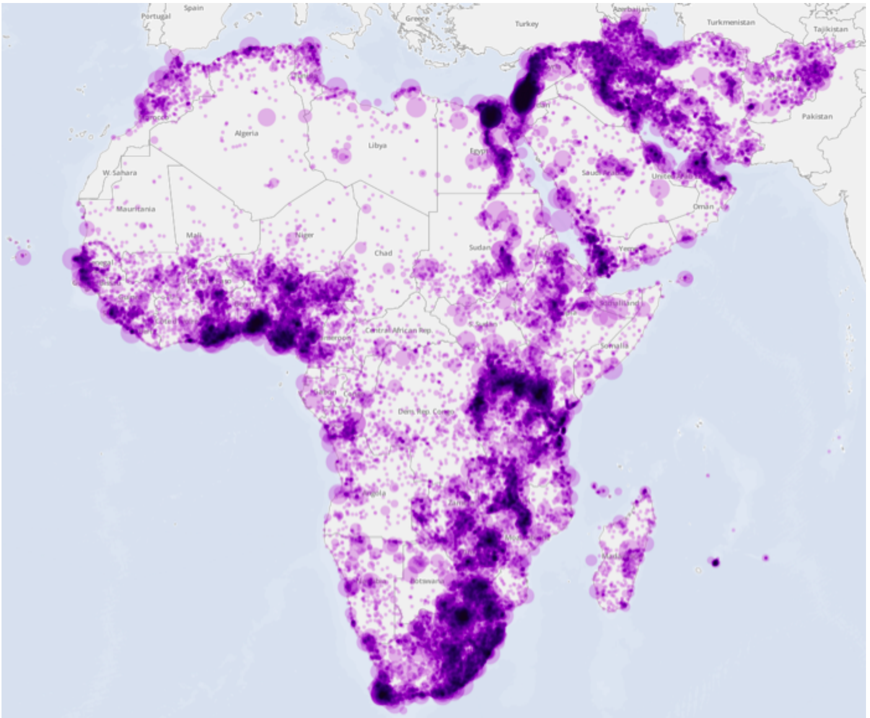
- Undiscovered Public Knowledge (Swanson, 1986).
- Mining of relationships for which there is “hidden” evidence in the research literature, yet they are not explicitly stated.
- Magnesium deficiency and migraine, fish oil and Raynaud's disease.
- Swanson's discoveries simulated by automated techniques (Weeber et al., 2001).

The mission of CORE

Aggregate all open access content distributed across different systems worldwide, enrich this content and provide access to it through a set of services ...

Example 1: Georgetown University researchers

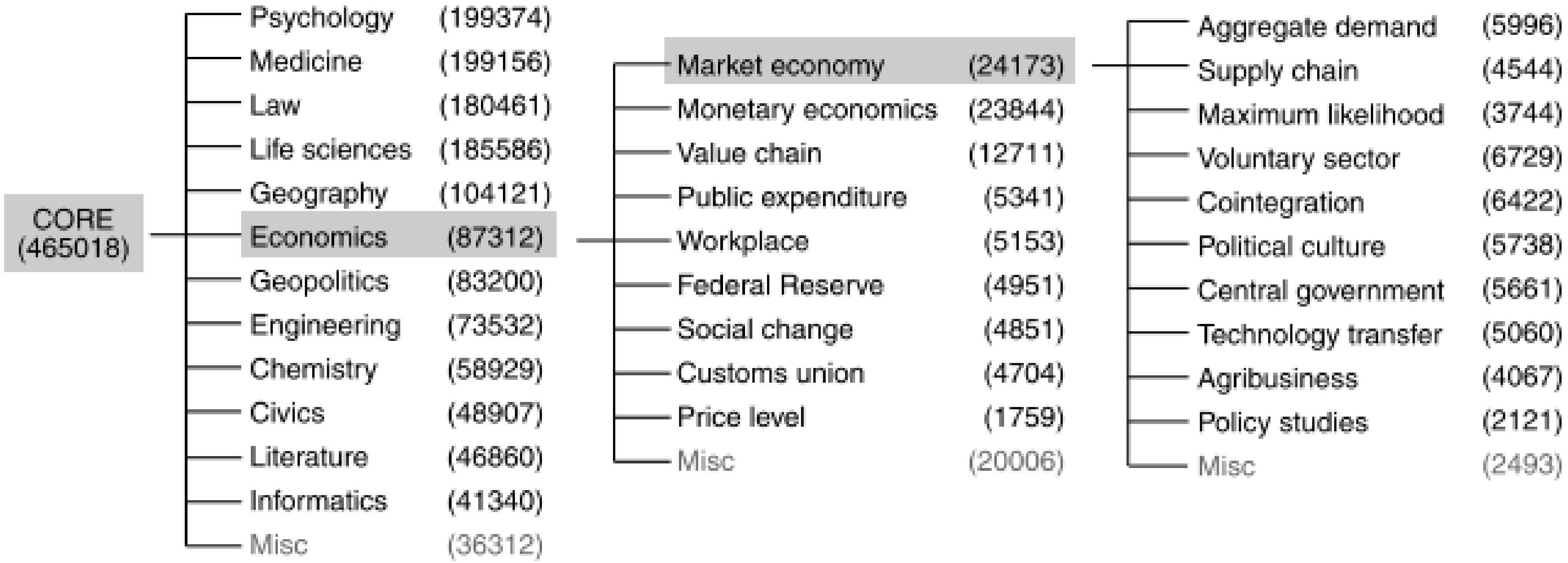
Analysing various popular political, social and cultural issues.



- Analysing the geographical coverage of scholarly literature.
- [Leetaru, Kalev H., Timothy K. Perkins, and Chris Rewerts. "Cultural Computing at Literature Scale." *D-Lib Magazine* 20.9/10 (2014).]

Example 2: Bauhaus University Weimar researchers

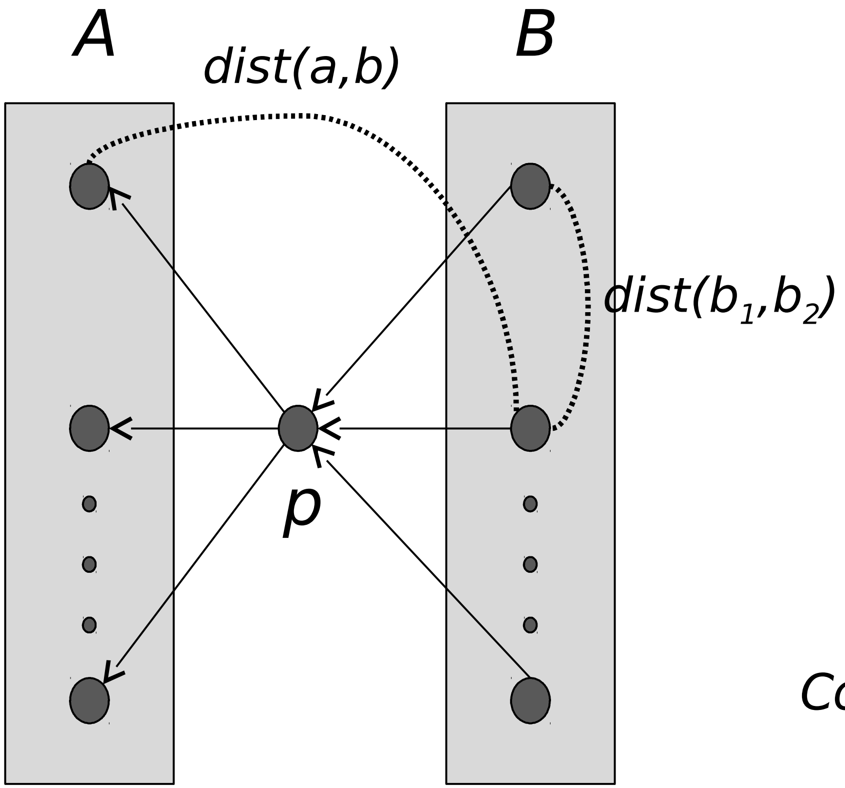
Automatically inducing a classification taxonomy using CORE data



- CORE (465018)
  - Psychology (199374)
  - Medicine (199156)
  - Law (180461)
  - Life sciences (185586)
  - Geography (104121)
  - Economics (87312)
    - Market economy (24173)
      - Aggregate demand (5996)
      - Supply chain (4544)
      - Maximum likelihood (3744)
      - Voluntary sector (6729)
      - Cointegration (6422)
      - Political culture (5738)
      - Central government (5661)
      - Technology transfer (5060)
      - Agribusiness (4067)
      - Policy studies (2121)
      - Misc (2493)
    - Monetary economics (23844)
    - Value chain (12711)
    - Public expenditure (5341)
    - Workplace (5153)
    - Federal Reserve (4951)
    - Social change (4851)
    - Customs union (4704)
    - Price level (1759)
    - Misc (20006)
  - Geopolitics (83200)
  - Engineering (73532)
  - Chemistry (58929)
  - Civics (48907)
  - Literature (46860)
  - Informatics (41340)
  - Misc (36312)

Volske, M., Gollub, T., Hagen, M. and Stein, B. (2014) A Key-Query Based Classification System for CORE.

Example 3: Semantometrics research



Premise: Full-text needed to assess publication’s research contribution.

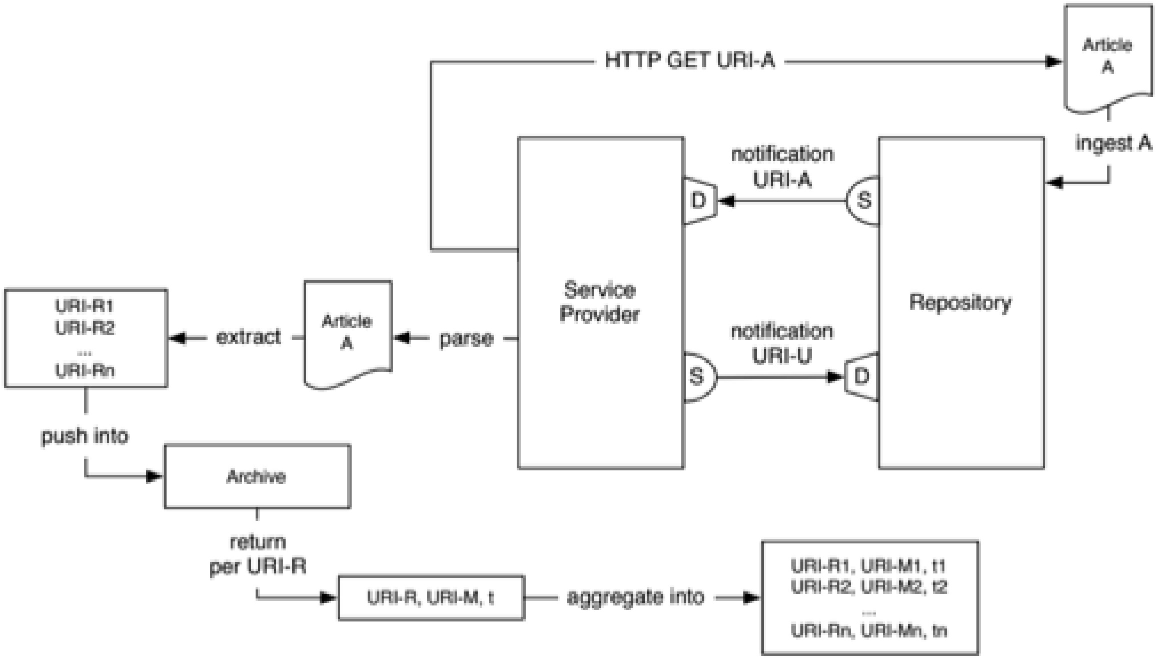
Hypothesis: Added value of publication *p* can be estimated based on the semantic distance from the publications cited by *p* to publications citing *p*.

$$Contribution(p) = \frac{B}{A} \times \frac{1}{|B| \times |A|} \times \sum_{a \in A, b \in B, a \neq b} dist(a,b)$$

Knoth, P. and Herrmannova, D. (2014) Towards Semantometrics: A New Semantic Similarity Based Measure for Assessing Research Contribution

Example 4: Los Alamos National Laboratory, University of Edinburgh

Archiving scholarly web resources, before they disappear ...



- High proportion of scholalry links are dead/do not point to its target after a few years.
- CORE used as a (repository) notification system
- Memento used as an archive

Klein, Martin, et al. "HiberActive: Pro-Active Archiving of Web References from Scholarly Articles." Open Repositories 2014 (2014).

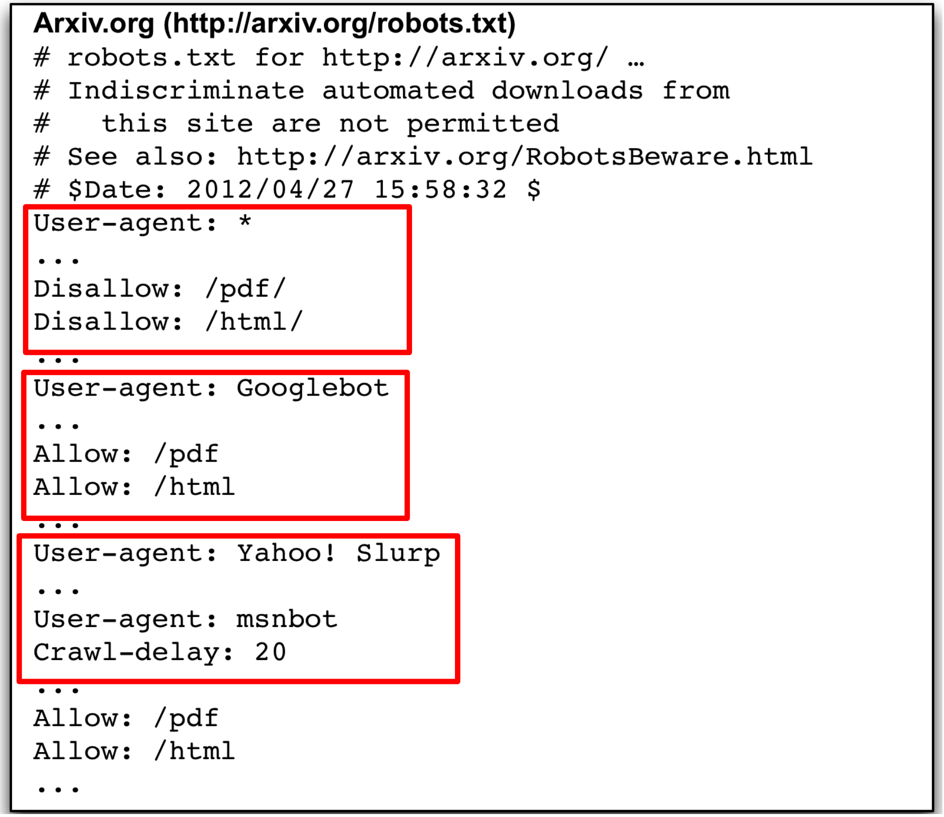
Outline

1. What is an aggregation?
2. Why is it important for TDM
3. Example: aggregating Open Access research papers
4. **What are the challenges?**

Challenges in aggregating for TDM

- Machine access to sources:
  - Restrictions on who can access
  - Limits on access frequency
- Scalability
- Lack of shared infrastructure
- Some organisations too risk-averse (Copyright)
- Misunderstandings with data providers
- Standardisation
- Financial resources (to run aggregations releasing Open Data)

Machine access restrictions



Arxiv.org (http://arxiv.org/robots.txt)

```
# robots.txt for http://arxiv.org/ …
# Indiscriminate automated downloads from
#   this site are not permitted
# See also: http://arxiv.org/RobotsBeware.html
# $Date: 2012/04/27 15:58:32 $
User-agent: *
...
Disallow: /pdf/
Disallow: /html/
...
User-agent: Googlebot
...
Allow: /pdf
Allow: /html
...
User-agent: Yahoo! Slurp
...
User-agent: msnbot
Crawl-delay: 20
...
Allow: /pdf
Allow: /html
...
```

Scalability issues – How Big is OA

- There are about 165M metadata records in OpenDOAR repositories => 16.4M green full-texts worldwide
- Green OA is about 21.1%, pure gold 6% and according to the study of (Laakso & Bjork, 2012) the proportion of hybrid gold is about 0.7% => 21M OA/public full-texts
- Rapid OA growth (size doubled in the last 4 years)
- Google Scholar estimated at 100M papers

Scalability issues – Infrastructure implications

- Average article about 1.8MB
- OA Storage: 36TB now, but possibly 72TB by 2016
- If the whole collection to be reharvested within 3 months: need to transfer and process 10MBit/s just for content 24/7
- High availability: need for architecture with no single point of failure (redundant and distributed)
- Europeana Cloud

Lack of shared infrastructure

- Many TDM algorithms should execute close to data to run efficiently.
- Not practical to transfer large datasets over the network.
- Need for a shared infrastructure for both data and TDM algorithms.

Risk-averse nature of some organisations

- Many (not-for-profit) organisations used to be worried about copyright infringement by TDM => the TDM experts, the best technology and resources are today in the commercial sector.
- 1st June 2014, UK TDM Copyright Exception - The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014, Regulation 3) => copies made for text-mining for non-commercial research purposes do not infringe copyright.

Misunderstandings on the issue of “control & credit”

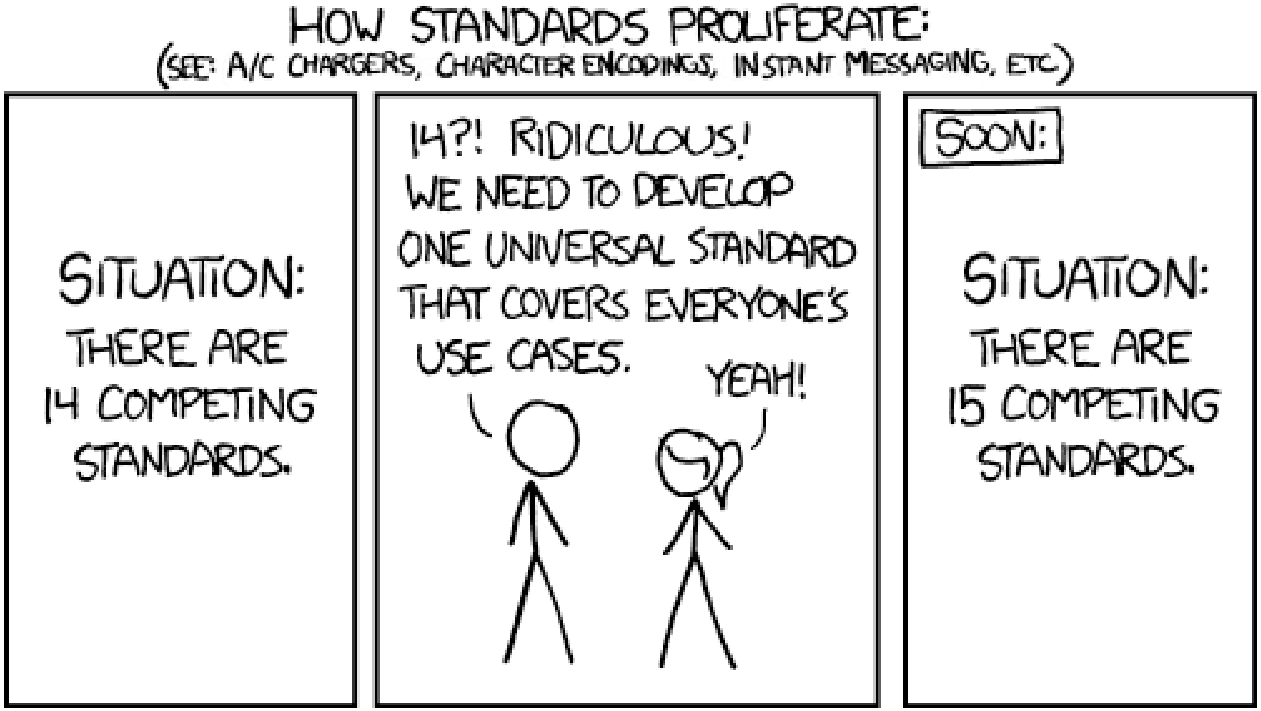
- Relationship between aggregators and content providers not always easy.

``I cannot say strongly enough: we support content mining. It can only work well if we are involved in the process and **managing the access**'’ – Richard Mollet (Head of the UK Publishers Association)

"*Aggregators and Google News are, to us, the worst offenders. They make money by living off the sweat of our brow.”*
https://www.techdirt.com/articles/20091014/1831246537.shtml

Standardisation

Standards do not exist, are not adopted, or are not consistently adopted or there are too many “standards” …


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

Financial resources (to run aggregations releasing open data)

- Typically few incentives for aggregators to release reusable Open Data, unless through paid for APIs.
- Access typically restricted to a fraction of the whole collection => makes TDM less useful.
- Aggregators worried opening data could undermine their business model.
- Open aggregators difficult to sustain in the long run, if investor interested primarily in profit.

Conclusions

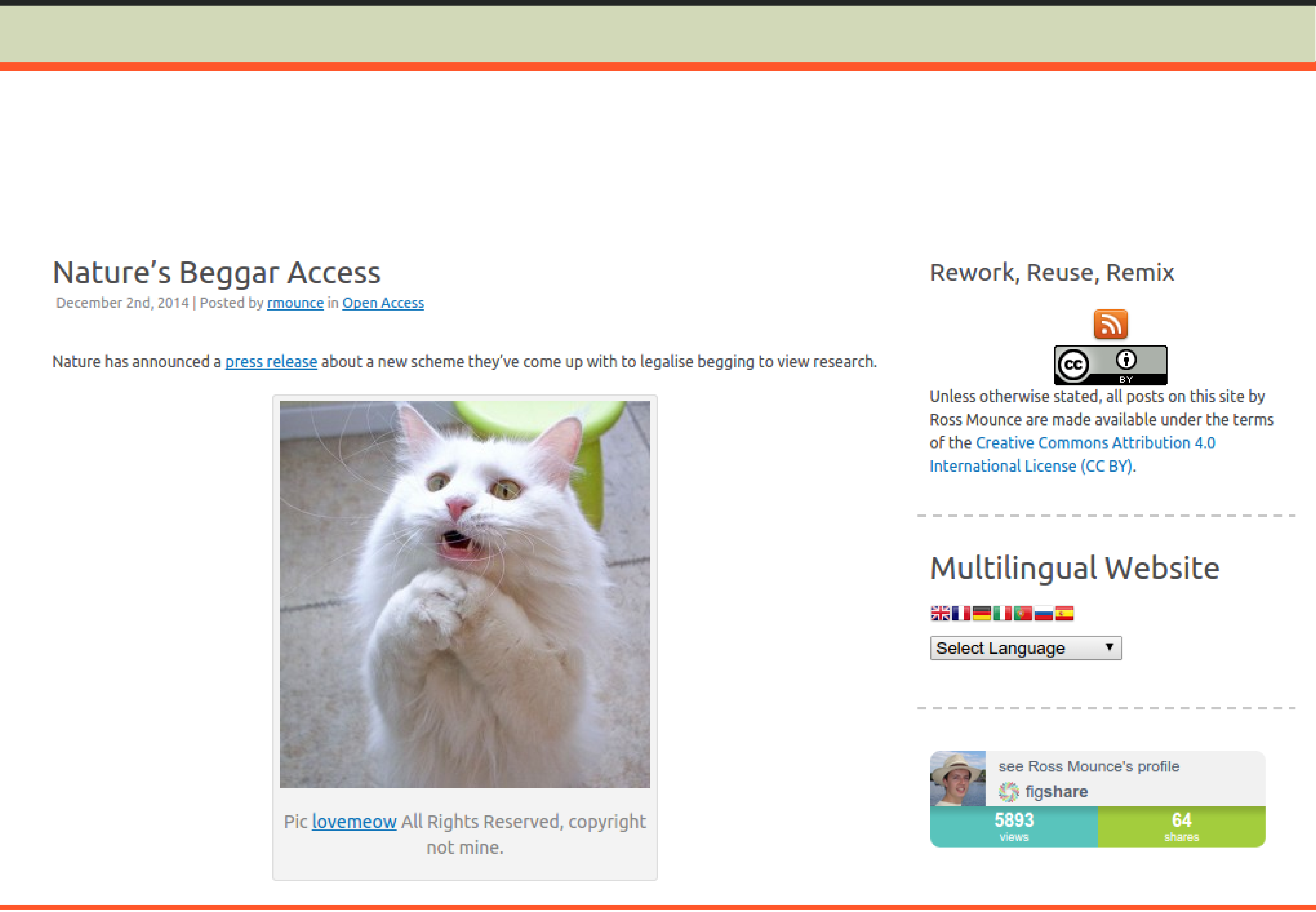
- TDM offers new possibilities for exploiting large amounts of textual resources.
- We need aggregators to make the ongoing process of data gathering efficient.
- We need to create incentives for data providers and aggregators to release open data and enable large scale machine processing.

Open Access is more than just free access

Do not get misled

The BOAI definition. See also 'BBB-definition'

http://www.budapestopenaccessinitiative.org/read



Nature’s Beggar Access

December 2nd, 2014 | Posted by rmounce in Open Access

Nature has announced a press release about a new scheme they’ve come up with to legalise begging to view research.

Pic lovemeow All Rights Reserved, copyright not mine.

Rework, Reuse, Remix

Unless otherwise stated, all posts on this site by Ross Mounce are made available under the terms of the Creative Commons Attribution 4.0 International License (CC BY).

Multilingual Website

Select Language ▼

see Ross Mounce's profile
figshare
5893
views
64
shares

Can blind or visually-impaired people make use of #BeggarAccess ? [compatibility with screen-readers]
No.

Can mobile users* make use of #BeggarAccess ?
No.

Can you print a paper copy from #BeggarAccess ?
No.

Can you copy/paste text, e.g. a quote from #BeggarAccess?
No.

* especially important in areas of the world where mobile phone internet access is the predominant form

We don't want Broken Access

We want Open Access