Not So Big Data

Crunching Wikipedia Referral Logs at Crossref

Joe Wass, Crossref
@joewass


WE'LL DO IT
LIVE
memegenerator.net

- Sorry
- for
- all
- the
- bullet
- points

Crossref

Crossref

- Set up in 2000
- Persistent IDs for scholarly publishing, mostly articles
- http://doi.org/10.5555/12345678
- The best way to cite
- Great metadata for scholarly works
- 80 million DOIs, 5,000 members (mostly publishers)
- DataCite also do DOIs for datasets

Wikipedia

References [edit]

1. ^ a b c "NIOSH Pocket Guide to Chemical Hazards #0167". National Institute for Occupational Safety and Health (NIOSH).
2. ^ "Cyclohexene". *Immediately Dangerous to Life and Health*. National Institute for Occupational Safety and Health (NIOSH).
3. ^ Michael T. Musser "Cyclohexanol and Cyclohexanone" in Ullmann's Encyclopedia of Industrial Chemistry, Wiley-VCH, Weinheim, 2005.doi:10.1002/14356007.a08_217
4. ^ Reed, Scott M.; Hutchison, James E. (2000). "Green Chemistry in the Organic Teaching Laboratory: An Environmentally Benign Synthesis of Adipic Acid". *J. Chem. Educ.* **77** (12): 1627–1629. doi:10.1021/ed077p1627.
5. ^ Jensen, Frederick R.; Bushweller, C. Hackett (1969). "Conformational preferences and interconversion barriers in cyclohexene and derivatives". *J. Am. Chem. Soc.* **91** (21): 5774–5782. doi:10.1021/ja01049a013.

- Citations all over the web
- Lots of citations to scholarly articles
- Big push to use DOIs in citations

History

- Crossref founded in 2000, about the same time as Wikipedia
- Some publishers that have been around for hundreds of years
- http://doi.org/10.1098/rstl.1672.0051 published 1672
- It's our job to keep track of traditional citations
- Things have changed a lot since 1672

Mr. Isaac Newtons *Answer to some Considerations upon his Doctrine of* Light *and* Colors ; *which Doctrine was printed in* Numb. 80. *of these Tracts.*

SIR, I have already told you, that at the perusal of the considerations, you sent me, on my Letter concerning *Refractions* and *Colors*, I found nothing, that, as I conceived, might not without difficulty be answer'd. And though I find the *Considerer* somewhat more concern'd for an *Hypothesis*, than I expected ; yet I doubt not, but we have one common design ; I mean, a sincere endeavour after knowledge, without valuing uncertain speculations for their subtleties, or despising certainties for their plainness : And on confidence of this it is, that I make this return to his discourse.*

* *Which* Discourse *was thought needless to be here printed at length, because in the body of this Answer are to be met with the chief particulars, wherein the* Answerer *was concern'd.*

1. *Of the Practique part of Opticques.*

The *first* thing that offers it self is less agreeable to me, and I begin with it because it is so. The considerer is pleased to reprehend me for laying aside the thoughts of improving Optiques by *Refractions*. If he had obliged me by a private Letter on this occasion, I would have acquainted him with my successes on the Tryals I have made of that kind, which I shall now say have been less than I sometimes expected, and perhaps than he at present hopes for. But since he is pleased to take it for granted, that I have let this subject pass without due examination, I shall refer him to my former Letter, * by which that conjecture will appear to be un-grounded. For, what I said there, was in respect of Telescopes of the ordinary construction, signifying, that their improvement is not to be expected from the *well-figuring* of Glasses, as Opticians have imagin'd ; but I despaired not of their improvement by other constructions ; which made me cautious to insert nothing that might

* *Printed in* Numb. 80. *of these Tracts.*

Tracking non-traditional scholarship

- Altmetrics
- Alternative metrics
- Article level metrics
- ALMs

Not another metric!?

No!

- But it's interesting to know what's going on.

Crossref Event Data

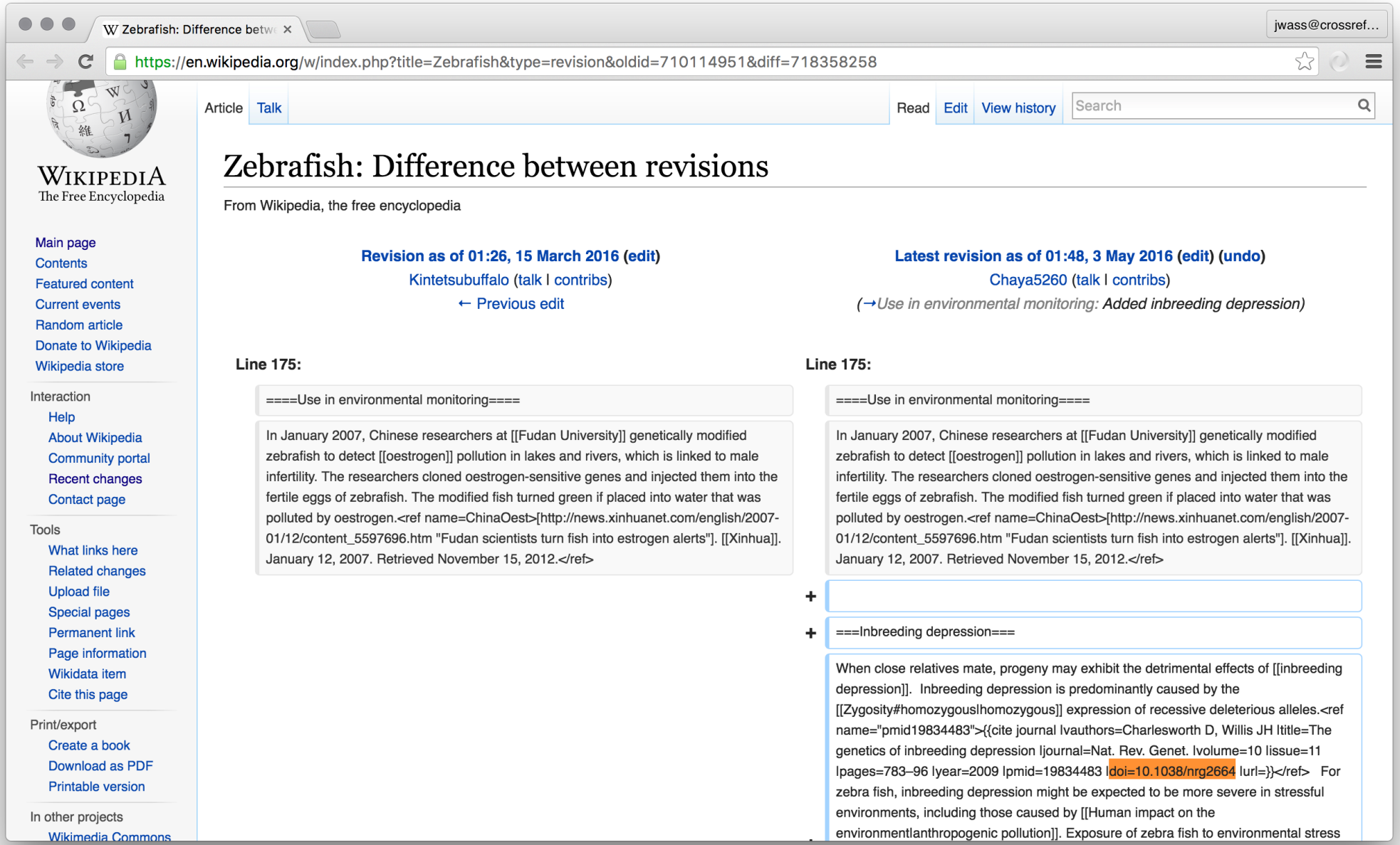
- Partnership with DataCite
- Early MVP stage
- “Tracking things that happen to our DOIs out in the wild”
- No-one tells us anything! Our DOIs never call!
- We want to change that (one way or another)

What happens on Wikipedia

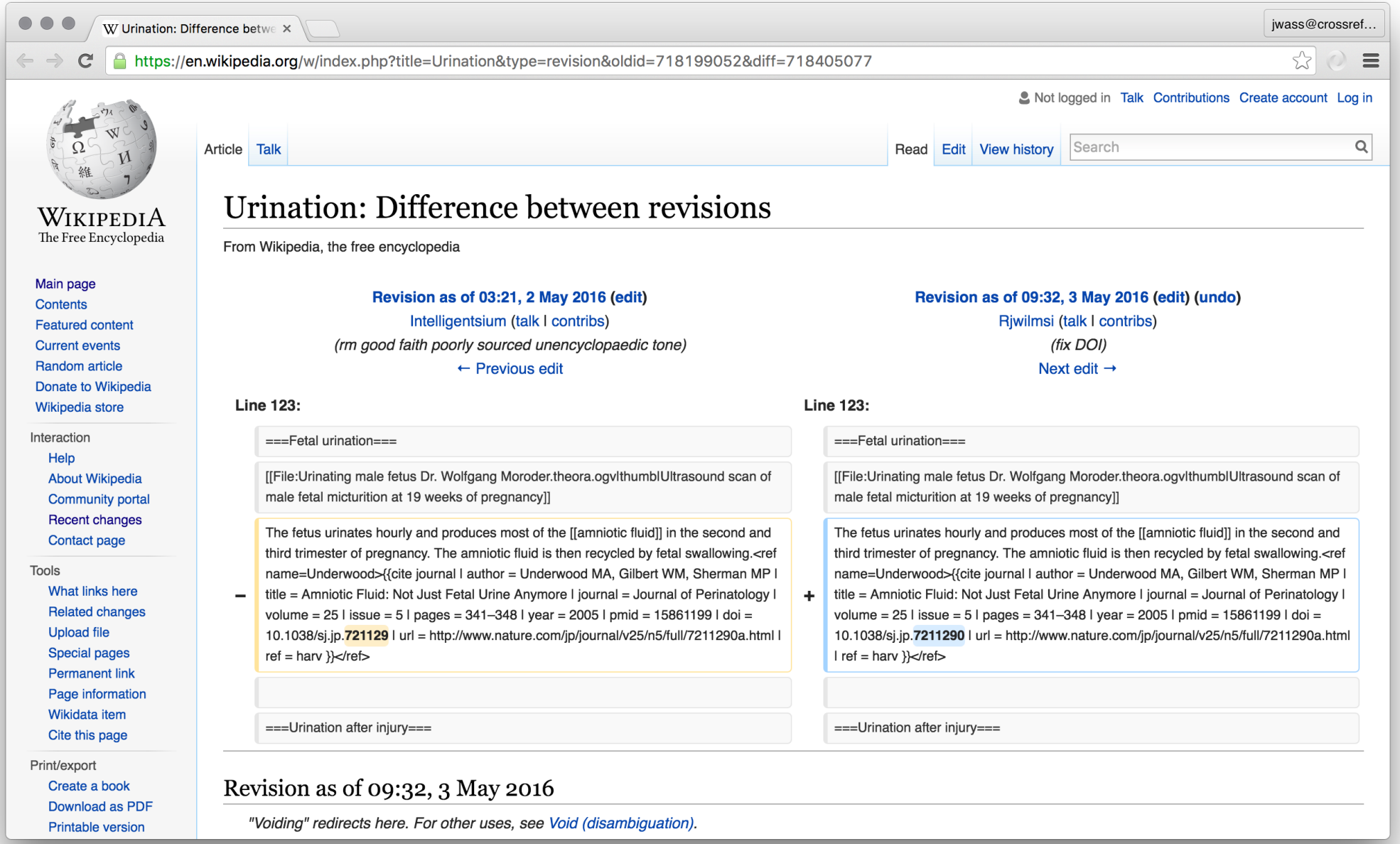
References [edit]

1. ^ a b c "NIOSH Pocket Guide to Chemical Hazards #0167". National Institute for Occupational Safety and Health (NIOSH).
2. ^ "Cyclohexene". *Immediately Dangerous to Life and Health*. National Institute for Occupational Safety and Health (NIOSH).
3. ^ Michael T. Musser "Cyclohexanol and Cyclohexanone" in Ullmann's Encyclopedia of Industrial Chemistry, Wiley-VCH, Weinheim, 2005.doi:10.1002/14356007.a08_217
4. ^ Reed, Scott M.; Hutchison, James E. (2000). "Green Chemistry in the Organic Teaching Laboratory: An Environmentally Benign Synthesis of Adipic Acid". *J. Chem. Educ.* **77** (12): 1627–1629. doi:10.1021/ed077p1627.
5. ^ Jensen, Frederick R.; Bushweller, C. Hackett (1969). "Conformational preferences and interconversion barriers in cyclohexene and derivatives". *J. Am. Chem. Soc.* **91** (21): 5774–5782. doi:10.1021/ja01049a013.

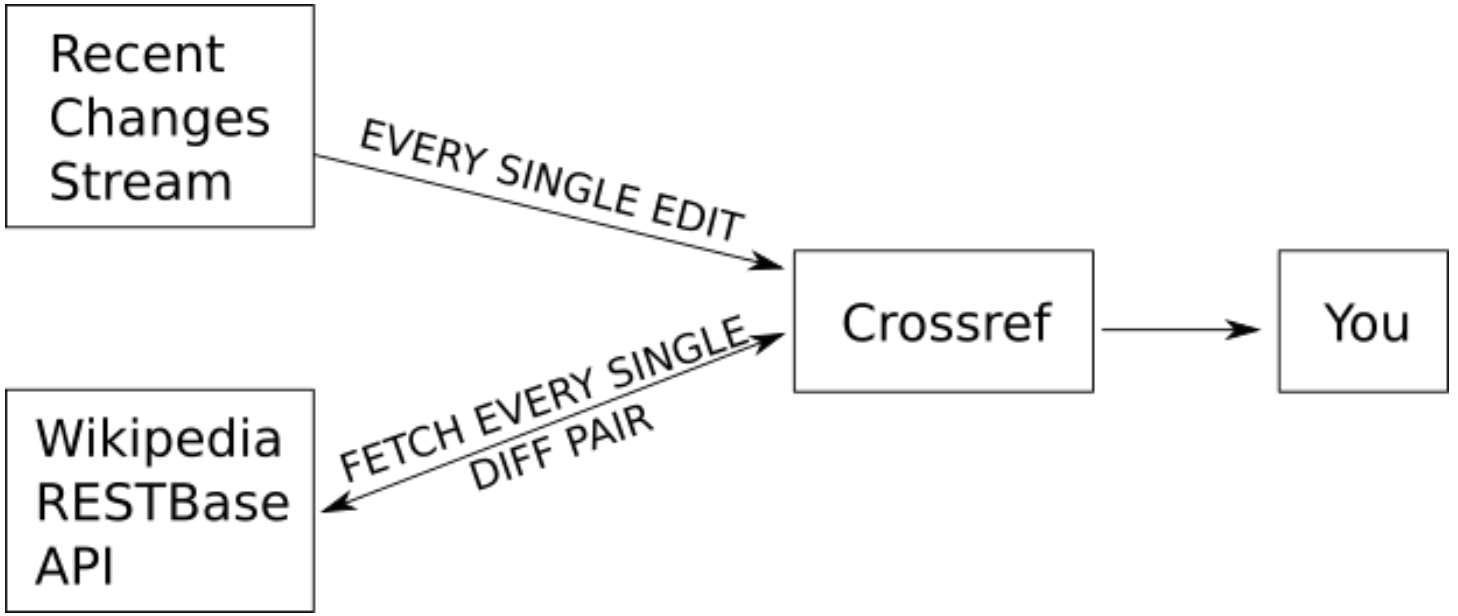
… doesn't stay on Wikipedia



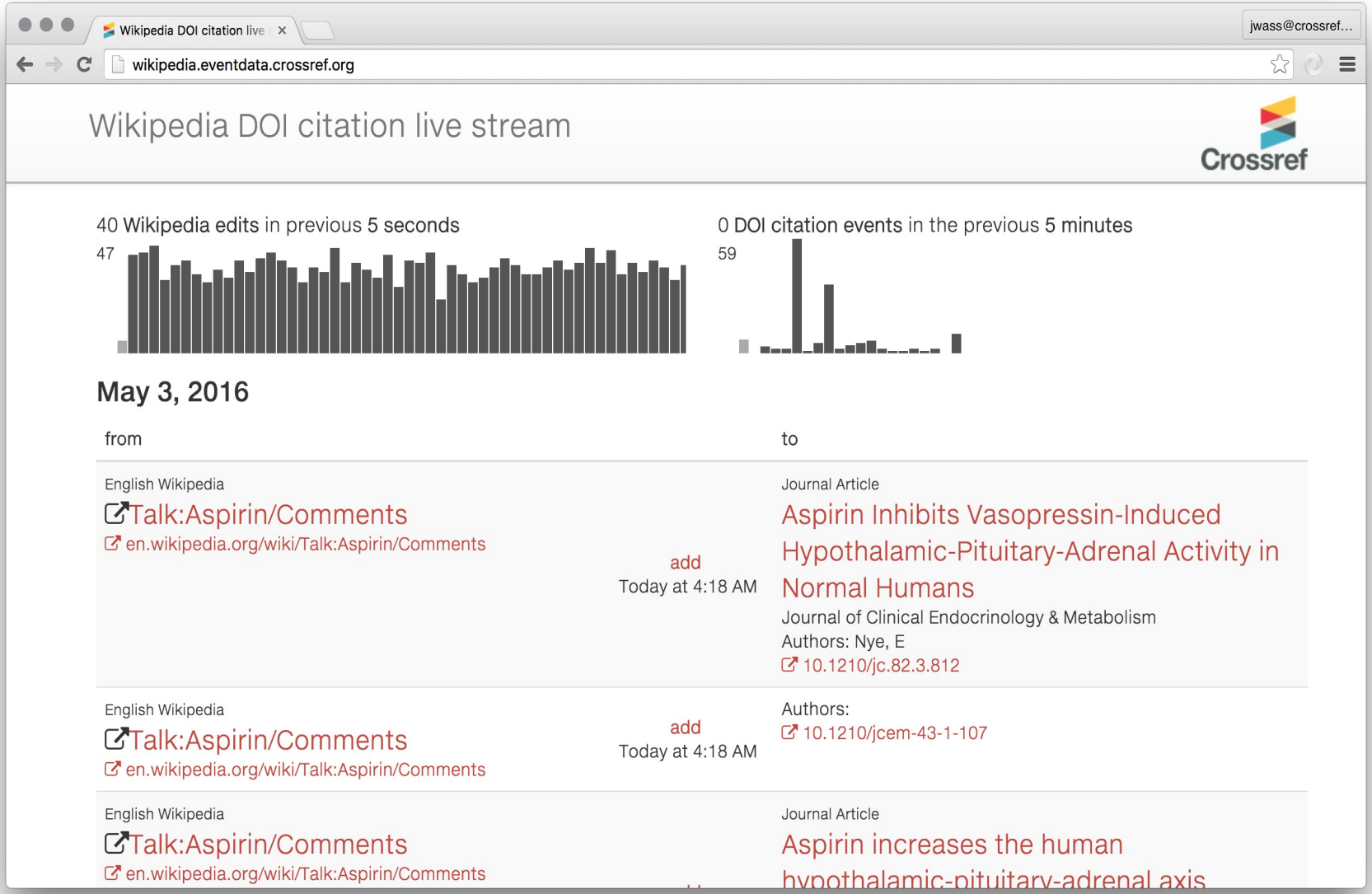
− that minoxidil-induced hair loss is a common side effect and describe the process as "shedding".	
− **[[Alcohol]] and [[propylene glycol]] present in some topical preparations may dry the scalp, resulting in [[dandruff]] and [[contact dermatitis]].<ref>{{cite web | url = http://www.medscape.com/viewarticle/407641 | title = Dandruff and Seborrheic Dermatitis | accessdate = 2009-10-09 | publisher = Medscape.com}}</ref> Some formulations of minoxidil substitute lipid Nanosomes in order to reduce contact dermatitis from the alcohol and propylene glycol vehicle.<ref>{{cite journal | author = Balakrishnan P, Shanmugam S, Lee WS, Lee WM, Kim JO, Oh DH, Kim DD, Kim JS, Yoo BK, Choi HG, Woo JS, Yong CS | title = Formulation and in vitro assessment of minoxidil Nanosomes for enhanced skin delivery | journal = International Journal of Phermaceutics | volume = 377 | pages = 1–8 | date = 1 February 2009 | pmid = 19394413 | doi = 10.1016/j.ijpharm.2009.04.020 | accessdate = 2012-06-30 | issue=1-2}}</ref><ref>{{cite journal | author = Padois K, Cantiéni C, Bertholle V, Bardel C, Pirot F, Falson F | title = Solid lipid nanoparticles suspension versus commercial solutions for dermal delivery of minoxidil | journal = International Journal of Pharmaceutics | volume = 416 | pages = 300–304 | date = 16 June 2011 | pmid = 21704140 | doi = 10.1016/j.ijpharm.2011.06.014 | accessdate = 2012-06-30 | issue=1}}</ref>**	+ **'''Customer Reviews'''**
	+ New Hair Growth within weeks Review by JazzyB
− **Side effects of "oral" minoxidil may include swelling of the face and extremities, rapid and irregular heartbeat, lightheadedness, cardiac lesions, and focal [[necrosis]] of the [[papillary muscle]] and subendocardial areas of the left ventricle.<ref name="Drugs.com"/> There have been cases of allergic reactions to minoxidil or the non-active ingredient propylene glycol, which is found in some topical minoxidil formulations. [[Pseudoacromegaly]] is an extremely rare side effect reported with large doses of oral minoxidil.<ref>{{cite journal | last1 = Nguyen | first1 = K. | last2 = Marks Jr | first2 = J. | title = Pseudoacromegaly induced by the long-term use of minoxidil | journal = Journal of the American Academy of Dermatology | volume = 48 | issue = 6 | pages = 962–965 | year = 2003 | pmid = 12789195 | doi = 10.1067/mjd.2003.325 }}</ref>**	+ **I have been using Minoxidil along with DR hair loss shampoo for two months now. I can clearly see new hair growth, dandruff is down to none, hair feel revitalized and fuller. Scalp has a fresh feeling and best of all I got my confidence back. The only downside to minoxidil treatment is that you have to use it continuous or the hair fall will happen again. But then again this is a small price to pay for getting your hair back. (Posted on 4/8/2015)**


Urination: Difference betwe
jwass@crossref...
https://en.wikipedia.org/w/index.php?title=Urination&type=revision&oldid=718199052&diff=718405077
Not logged in Talk Contributions Create account Log in
Article Talk
Read Edit View history
Search
WIKIPEDIA
The Free Encyclopedia
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page
Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page
Print/export
Create a book
Download as PDF
Printable version
Urination: Difference between revisions
From Wikipedia, the free encyclopedia
Revision as of 03:21, 2 May 2016 (edit)
Intelligentsium (talk | contribs)
(rm good faith poorly sourced unencyclopaedic tone)
← Previous edit
Revision as of 09:32, 3 May 2016 (edit) (undo)
Rjwilmsi (talk | contribs)
(fix DOI)
Next edit →
Line 123:
Line 123:
===Fetal urination===
[[File:Urinating male fetus Dr. Wolfgang Moroder.theora.ogv|thumb|Ultrasound scan of male fetal micturition at 19 weeks of pregnancy]]
The fetus urinates hourly and produces most of the [[amniotic fluid]] in the second and third trimester of pregnancy. The amniotic fluid is then recycled by fetal swallowing.<ref name=Underwood>{{cite journal | author = Underwood MA, Gilbert WM, Sherman MP | title = Amniotic Fluid: Not Just Fetal Urine Anymore | journal = Journal of Perinatology | volume = 25 | issue = 5 | pages = 341–348 | year = 2005 | pmid = 15861199 | doi = 10.1038/sj.jp.721129 | url = http://www.nature.com/jp/journal/v25/n5/full/7211290a.html | ref = harv }}</ref>
The fetus urinates hourly and produces most of the [[amniotic fluid]] in the second and third trimester of pregnancy. The amniotic fluid is then recycled by fetal swallowing.<ref name=Underwood>{{cite journal | author = Underwood MA, Gilbert WM, Sherman MP | title = Amniotic Fluid: Not Just Fetal Urine Anymore | journal = Journal of Perinatology | volume = 25 | issue = 5 | pages = 341–348 | year = 2005 | pmid = 15861199 | doi = 10.1038/sj.jp.7211290 | url = http://www.nature.com/jp/journal/v25/n5/full/7211290a.html | ref = harv }}</ref>
===Urination after injury===
===Urination after injury===
Revision as of 09:32, 3 May 2016
"Voiding" redirects here. For other uses, see Void (disambiguation).

Time to investigate



wikipedia.eventdata.crossref.org



- Wikipedia is important for DOI Citations
- We should pay close attention!
- There's an edit every few minutes that concerns a DOI somehow
- Citations come and go
- Different to traditional citation
- But do people follow the citations?

Big Data

- Wouldn't it be cool to have a Big Data problem?
- DOI Resolution Logs
- ~1.5 Terabytes for 5 years of logs
- Every time a DOI is clicked…
- … by a human
- … by a machine
- … by something else
- Not necessarily human activity
- A indicator of existence and proxy of use
- Check the blog for full details

```
184.73.49.84 HTTP:HDL "2015-01-01 00:00:01.084Z" 1 1 108ms 10.1038/nature14019 "200:0.na/10.1038" ""

201.80.126.36 HTTP:HDL "2015-01-01 00:00:01.297Z" 1 1 189ms 10.1179/2047058414Y.0000000138
 "200:0.na/10.1179" "http://buscatextual.cnpq.br/buscatextual/visualizacv.do?metodo=apresentar&id=K4763600A7"

68.180.228.178 HTTP:HDL "2015-01-01 00:00:01.333Z" 1 1 209ms 10.1016/j.geomorph.2008.01.009 "200:0.na/10.1016" ""

100.43.85.19 HTTP:HDL "2015-01-01 00:00:01.406Z" 1 1 198ms 10.1016/j.jnutbio.2008.10.008 "200:0.na/10.1016" ""

68.180.228.178 HTTP:HDL "2015-01-01 00:00:01.562Z" 1 1 174ms 10.1080/14650040601031230 "200:0.na/10.1080" ""

68.180.228.178 HTTP:HDL "2015-01-01 00:00:01.655Z" 1 1 187ms 10.1111/j.1365-313X.2010.04248.x "200:0.na/10.1111" ""

194.140.240.28 HTTP:HDL "2015-01-01 00:00:01.764Z" 1 1 175ms 10.1787/5k3xz6hc2z0x-en "200:0.na/10.1787"
 "http://www.oecd.org/tax/index.xml"

205.142.197.84 HTTP:HDL "2015-01-01 00:00:01.782Z" 1 1 277ms 10.1007/s12026-014-8526-z "200:0.na/10.1007"
 "http://www.ncbi.nlm.nih.gov/pubmed/24838142"

207.46.13.58 HTTP:HDL "2015-01-01 00:00:02.026Z" 1 1 268ms 10.1002/1097-0142(19921015)70:8%3C2121::AID-CNCR2820700819%3E3.0.CO;2-S
 "200:0.na/10.1002" ""

192.138.89.30 HTTP:HDL "2015-01-01 00:00:02.256Z" 1 1 196ms 10.1016/j.jcrimjus.2010.04.019 "200:0.na/10.1016"
 "http://sfx.carli.illinois.edu/sfxwhe?genre=article&isbn=&issn=00472352&title=Journal%20of%20Criminal%20Justice&volume=38&issue=4&date=
```

Analysis

Gather stats for

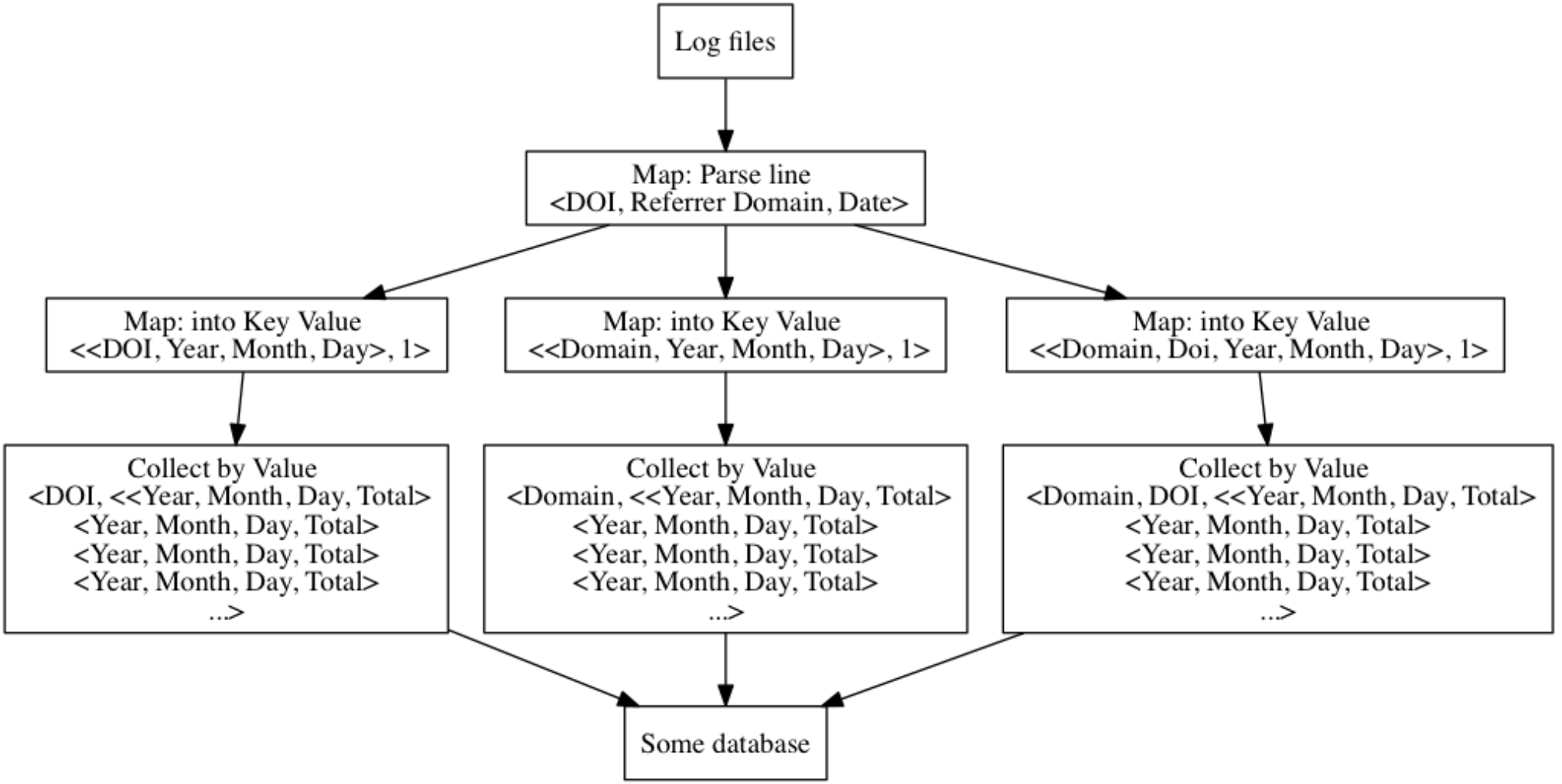
- DOI
- Date
- Referrer domain if present
- Referrer HTTP or HTTPs, if there is a referrer

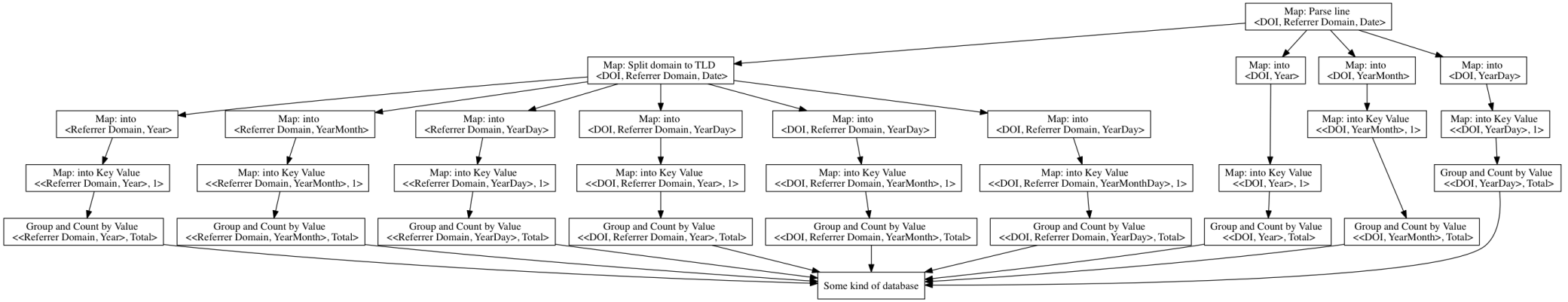
Remove

- IP address
- Precise referrer URL
- Precise time
- Hat-tip to Zara's talk

Apache Spark

- Big data tool
- Map Reduce, but more flexible
- Graph of transforms
- Clever partitioning into Resilient Distributed Datasets
- Partitions of data in streams, distributed to nodes


Log files
Map: Parse line
<DOI, Referrer Domain, Date>
Map: into Key Value
<<DOI, Year, Month, Day>, 1>
Map: into Key Value
<<Domain, Year, Month, Day>, 1>
Map: into Key Value
<<Domain, Doi, Year, Month, Day>, 1>
Collect by Value
<DOI, <<Year, Month, Day, Total>
<Year, Month, Day, Total>
<Year, Month, Day, Total>
<Year, Month, Day, Total>
...>
Collect by Value
<Domain, <<Year, Month, Day, Total>
<Year, Month, Day, Total>
<Year, Month, Day, Total>
<Year, Month, Day, Total>
...>
Collect by Value
<Domain, DOI, <<Year, Month, Day, Total>
<Year, Month, Day, Total>
<Year, Month, Day, Total>
<Year, Month, Day, Total>
...>
Some database


Map: Parse line
<DOI, Referrer Domain, Date>
Map: Split domain to TLD
<DOI, Referrer Domain, Date>
Map: into
<DOI, Year>
Map: into
<DOI, YearMonth>
Map: into
<DOI, YearDay>
Map: into
<Referrer Domain, Year>
Map: into
<Referrer Domain, YearMonth>
Map: into
<Referrer Domain, YearDay>
Map: into
<DOI, Referrer Domain, YearDay>
Map: into
<DOI, Referrer Domain, YearDay>
Map: into
<DOI, Referrer Domain, YearDay>
Map: into Key Value
<<DOI, YearMonth>, 1>
Map: into Key Value
<<DOI, YearDay>, 1>
Map: into Key Value
<<Referrer Domain, Year>, 1>
Map: into Key Value
<<Referrer Domain, YearMonth>, 1>
Map: into Key Value
<<Referrer Domain, YearDay>, 1>
Map: into Key Value
<<DOI, Referrer Domain, Year>, 1>
Map: into Key Value
<<DOI, Referrer Domain, YearMonth>, 1>
Map: into Key Value
<<DOI, Referrer Domain, YearMonthDay>, 1>
Map: into Key Value
<<DOI, Year>, 1>
Group and Count by Value
<<DOI, YearDay>, Total>
Group and Count by Value
<<Referrer Domain, Year>, Total>
Group and Count by Value
<<Referrer Domain, YearMonth>, Total>
Group and Count by Value
<<Referrer Domain, YearDay>, Total>
Group and Count by Value
<<DOI, Referrer Domain, Year>, Total>
Group and Count by Value
<<DOI, Referrer Domain, YearMonth>, Total>
Group and Count by Value
<<DOI, Referrer Domain, YearDay>, Total>
Group and Count by Value
<<DOI, Year>, Total>
Group and Count by Value
<<DOI, YearMonth>, Total>
Some kind of database

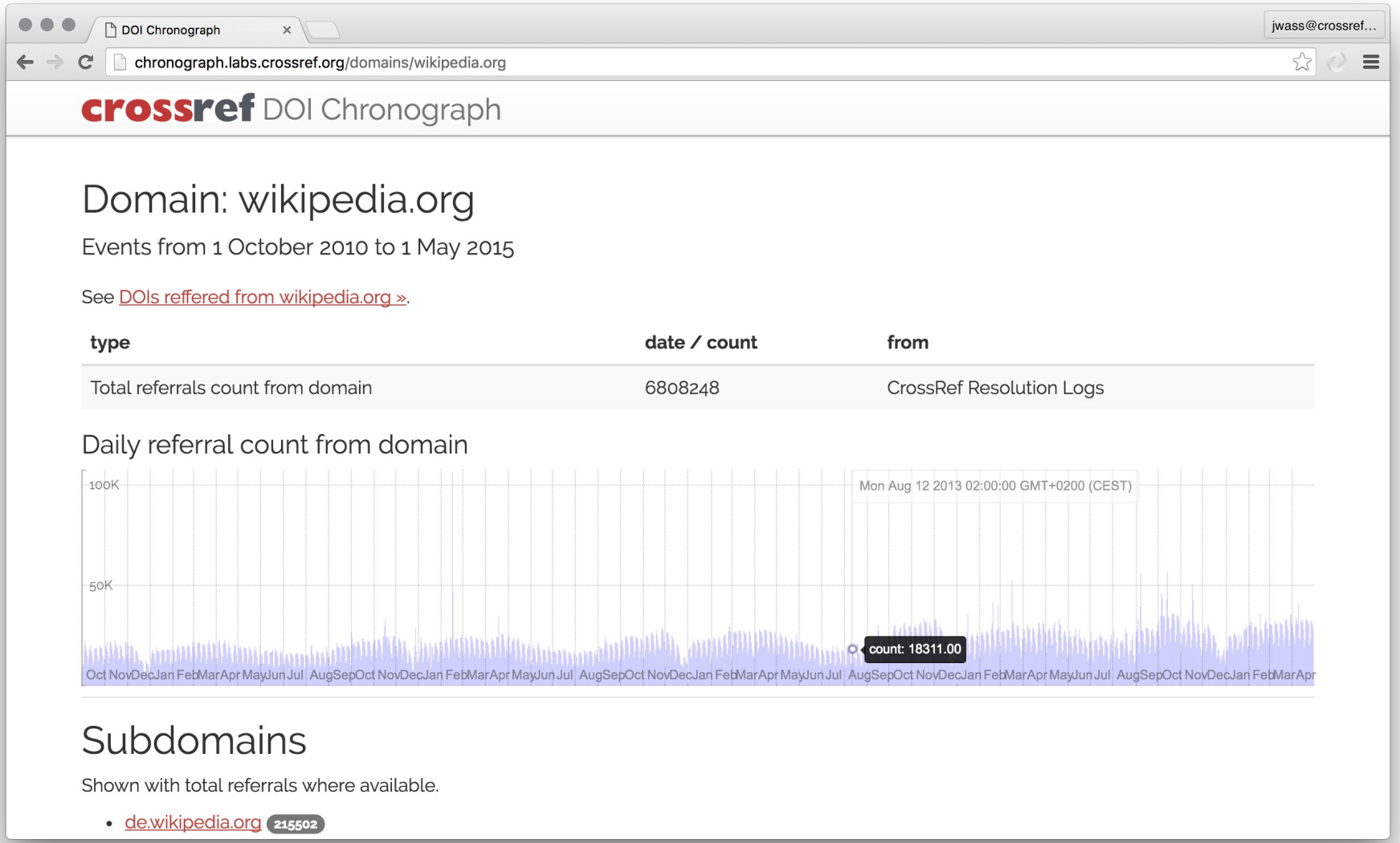
Magic!

- Write code in Scala / Clojure
- Works on my laptop
- Scales up to multi-node distribution in the cloud! I'm doing Big Data!
- Sometimes it would fail after 12 hours
- Heap explodes after a long time, slow to refine
- One by one nodes might die
- Fetching 10s of GB of data from S3 and unzipping slow painful
- EC2 is cheap, but gets expensive
- Got the results!

DOI Chronograph

- Stats for all DOIs and all referral domains
- Showed that Wikipedia was the 5th largest non-traditional referrer
- Over a year ago

DOI Chronograph



DOI Chronograph

jwass@crossref...

chronograph.labs.crossref.org/subdomains/en.wikipedia.org

crossref DOI Chronograph

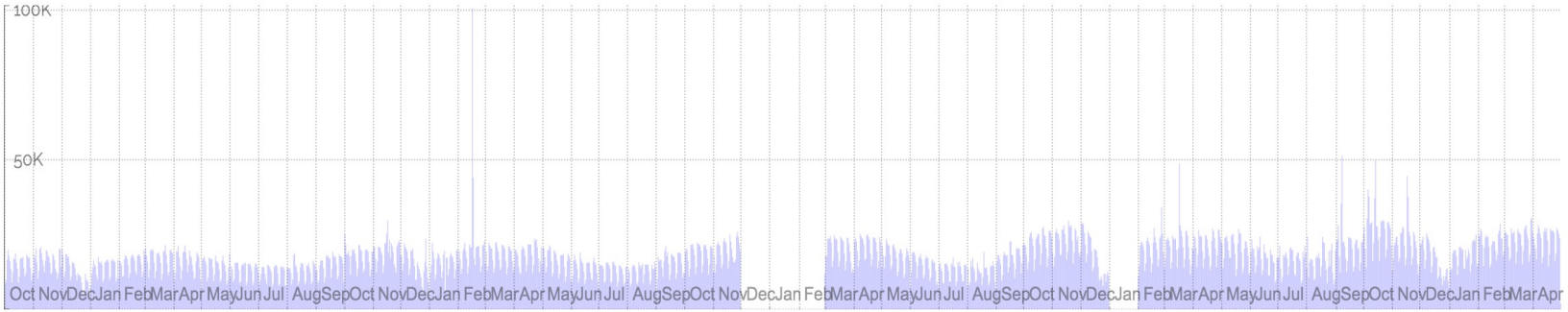
Subdomain: en.wikipedia.org

Main domain: wikipedia.org.

Events from 1 October 2010 to 1 May 2015

type	date / count	from
Total referrals count from subdomain	1	CrossRef Resolution Logs

Daily referral count from domain



Subdomains

Shown with total referrals where available.

- de.wikipedia.org 215502

DOI Chronograph

jwass@crossref...

chronograph.labs.crossref.org/subdomains/en.m.wikipedia.org

crossref DOI Chronograph

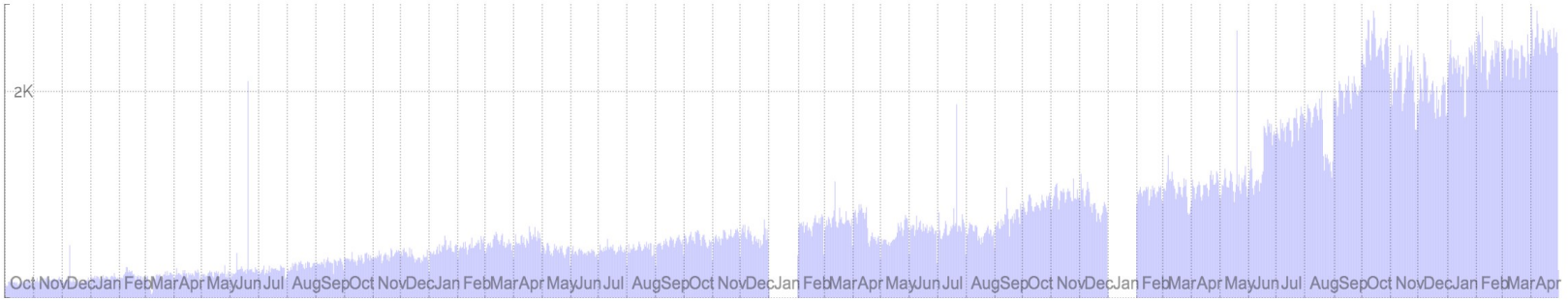
Subdomain: en.m.wikipedia.org

Main domain: wikipedia.org.

Events from 1 October 2010 to 1 May 2015

type	date / count	from
Total referrals count from subdomain	190455	CrossRef Resolution Logs

Daily referral count from domain



Subdomains

Shown with total referrals where available.

- de.wikipedia.org 215502

An interesting twist

Changes to Wikipedia

- Snowden, government censorship, Russia
- Gradual change to HTTPS only
- Browsers don't send referrer data from HTTPS to HTTP
- Suddenly all this valuable data evaporates

HTTPS for all!

- DOI supports HTTPS
- But only if the link is HTTPS
- Change to use protocol-relative DOIs
- Crossref collaborating with Wikipedia Research
- Data valuable to everyone
- Time for some reanalysis!

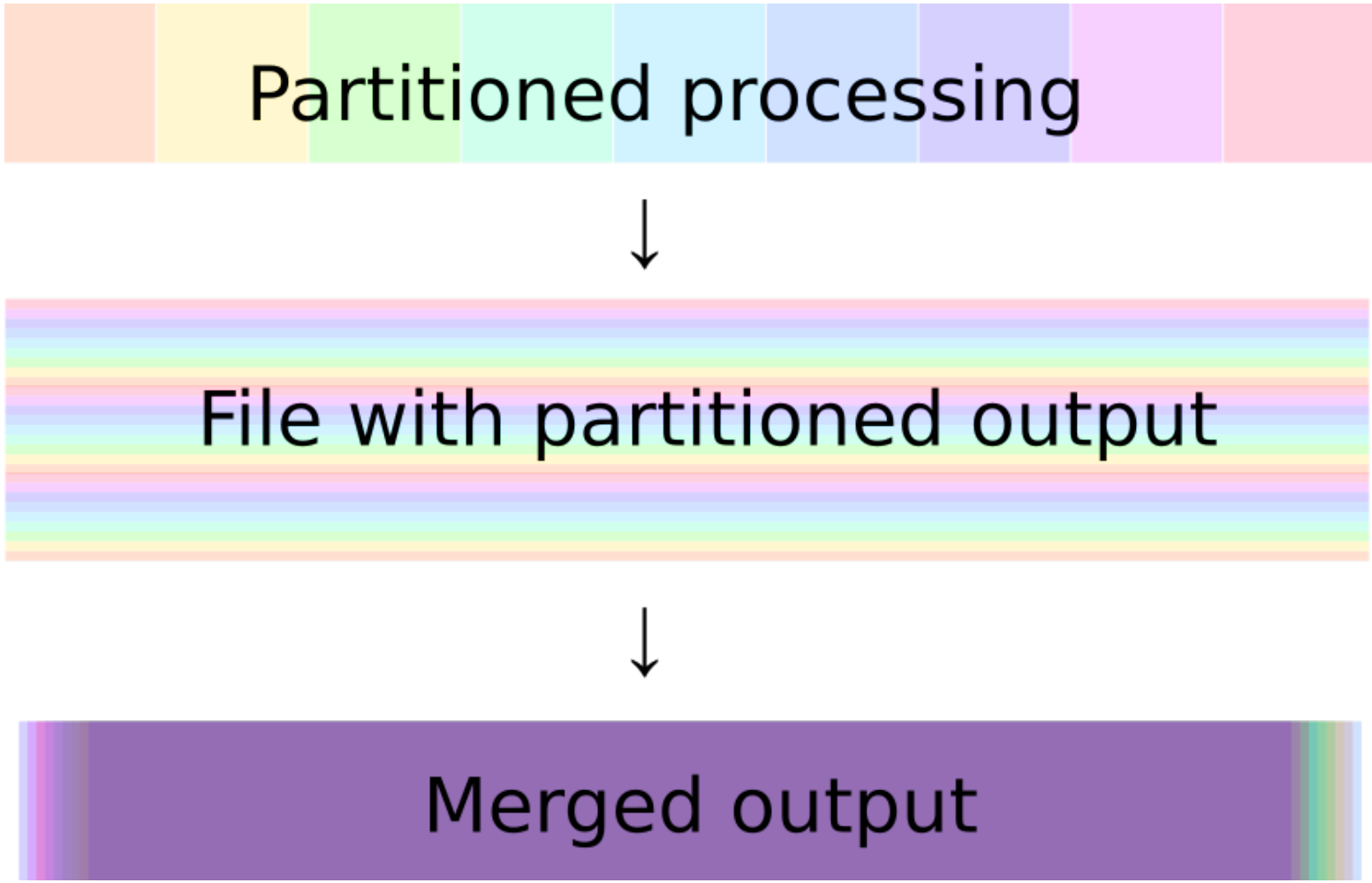
Time for some reanalysis!

- Did the change to the policy change referrals?
- Are we still getting any data?
- Can we observe what's happening?
- Did the change to the DOI URLs happen?
- Are people and machines still following them?

A moment of honest reflection

- “Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate.”
- Learn some things from Spark's architecture but do it myself more simply.
- Plain old Java
- Runs over night
- Handles millions of lines
- Is this big data?
- No,
- It runs on my laptop.

Partition data


Partitioned processing
File with partitioned output
Merged output

Approaches

- Partition the data for RAM
- Expensive parsing as offline first stage
- Plain old CSV files
- Read from and write to gzip directly so the dataset can hang around on disk
- Be ok with running for a few hours once a month

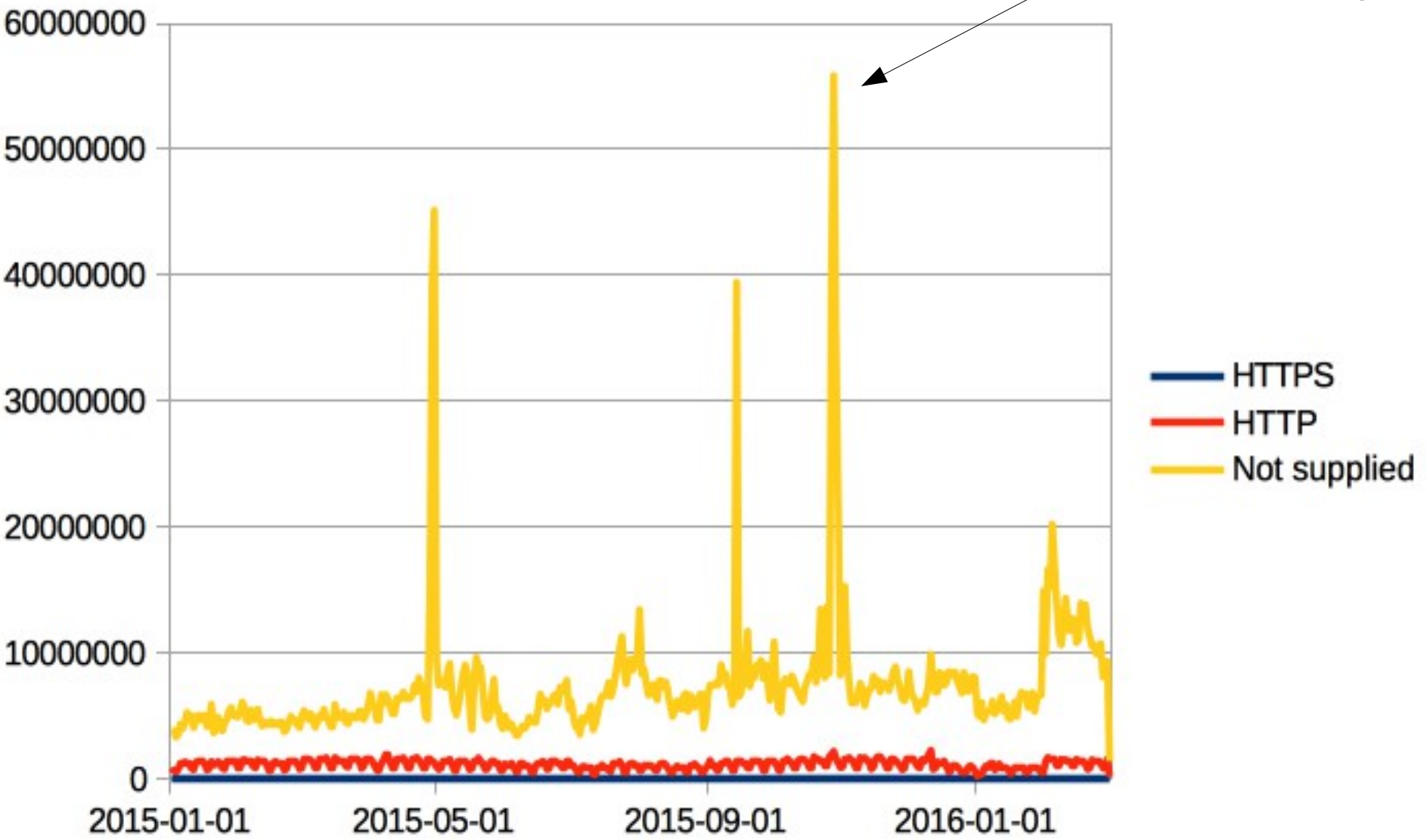
```
ascelibrary.org.er.lib.k-state.edu
2015-05-05,4

stwww.weizmann.ac.il
2015-05-13,1
2015-05-06,1

uat.scholarmate.com
2015-05-14,1

202.204.48.82
2015-05-10,1
2015-05-21,1
2015-05-25,1
2015-05-15,1
2015-05-06,1
2015-05-07,1
2015-05-19,1
2015-05-08,3
2015-05-09,1

energystorage.today
2015-05-21,1
2015-05-23,1

scitation.aip.org.proxy.uba.uva.nl
2015-05-13,1
```

On My Laptop™

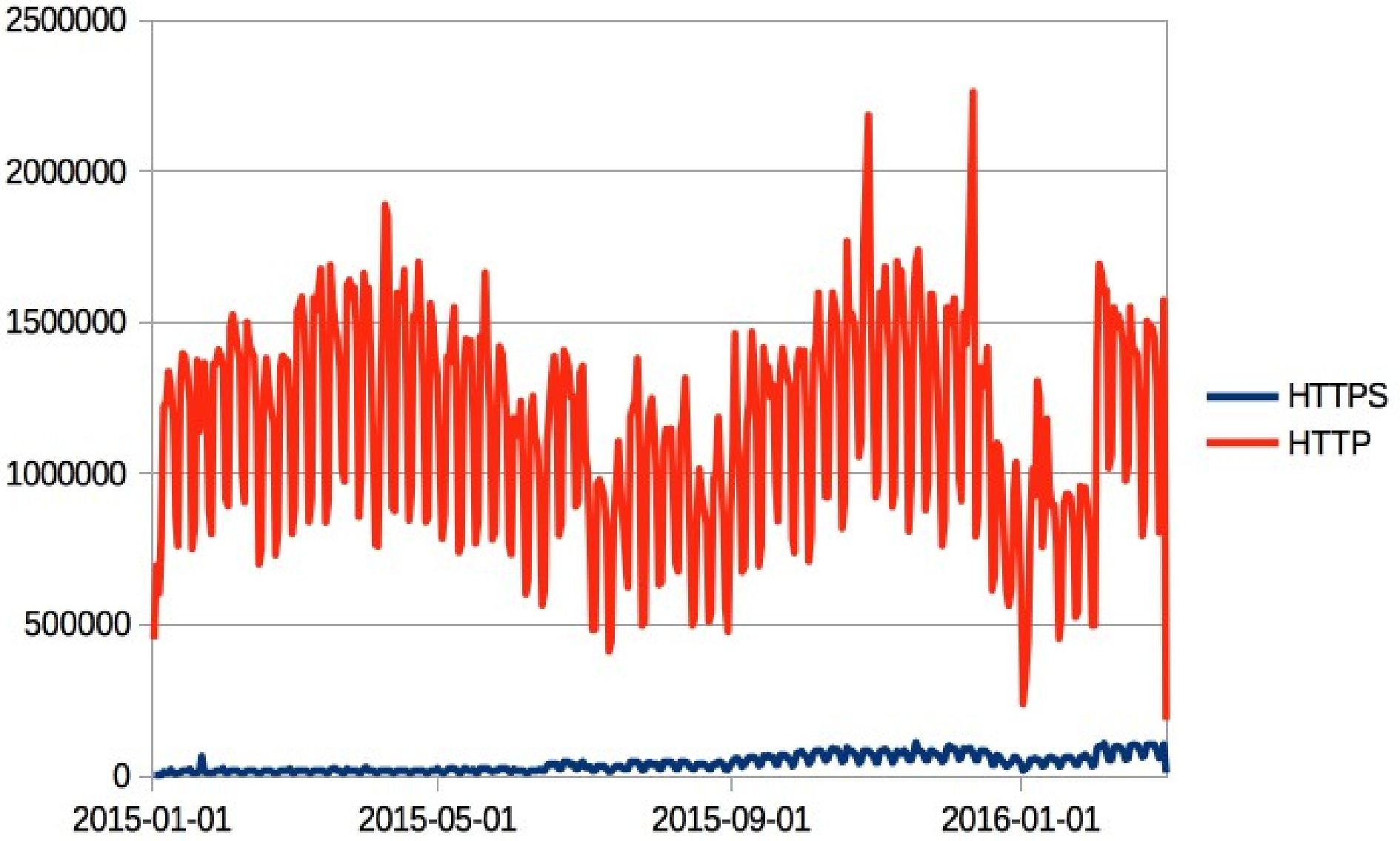
- Parsing / normalizing 15 months
- 15 hrs = 1 million lines every few seconds
- Input 100 GB compressed
- Input 500 GB uncompressed per year
- Output 43 GB compressed
- Input 4,000,000,000 lines
- 0.004 billion if you're British
- FOUR BILLION if you're American
- If you're European you can choose

All DOIs per day

Spiders?
Metadata
lookup?

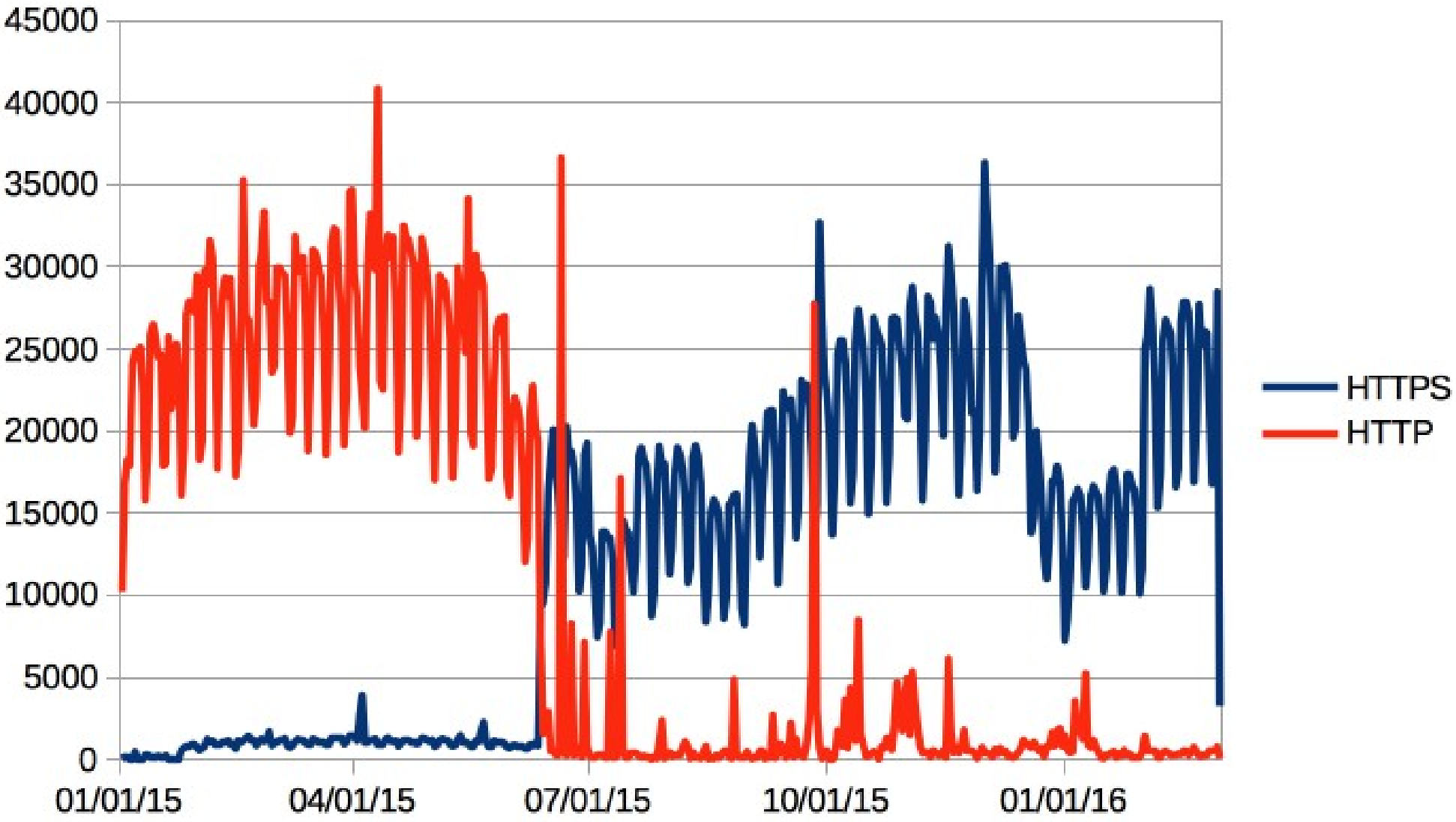


All DOIs HTTP & HTTPS per day



Big reveal...

Wikipedia HTTP & HTTPS per day



Success!

- We can observe the changeover to HTTPS links
- The changeover appeared to work for most of the DOIs
- We have approximately the same volume of data

- Not all large data is Big Data
- Try normal techniques before Big ones.

http://blog.crossref.org

http://eventdata.crossref.org

@joewass