

# On Archiving Linked and Open Data

Axel Polleres

(with a lot of input by **Javier Fernandez & Jürgen Umbrich** ;-)

30.05.2016 - 2nd Workshop on Managing the Evolution and Preservation of the Data Web, MEPDaW@ESWC16

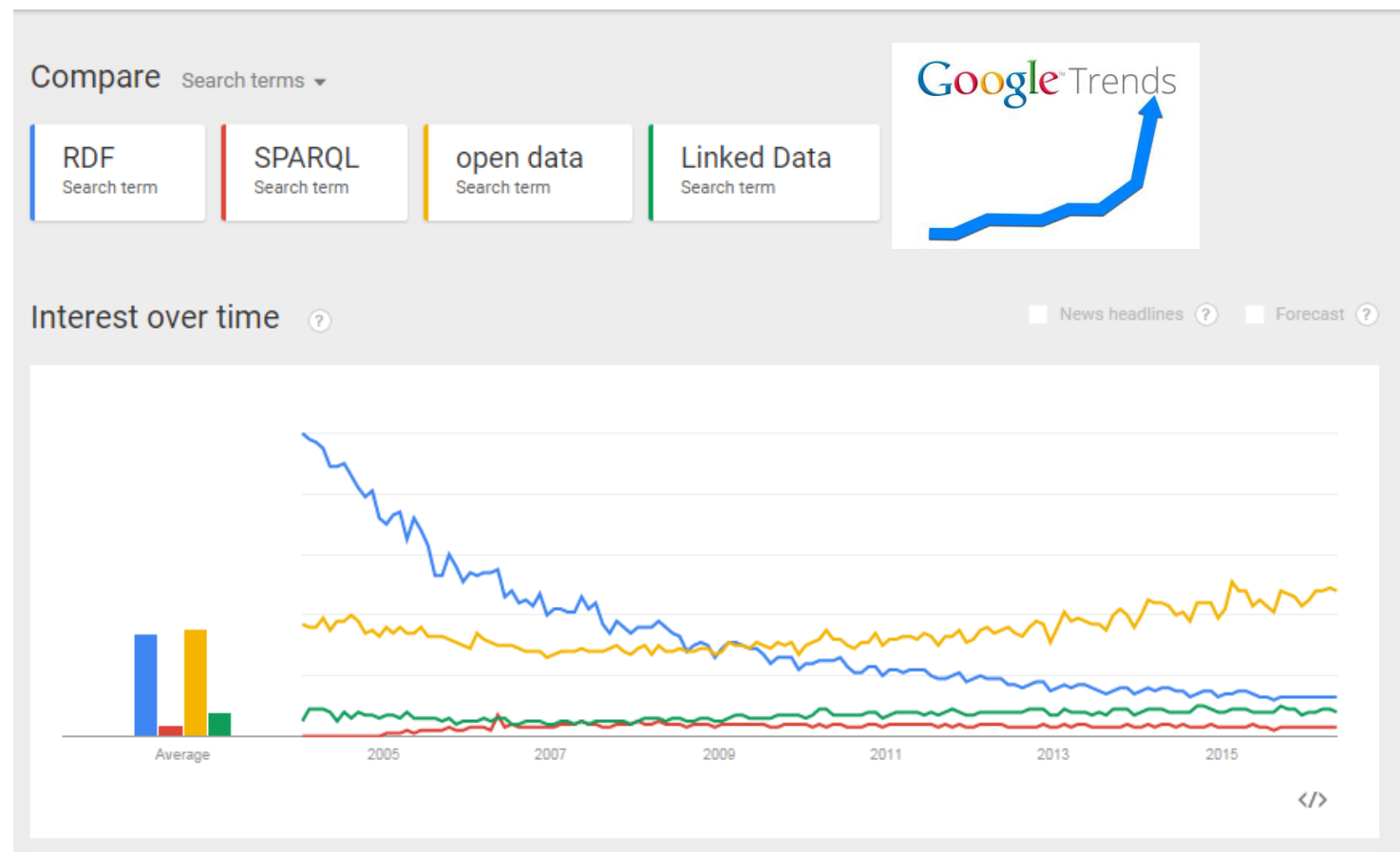
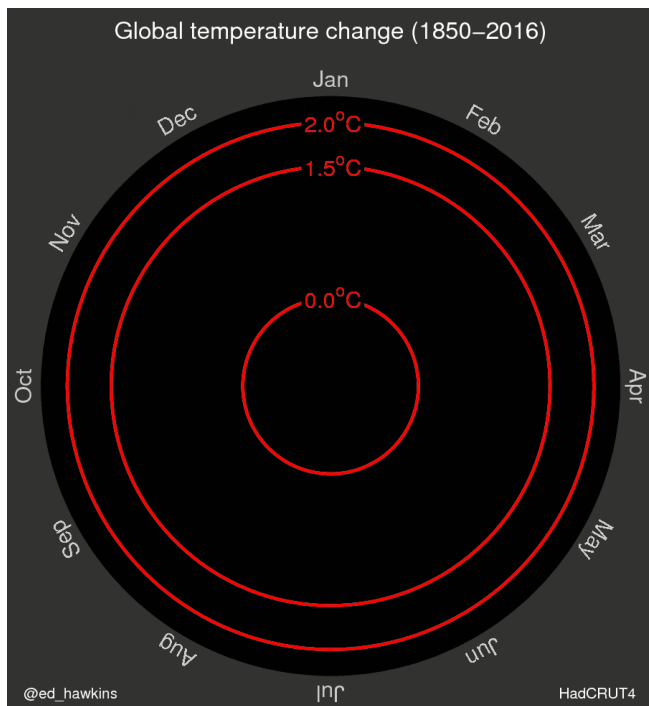
# What we will talk about...

- ***Monitoring Evolution and Archiving ... why is it important? Some examples...***
- ***General Challenges of Archiving the Web of Data***
- ***Some challenges more in-depth***
- ***Discussion... (hopefully 😊 )***

# Why evolution matters

*(Creationists: please ignore this slide...)*

- Monitoring evolution is relevant



# Evolution matters

- Changes tell us “something”
  - Uncertain information
  - Validity of the information

## Donald Trump: Difference between revisions

From Wikipedia, the free encyclopedia

Line 246:

[[File:Donald Trump by Gage Skidmore 3.jpg|thumb|Trump speaking at the 2015 [[Conservative Political Action Conference]] (CPAC) in [[National Harbor, Maryland]]]]

In April 2011, Trump questioned President [[Barack Obama]]'s [[Barack Obama citizenship conspiracy theories|proof of citizenship]].Trump is said to have sent a team of private investigators to [[Hawaii]], Obama's documented birthplace.<ref name=Factcheck2011Apr>{{cite news |url=http://www.factcheck.org/2011/04/donald-youre-fired/ |title=Donald, You're Fired! Trump repeats false claims about Obama's birthplace. |publisher=Factcheck.org |date=April 9, 2011 |accessdate=September 13, 2015}}</ref> and told "[[Today (U.S. TV program)|The Today Show]]" "they cannot believe what they're finding."<ref name=Elliott8Apr>{{cite news |url=http://www.salon.com/2011/04/08/trump\_hawaii\_investigators/ |title=Did Trump really send investigators to Hawaii? |date=April 8, 2011 |accessdate=September 13, 2015 |first=Justin |last=Elliott |work=Salon}}</ref> On April 25, 2011, Trump called for Obama to end the citizenship issue by releasing the long form of his birth certificate.<ref>{{cite news |url=http://ac360.blogs.cnn.com/2011/04/25/trump-claims-obama-birth-certificate-missing/ |title=Trump claims Obama birth certificate 'missing' |date=April 25, 2011

Line 246:

[[File:Donald Trump by Gage Skidmore 3.jpg|thumb|Trump speaking at the 2015 [[Conservative Political Action Conference]] (CPAC) in [[National Harbor, Maryland]]]]

In April 2011, Trump questioned President [[Barack Obama]]'s [[Barack Obama citizenship conspiracy theories|proof of citizenship]].

# Evolution matters

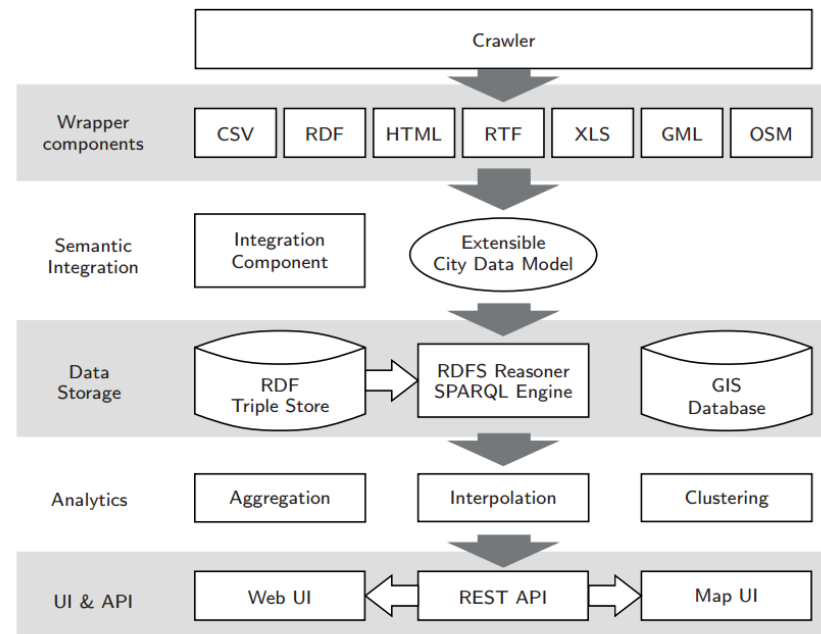
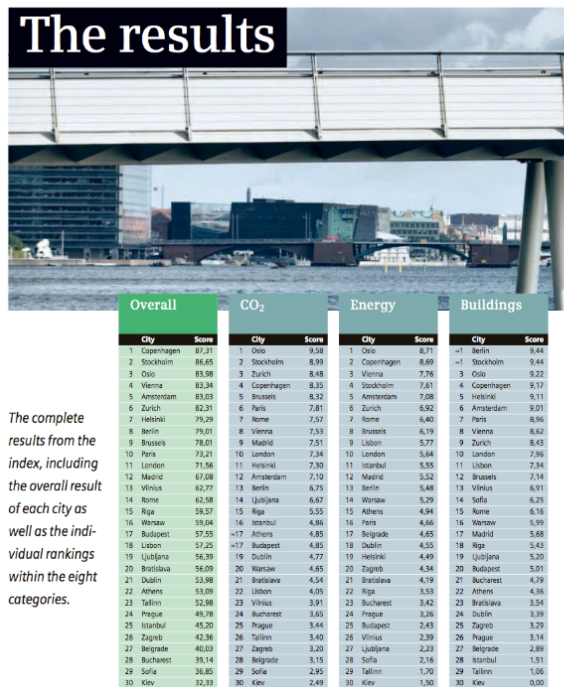
- Evolution may reveal actions vs. consequences
  - E.g. CityDataPipeline, <http://citydata.wu.ac.at/>
    - Collecting, Integrating and Predicting Open City Data

## Vienna Environment

### Accumulated ozone concentration

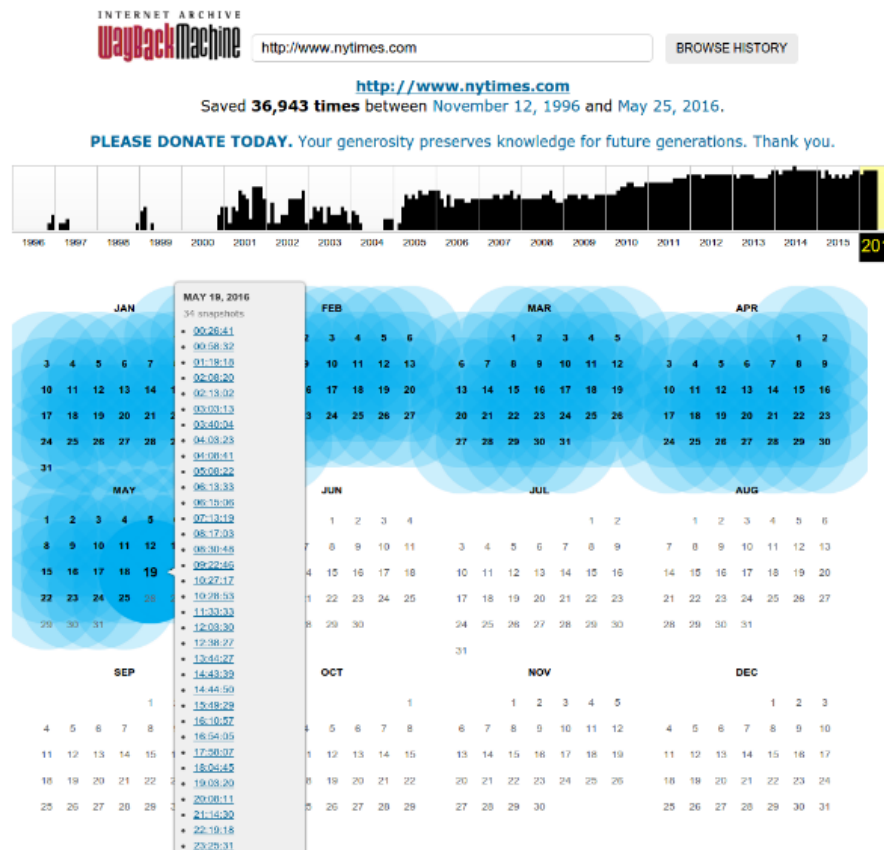
- 1992:** 6112.40 microgram per m3 (from <http://epp.eurostat.ec.europa.eu/>)
- 1994:** 5998.80 microgram per m3 (from <http://epp.eurostat.ec.europa.eu/>)
- 1995:** 5513.80 microgram per m3 (from <http://epp.eurostat.ec.europa.eu/>)
- 1997:** 5181.80 microgram per m3 (from <http://epp.eurostat.ec.europa.eu/>)
- 1998:** 5785.40 microgram per m3 (from <http://epp.eurostat.ec.europa.eu/>)
- 1999:** 5141.60 microgram per m3 (from <http://epp.eurostat.ec.europa.eu/>)
- 2000:** 6711.00 microgram per m3 (from <http://epp.eurostat.ec.europa.eu/>)
- 2001:** 4940.40 microgram per m3 (from <http://epp.eurostat.ec.europa.eu/>)
- 2002:** 5943.20 microgram per m3 (from <http://epp.eurostat.ec.europa.eu/>)
- 2003:** 7939.40 microgram per m3 (from <http://epp.eurostat.ec.europa.eu/>)
- 2004:** 4755.60 microgram per m3 (from <http://epp.eurostat.ec.europa.eu/>)
- 2005:** 5585.00 microgram per m3 (from <http://epp.eurostat.ec.europa.eu/>)

European Green City Index | The results



# Preservation matters

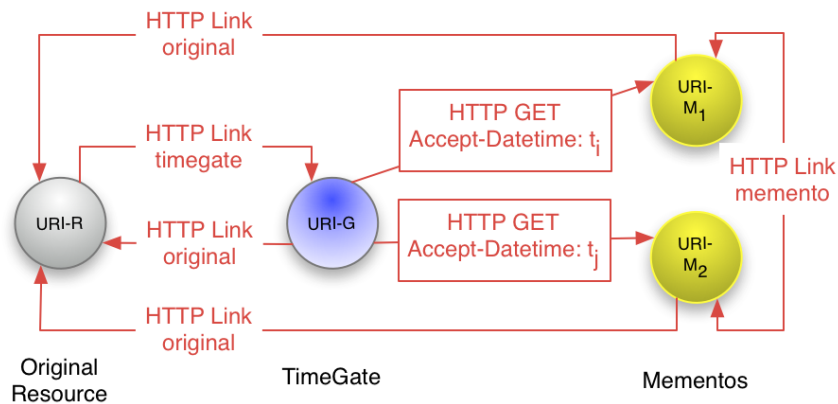
- Web archives: Common Crawl, Internet Memory, Internet Archive, ...



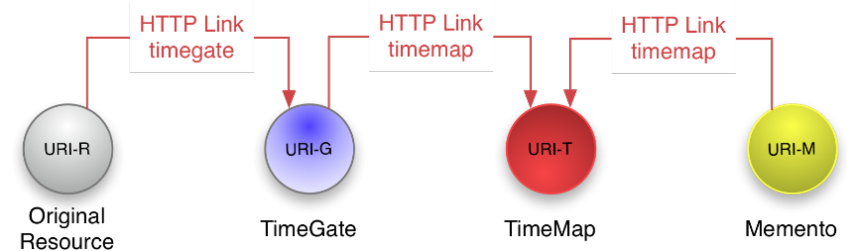
# Time-based access matters

- The Memento protocol RFC 7089

Follow your nose  
(HTTP content negotiation with datetime)



Batch discovery  
(list of URIs of Mementos of the Original Resource)



But...

# Challenges

- Poor **granularity** (“some” snapshots)
- **Aggregated** data, only, rather than raw data access
  - (e.g. in Google trends)
- Few work on archiving the **Web of Data**, or on integrating archives
- What is the right **query language**?
  - basic retrieval features (get version at timestamp **t**)
  - when did a certain information disappear?
  - when was it changed?
  - structured queries?
- Scalability problems

Donald Trump: Revision history

From year (and earlier): 2016 From months (and earlier): all Tag filter: Show

For any version listed below, click on its data to view it. For more help, see Help:Page history and Help:Edit summary.

External tools: Revision history statistics · Revision history search · Edits by user · Number of watchers · Page view statistics

(cur = difference from current version, (prev) = difference from preceding version, m = minor edit, → = section edit, ← = automatic edit summary (newest | oldest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

Compare selected revisions

- (cur | prev) 01:45, 25 May 2016 Anythingyouwant (talk | contribs) ... (282,961 bytes) (+27) ... (←Early life: clarify who was Frederick and who was Fred)
- (cur | prev) 00:32, 25 May 2016 Anythingyouwant (talk | contribs) ... (282,934 bytes) (-27) ... (per talk page, unclustering huge clusters of hidden footnotes)
- (cur | prev) 00:27, 25 May 2016 BarneProof (talk | contribs) ... (282,961 bytes) (+57) ... (←Religious views: Younger than what?)
- (cur | prev) 00:06, 25 May 2016 AnonWBOT (talk | contribs) ... (282,904 bytes) (+548) ... (Rescuing orphaned refs ("msnbc" from rev 721935275; "CBC, August29, 2015" from rev 721935275))
- (cur | prev) 23:30, 24 May 2016 Anythingyouwant (talk | contribs) ... (282,355 bytes) (-1) ... (←Controversial immigration policies: spelling, merge two paragraphs)
- (cur | prev) 23:35, 24 May 2016 Anythingyouwant (talk | contribs) ... (282,354 bytes) (-661) ... (←Controversial immigration policies: swap hidden cluster of footnotes for two footnotes from main article on political positions, paraphrase accordingly)
- (cur | prev) 23:19, 24 May 2016 Anythingyouwant (talk | contribs) ... (283,015 bytes) (-2) ... (move reference out of huge hidden cluster of footnotes, to appropriate spot in 2016 campaign section)
- (cur | prev) 23:13, 24 May 2016 Anythingyouwant (talk | contribs) ... (283,017 bytes) (-1,626) ... (←Controversial immigration policies: trim some footnotes from big hidden cluster of footnote)
- (cur | prev) 23:07, 24 May 2016 Anythingyouwant (talk | contribs) ... (284,043 bytes) (-7) ... (←Controversial immigration policies: move reference out of cluster)

Wikipedia habits



INTERNET ARCHIVE  
waybackmachine

https://en.wikipedia.org/wiki/Donald\_Trump BROWS

https://en.wikipedia.org/wiki/Donald\_Trump  
Saved 790 times between July 24, 2004 and May 23, 2016.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations.

1966 1967 1968 1969 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012

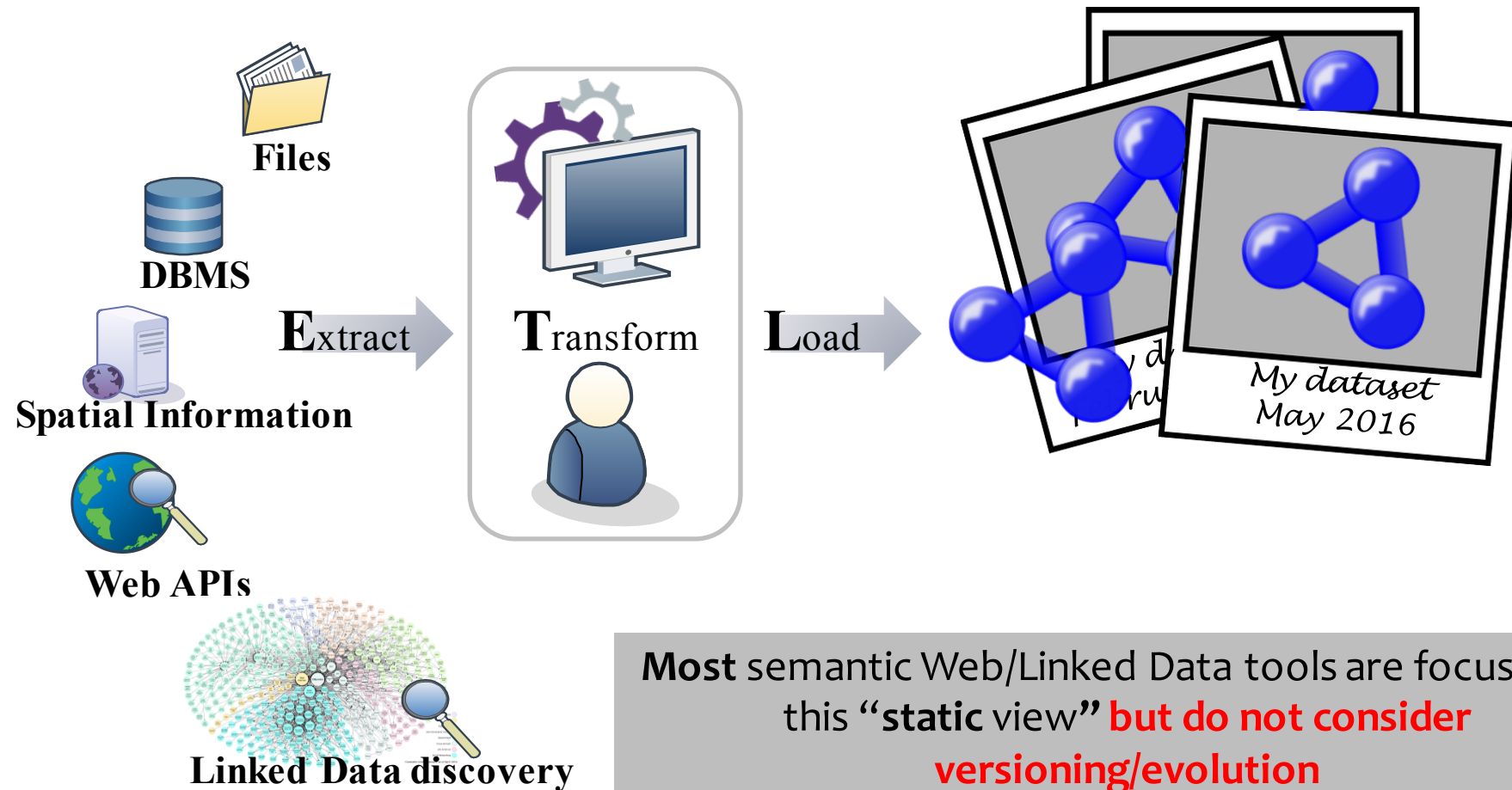
JAN FEB MAR APRIL

MAY JUN JUL

Is it easier/better for RDF/Linked Data?



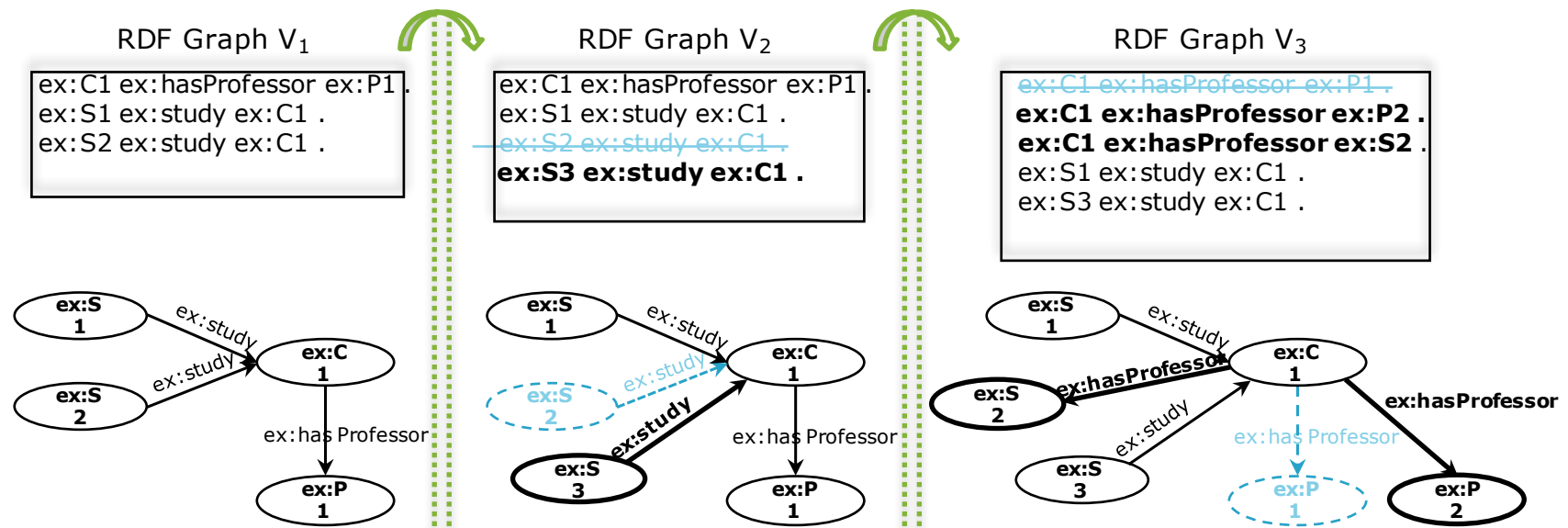
# Linked Data Archives: The missing link in the RDF evolution



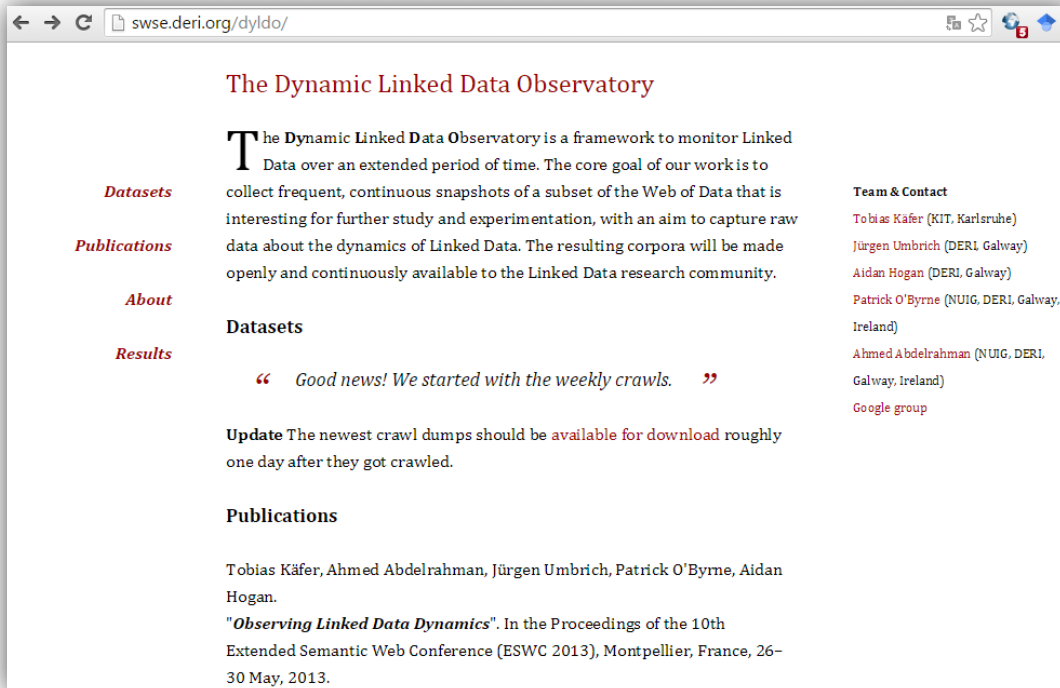
Most semantic Web/Linked Data tools are focused on this “static view” **but do not consider versioning/evolution**

Sindice, SWSE, Swoogle, LOD Cache, LOD-Laundromat... so far, no versions!

# RDF Archiving. Example



# One of the first real LOD use cases: The Dynamic Linked Data Observatory (*evolving Linked Data since 2012*)



The Dynamic Linked Data Observatory

The Dynamic Linked Data Observatory is a framework to monitor Linked Data over an extended period of time. The core goal of our work is to collect frequent, continuous snapshots of a subset of the Web of Data that is interesting for further study and experimentation, with an aim to capture raw data about the dynamics of Linked Data. The resulting corpora will be made openly and continuously available to the Linked Data research community.

**Datasets**

**Publications**

**About**

**Results**

**Team & Contact**  
Tobias Käfer (KIT, Karlsruhe)  
Jürgen Umbrich (DERI, Galway)  
Aidan Hogan (DERI, Galway)  
Patrick O'Byrne (NUIG, DERI, Galway, Ireland)  
Ahmed Abdelrahman (NUIG, DERI, Galway, Ireland)  
Google group

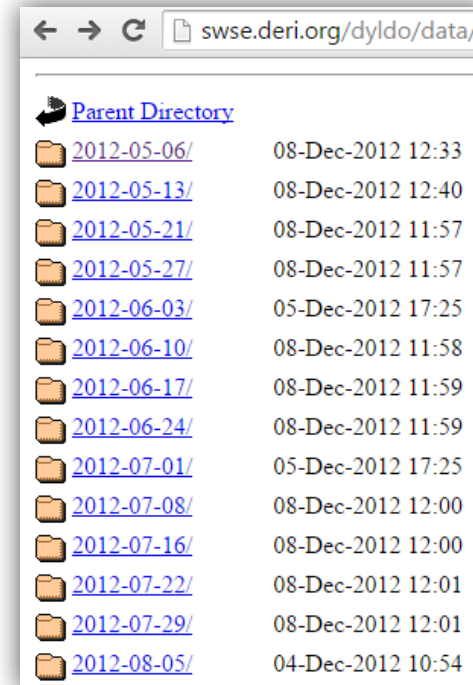
**Datasets**

“ Good news! We started with the weekly crawls. ”

**Update** The newest crawl dumps should be available for download roughly one day after they got crawled.

**Publications**

Tobias Käfer, Ahmed Abdelrahman, Jürgen Umbrich, Patrick O'Byrne, Aidan Hogan.  
"Observing Linked Data Dynamics". In the Proceedings of the 10th Extended Semantic Web Conference (ESWC 2013), Montpellier, France, 26-30 May, 2013.



Parent Directory

2012-05-06/	08-Dec-2012 12:33
2012-05-13/	08-Dec-2012 12:40
2012-05-21/	08-Dec-2012 11:57
2012-05-27/	08-Dec-2012 11:57
2012-06-03/	05-Dec-2012 17:25
2012-06-10/	08-Dec-2012 11:58
2012-06-17/	08-Dec-2012 11:59
2012-06-24/	08-Dec-2012 11:59
2012-07-01/	05-Dec-2012 17:25
2012-07-08/	08-Dec-2012 12:00
2012-07-16/	08-Dec-2012 12:00
2012-07-22/	08-Dec-2012 12:01
2012-07-29/	08-Dec-2012 12:01
2012-08-05/	04-Dec-2012 10:54

Weekly dumps of  
crawl snapshots...

Granularity?  
Queries?  
Completeness?  
Crawl failures?



# Research challenges on evolving structured interlinked data

- How can we **represent archives** of continuously evolving linked datasets? (efficiency vs. compact representation)
- How can we **minimize the redundant information** of archives? (e.g. duplicates in snapshots)
- How can we improve **completeness** of archiving?
- How can emerging retrieval demands in archiving be satisfied?
  - e.g. time-traversing and traceability? Avoiding bottlenecks?
- How can certain **time-specific queries** over archives be answered?
  - Can we re-use existing technologies (e.g. SPARQL or temporal extensions)?
  - *What is the right **query language** for such queries?*
  - e.g. knowing if a dataset has changed, and how, in a certain time period?

# General archiving challenges

- ***The synchronisation problem***
  - how can we monitor changes?
- ***The appraisal problem***
  - how can we assess the quality of a dataset? (and does archiving help with that?)
- ***The archiving & query problem***
  - how can we efficiently archive and perform time-based retrieval queries of a dataset?

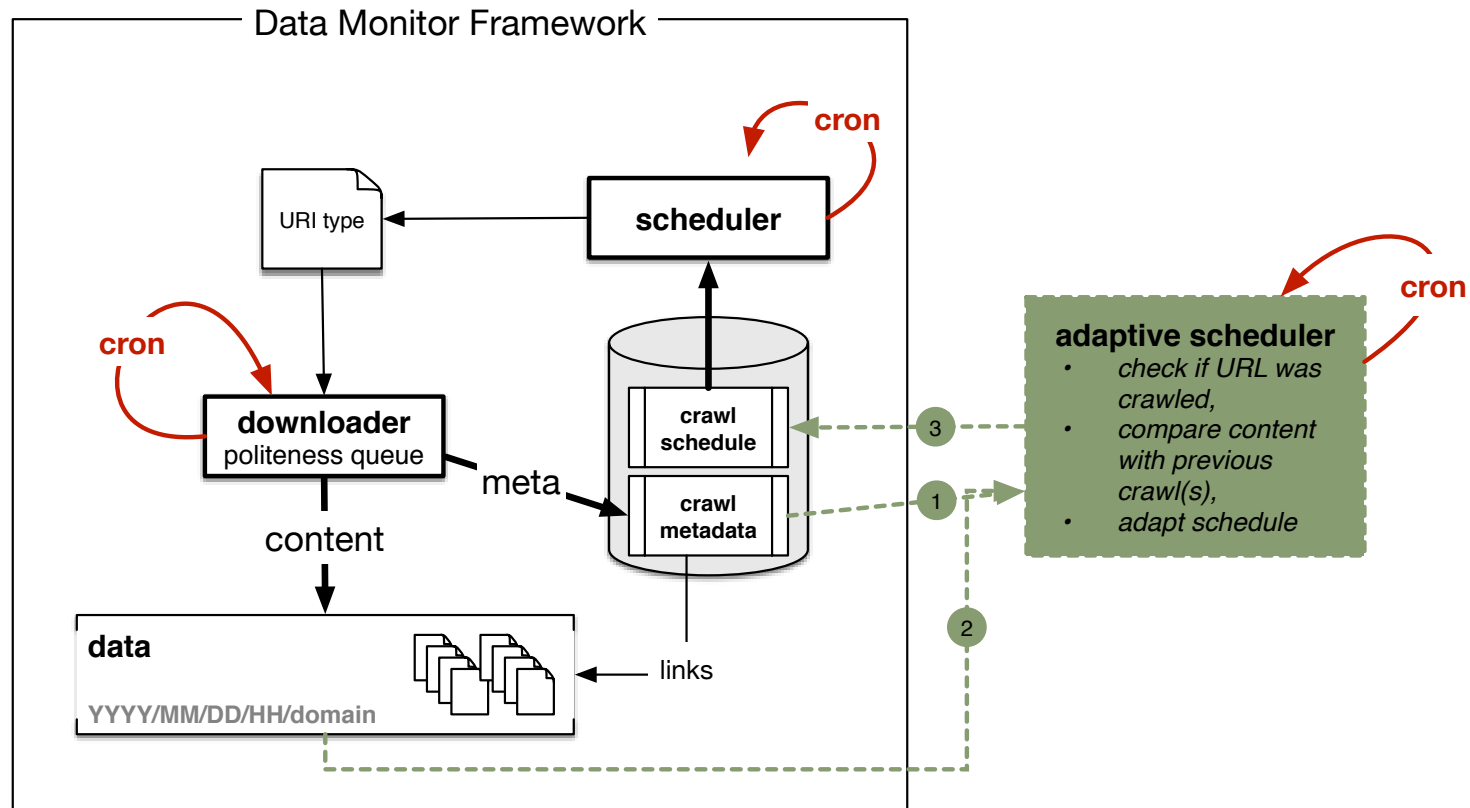
# The synchronization problem

how can we monitor changes?

# Pull changes (crawl) vs. Push changes (notify)

- Observations:
  - Some services that publish or are mapped to RDF change **regularly**, but we don't know the frequency upfront!
  - Some services mapped to RDF **announce/archive their changes already**, so they already keep an archive...

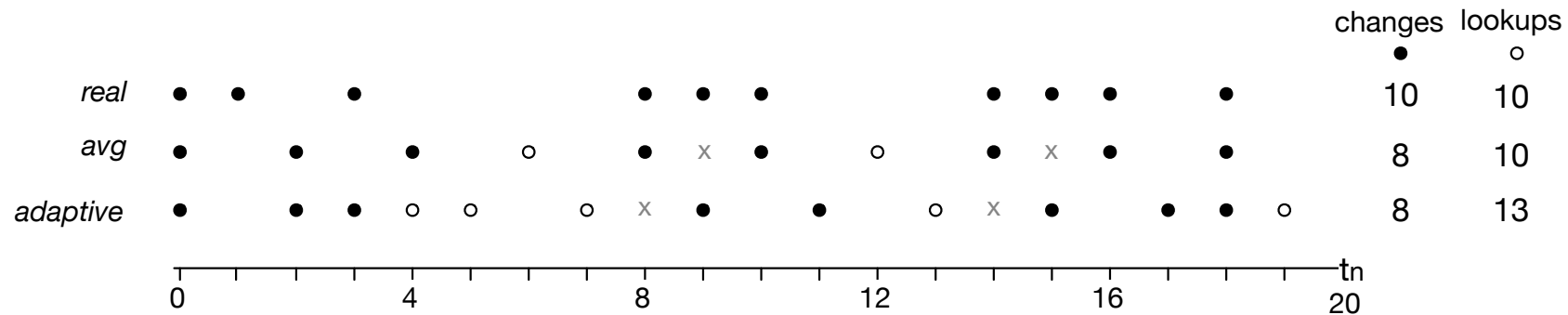
# 1- An adaptive archiver (recall: DIACHRON WS 2015)





# Experiment: Rescheduling

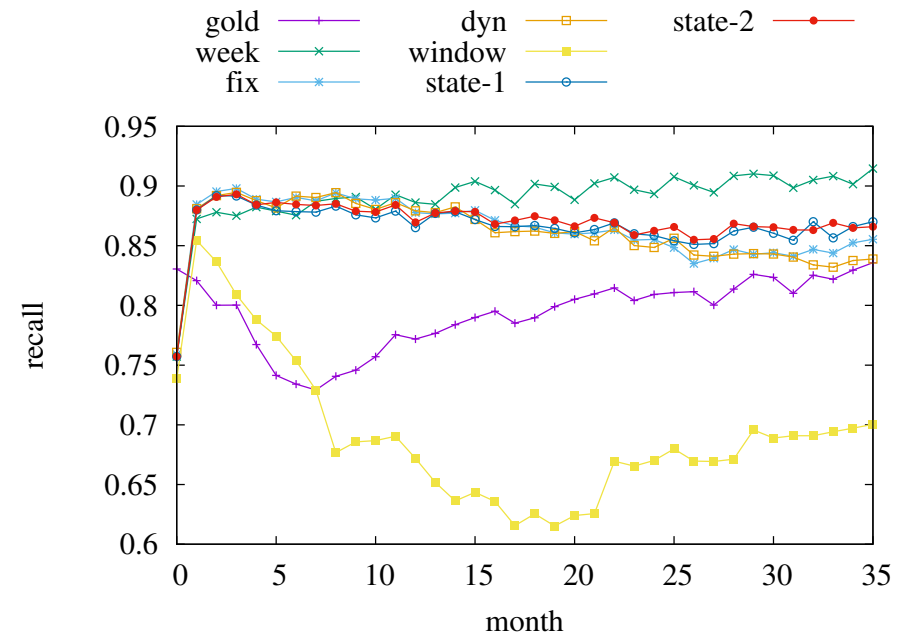
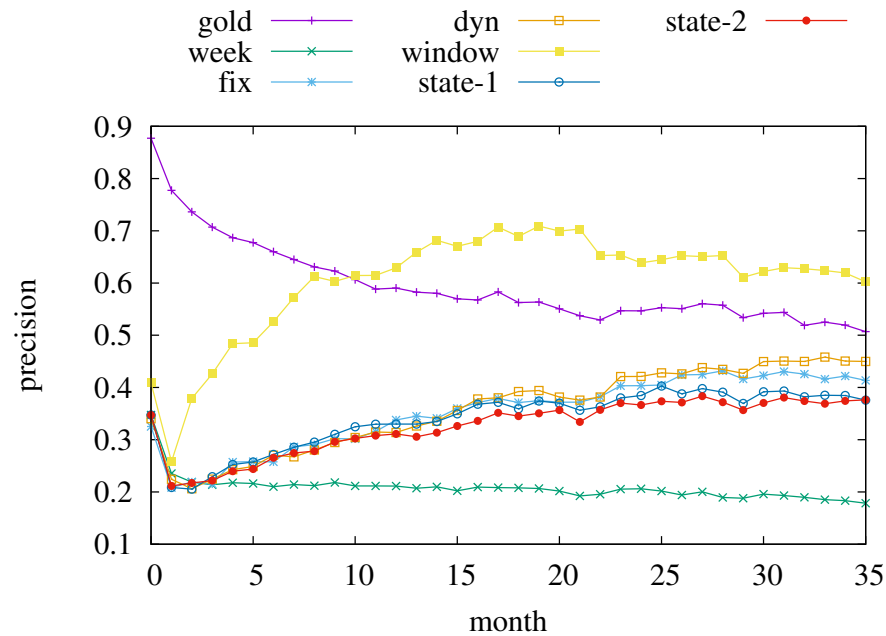
- Evaluating strategies to compute next crawl time for URLs to accurately capture content change



# Test setup

- Revision history of 2660 wikipedia-articles
  - Wiki-changes do not follow a typical Poisson distribution
- Several heuristics; e.g.:  
e.g., increase the crawl frequency,..."
  - "if you observe several changes in a row"
  - "if the probability is high that the content changes after we observed a change in the last snapshot" → Markov models
  - "If the document changed more often than 50% in the last 10 days .... "

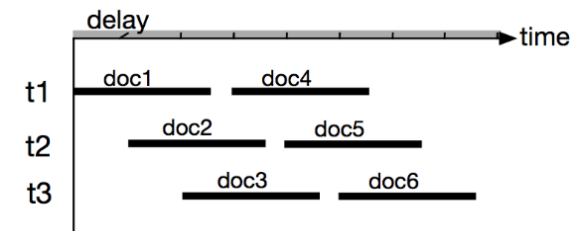
# Results



We observed the typical trade-off between recall and precision...  
Strategies based on Markov models seems to provide best and most stable trade-off

# Experiment: Improve completeness - Crawl time estimation

- Aim: Estimate the overall crawl time and needed number of threads for a set of URLs from different domains
- Heuristic
  1. Estimate the crawl time per domain
    - Average download time, domain delay etc...
  2. First-fit bin-packing algorithms to determine overall crawl time



# Results

**Table 3.** Results of crawl time estimation

Actual crawl time	Overestimate (crawls)	Underestimate (crawls)	total crawls
$t < 30min$	78% (202)	-16% (154)	356
$30min < t < 60min$	134% (25)	-18% (46)	71
$60min < t < 120min$	638% (19)	-11% (51)	70
$120min < t < 180min$	1.3% (15)	-1% (215)	230
$180min < t < 240min$	- (0)	-36% (4)	4
$t > 240min$	- (0)	-44% (28)	28

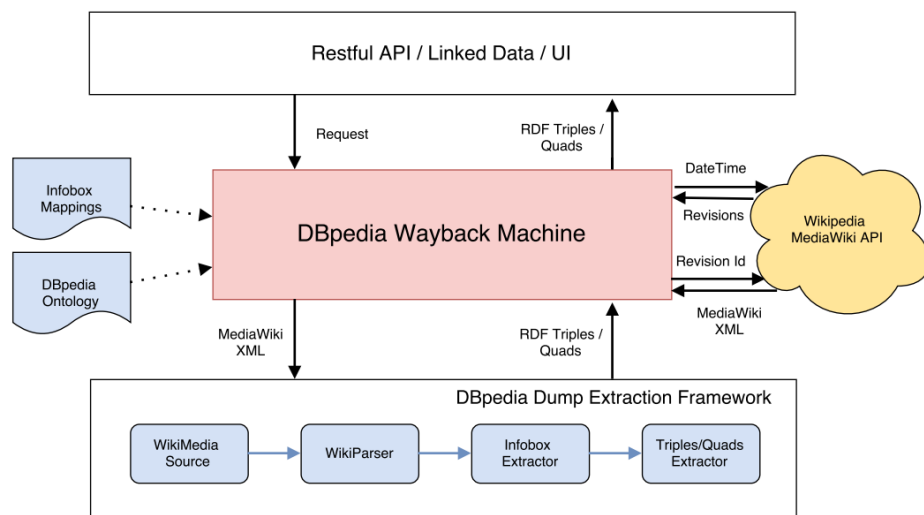
- Large over estimation (is acceptable)
- Small underestimation (e.g. 10mins for 60mins crawl)

# Pull changes (crawl) vs. Push changes (notify)

- Observations:
  - Some services that publish or are mapped to RDF change regularly, but we don't know the frequency upfront!
  - **Some services mapped to RDF announce/archive their changes already, so they already keep an archive...**

## 2- “Recreate” the versions from sources (SEMANTiCS 2015 demo...)

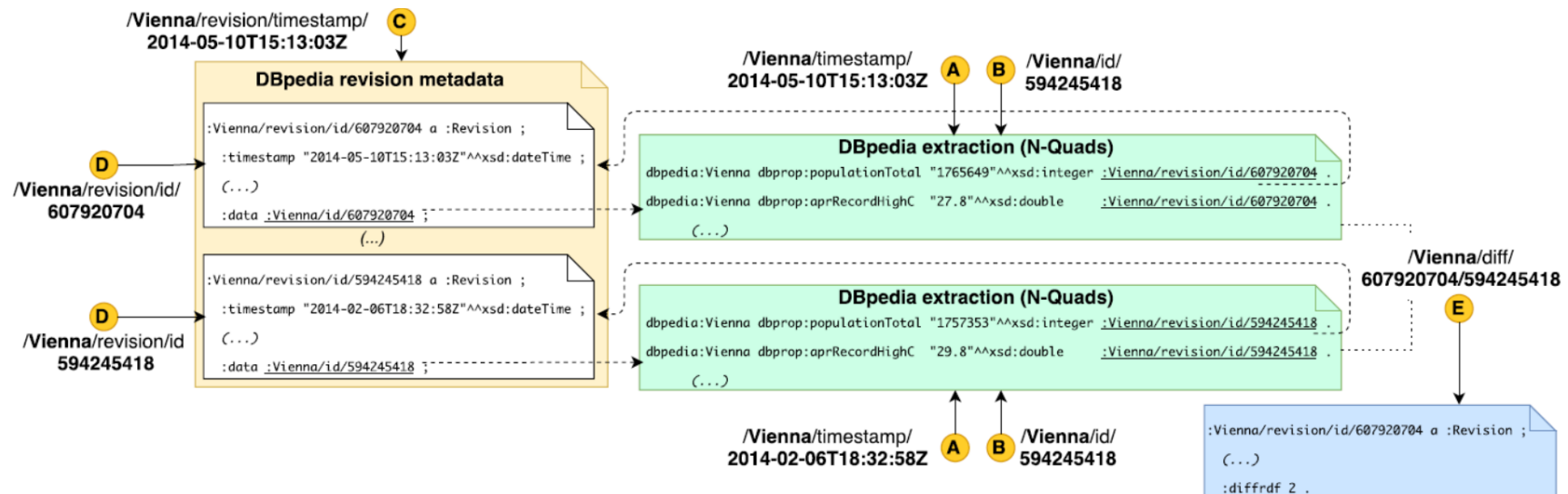
- If raw historical data on changes is available...
- Aim: Fine grained access to previous versions, re-applying X2RDF transformations on the original source
- Example: **DBpedia Wayback machine**
  - Re-apply mappings on the Wikipedia revision history



<http://data.wu.ac.at/wayback/>

## 2- "Recreate" the versions

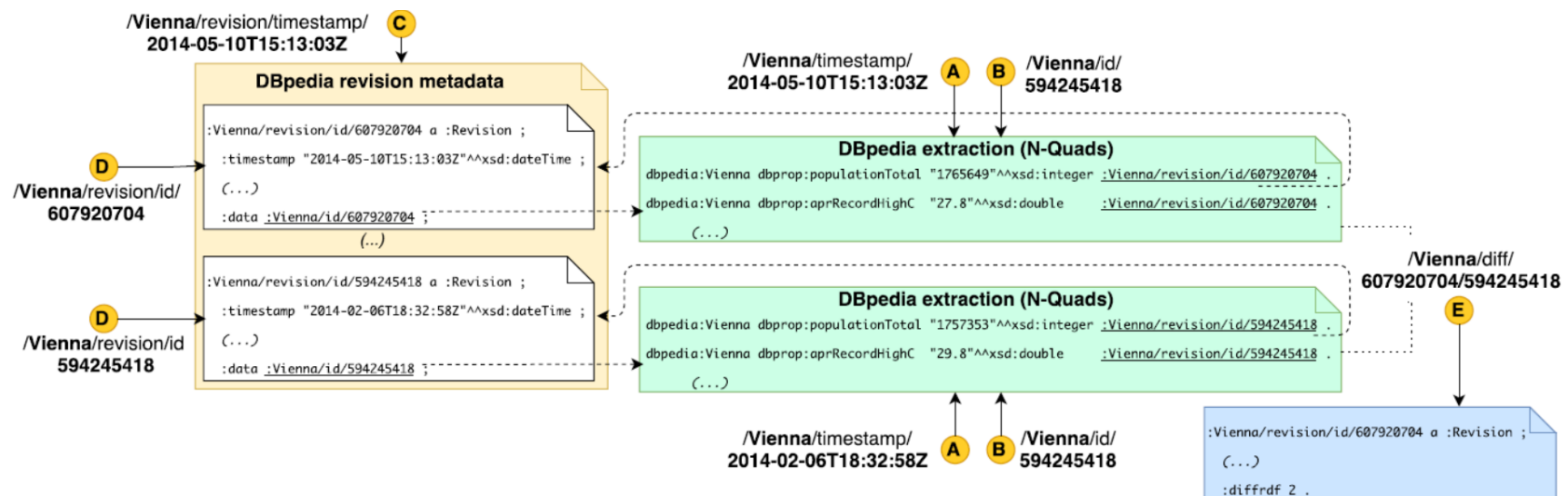
- How can one represent revisions while respecting DBpedia?
  - a) quads → <dbpediaSubject> <pred> <obj> <Revision> .
  - b) proprietary triples → <ownSubject/Revision> <pred> <obj> .
- Operations?
  - Get revisions meta-data for one resource (by revisionID or timestamp)
  - Get "materialised" versions of a resource (by revisionID or timestamp)
  - Get difference between two revisions





## 2- "Recreate" the versions

- More complex operations/queries? **Open challenge**
  - a) **On-demand? Query rewriting, similar to RDB2RDF**
  - b) **Batch: Fetch the desired information, then store and query it**



# We are (obviously) not the only ones looking into this...

The screenshot shows a SlideShare interface. At the top, there is a navigation bar with the LinkedIn logo, the text 'SlideShare', a search bar containing 'Suche', and several menu items: 'Startseite', 'Technologie', 'Bildung', 'Mehr Themen', and 'My Clipboards'. Below the navigation bar, a blue banner indicates '1 person clipped this slide'. The main content area features a slide with the following elements:

- Title:** Access to DBpedia Versions using Memento and Triple Pattern Fragments
- Image:** A circular logo for 'memento' featuring a clock face.
- Text:** Herbert Van de Sompel  
@hvdsomp  
Los Alamos National Laboratory
- Image:** A logo for '#LD Linked Data Fragments'.
- Text:** Miel Vander Sande  
@Miel\_vds  
Ghent University
- Text:** Acknowledgments: Lyudmila Balakireva, Harihar Shankar, Ruben Verborgh

However:  
Only one HDT per  
“irregular” dbpedia  
dump

# The appraisal problem

How can we assess the quality of a dataset?

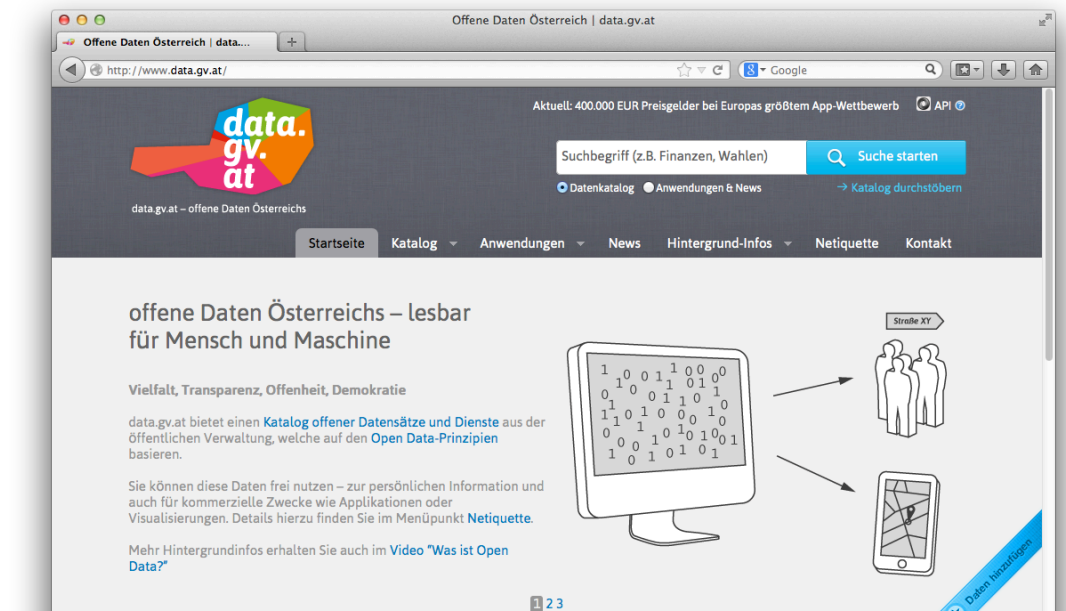
# Data Quality issues:

- Missing
- Outdated data
- Wrong data
- Ambiguous Data
- Wrong meta-data
- Data source offline/not reachable
  
- → **Archiving** & looking at the **history** of datasets helps!

# Open Data Portals

CKAN ... <http://ckan.org/>

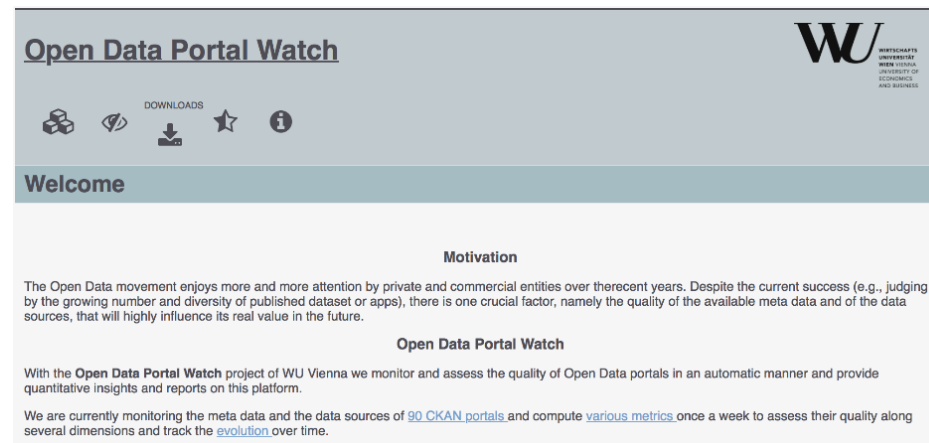
- almost „de facto“ standard for Open Data Portals
- facilitates search, metadata (publisher, format, publication date, license, etc.) for datasets
- <http://datahub.io/>
- <http://data.gv.at/>
- machine-processable? ...  
... **partially**



# OPEN DATA PORTAL WATCH ... a first step.

<http://data.wu.ac.at/portalwatch/>

- Periodically monitoring a list of Open Data Portals
  - 90 CKAN powered Open Data Portals
- Quality assessment
- Evolution tracking
  - Meta data
  - Data



The screenshot shows the homepage of the 'Open Data Portal Watch' project. At the top, there is a header with the project name 'Open Data Portal Watch' on the left and the WU logo on the right. Below the header, there is a navigation bar with icons for a recycling symbol, a magnifying glass, a download arrow, a star, and an information icon, with the word 'DOWNLOADS' centered above them. The main content area starts with a 'Welcome' section, followed by a 'Motivation' section. The 'Motivation' section contains text explaining the project's goal: to monitor and assess the quality of Open Data portals. Below this is the 'Open Data Portal Watch' section, which provides more details about the project's methodology and current status, mentioning that they monitor 90 CKAN portals and compute various metrics weekly.

# Open Data Portal list

## Open Data Portal Watch



### Brief overview of 89 Open Data CKAN portals

Sort by [Domain](#) [Country](#) [Datasets](#) [Resources](#) Filter:  [Tile view](#) [Table View](#)

<p>annuario.comune.fi.it Italy</p> <p>358 DATASETS 1363 RESOURCES</p>	<p>catalogue.datalocale.fr France</p> <p>303 DATASETS 751 RESOURCES</p>	<p>dados.gov.br Brazil</p> <p>501 DATASETS 4344 RESOURCES</p>	<p>data.buenosaires.gob.ar Argentina</p> <p>123 DATASETS 626 RESOURCES</p>
<p>data.edostate.gov.ng Nigeria</p> <p>164 DATASETS 207 RESOURCES</p>	<p>data.glasgow.gov.uk United Kingdom (common practice)</p> <p>384 DATASETS 1943 RESOURCES</p>	<p>datagm.org.uk United Kingdom (common practice)</p> <p>360 DATASETS 506 RESOURCES</p>	<p>data.gov.sk Slovakia</p> <p>216 DATASETS 556 RESOURCES</p>
<p>ckan.data.graz.gv.at Austria</p> <p>151 DATASETS 341 RESOURCES</p>	<p>data.kk.dk Denmark</p> <p>102 DATASETS 346 RESOURCES</p>	<p>data.lexingtonky.gov government</p> <p>93 DATASETS 186 RESOURCES</p>	<p>data.nsw.gov.au Australia</p> <p>311 DATASETS 458 RESOURCES</p>
<p>data.ohouston.org non-commercial</p> <p>227 DATASETS 361 RESOURCES</p>	<p>data.ottawa.ca Canada</p> <p>119 DATASETS 493 RESOURCES</p>	<p>data.cityofsantacruz.com commercial</p> <p>52 DATASETS 72 RESOURCES</p>	<p>dados.recife.pe.gov.br Brazil</p> <p>43 DATASETS 318 RESOURCES</p>

# QUALITY DIMENSIONS

---

<b>DIMENSION</b>	<b>DESCRIPTION</b>
Retrievability	The extent to which meta data and resources can be retrieved.
Usage	The extent to which available meta data keys are used to describe a dataset.
Completeness	The extent to which the used meta data keys are non empty.
Accuracy	The extent to which certain meta data values accurately describe the resources.
Openness	The extent to which licenses and file formats conform to the open definition.
Contactability	The extent to which the data publisher provide contact information.

---

Objective measures which can be automatically computed in a scalable way



# Portal Overview

## Open Data Portal Watch

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Portal: GovData | Datenportal für Deutschland - GovData

OVERVIEW DETAILS EVOLUTION

Available Snapshots

Snapshot: Sun Feb 22 2015 23:52:47 GMT+0100 (CET)

### QUALITY

Metric	Value
Qr(ds)	100%
Qi	~80%
Qc	~80%
Qu	~80%

### SIZE

Metric	Value
DATASETS	13195
RESOURCES	37256

### OPENNESS

Metric	Avg. Value
LICENSE	0.17
FORMAT	0.94

### RETRIEVABILITY

Metric	Avg. Value
DATASETS	1.00
RESOURCES	0.79

### CONTACTABILITY

Metric	Avg. Value
EMAIL	0.92
URL	0.00

# ODP Evolution

## Open Data Portal Watch



Portal: GovData | Datenportal für Deutschland - GovData

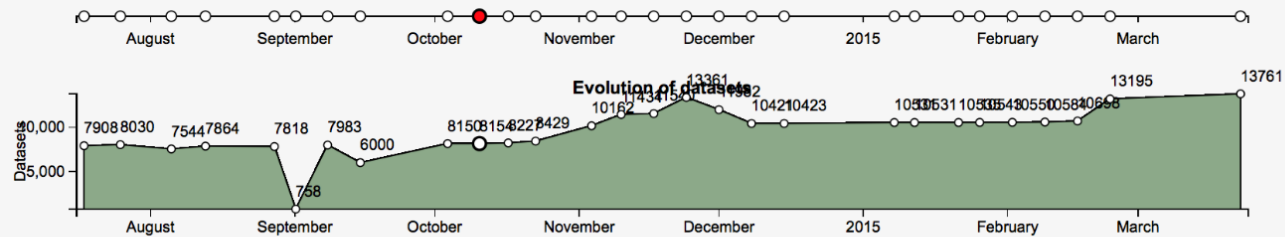
OVERVIEW

DETAILS

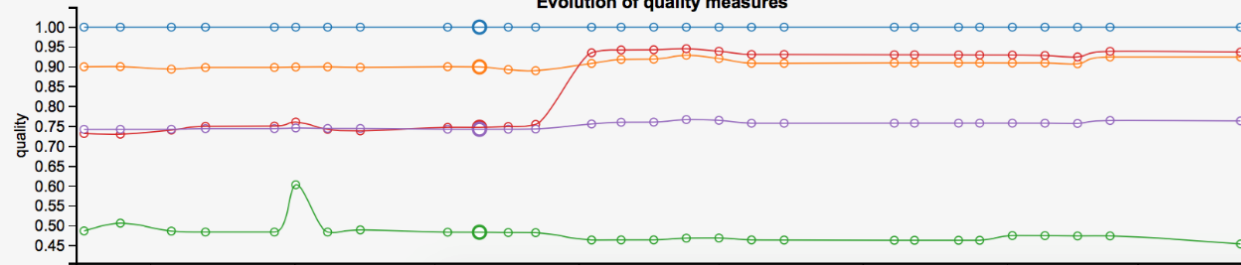
EVOLUTION

Available snapshots

Fri Oct 10 2014 14:51:16 GMT+0200 (CEST)



Evolution of quality measures



# ODP CHANGES

## Changes between the first and last snapshots

### dataset changes

70 PORTALS WITH DATASET CHANGES

- Avg. increase by 87.05% for 60 portals
- Avg. decrease by -64.16% for 10 portals

Show  entries









Search:

↑ PORTAL	↑ FROM	↑ TO	↑ CHANGE	↓ CHANGE PERCENTAGE
<b>data.sa.gov.au</b> <i>(2014-07-17)→(2015-03-15)</i>	484	5721	5237	1082.02%
<b>datos.codeandomexico.org</b> <i>(2014-07-17)→(2015-03-15)</i>	94	715	621	660.64%
<b>data.opendataportal.at</b> <i>(2014-07-17)→(2015-03-16)</i>	46	323	277	602.17%
<b>annuario.comune.fi.it</b> <i>(2014-08-07)→(2015-03-15)</i>	50	351	301	602.00%
<b>udct-data.aigid.jp</b> <i>(2014-08-07)→(2015-03-16)</i>	431	2110	1679	389.56%
<b>catalogo.datos.gob.mx</b> <i>(2014-08-08)→(2015-03-15)</i>	111	360	249	224.32%










# Data Dumps

- OPEN DATA PORTAL WATCH provides an archive of Open Data portal crawls (weekly snapshots/dynamic crawling framework):

## Open Data Portal Watch Dumps

Name	Last modified	Size
 Parent Directory		-
 africaopendata.org/	16-Mar-2015 13:03	-
 annuario.comune.fi.it/	16-Mar-2015 13:03	-
 bermuda.io/	16-Mar-2015 13:14	-
 catalog.data.gov/	05-Feb-2015 15:28	-
 catalog.data.ug/	16-Mar-2015 13:07	-
 catalogo.datos.gob.mx/	16-Mar-2015 13:08	-
 catalogodatos.gub.uy/	16-Mar-2015 13:15	-

## Open Data Portal Watch Dumps

Name	Last modified	Size
 Parent Directory		-
 2014-07-17.gz	05-Feb-2015 15:13	2.2M
 2014-07-25.gz	05-Feb-2015 15:13	2.2M
 2014-08-05.gz	05-Feb-2015 15:13	2.2M
 2014-08-12.gz	05-Feb-2015 15:13	2.2M
 2014-08-27.gz	05-Feb-2015 15:13	2.2M
 2014-09-01.gz	05-Feb-2015 15:14	2.2M
 2014-09-07.gz	05-Feb-2015 15:14	2.2M
 2014-09-14.gz	05-Feb-2015 15:14	2.2M

## Quality assessment & evolution of Open Data portals

Portal Overview

Jürgen Umbrich\*, Sebastian Neumaier\*, Axel Polleres\*  
Vienna University of Economics and Business, Vienna, Austria  
Email: \*firstname.lastname@wu.ac.at

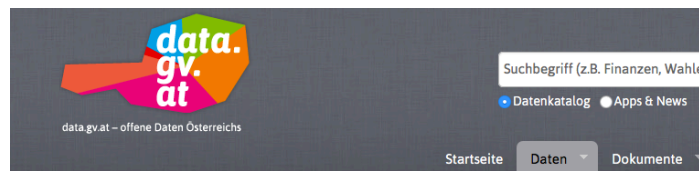
Best paper award  
at IEEE OBD-  
2015 ☺

<http://data.wu.ac.at/portalwatch/>

- Key findings:
  - Significantly varying quality across portals
  - Rapid growth for some portals
  - Huge variety and range of datasets
  - Open Data Portal **search** is a big problem
  - Time: many datasets only provide **current**, but no **historical** data

# Historical vs. current-only data (monotonic changes vs. non-monotonic changes)

- Weather data (every 15min) from 21 Austrian weather stations...
- VS.
- Population per gender and age in Vienna districts



## Katalog Meteorologische Messdaten der ZAMG

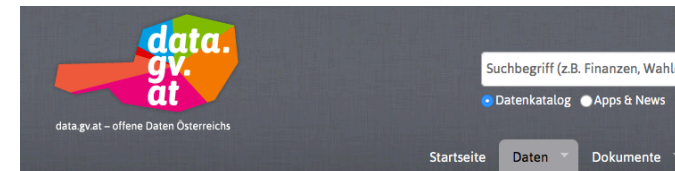
Aktuelle Messwerte von 21 Wetterstationen in Österreich. Die Daten werden stündlich aktualisiert. Sie beinhalten neben Stationsnummer, Stationsname, Seehöhe der Station, Messdatum und Messzeit (Lokalzeit) die meteorologischen Messwerte von Temperatur, Taupunkt, relative Luftfeuchtigkeit, Richtung und Geschwindigkeit des Windmittels und der Windspitze, Niederschlagssumme der letzten Stunde, Luftdruck reduziert auf Meeresniveau und Luftdruck auf Stationsniveau sowie die Sonnenscheindauer der letzten Stunde (in Prozent). Die Messstationen, die diese Daten liefern, sind über das Bundesgebiet verteilt und beinhalten alle Landeshauptstädte sowie die wichtigsten Bergstationen.

Eindeutiger Identifikator	9b40a0af-a6fe-47ff-9624-2ea8f40c746f
Datum des Metadatensatzes	2016-03-03
Kategorie	Umwelt
Datenverantwortliche Stelle	Zentralanstalt für Meteorologie und Geodynamik
Lizenz	Creative Commons Namensnennung 3.0 Österreich
Zeitliche Ausdehnung (Anfang)	2012-08-05

Updated every  
15min,  
only current  
data



Updated  
annually,  
historical data  
since 2011



## Katalog Bevölkerung in Wien: Geburtsland - Geschlecht

Eindeutiger Identifikator	
Datum des Metadatensatzes	2016-02-29
Kategorie	Bevölkerung
Datenverantwortliche Stelle	Magistratsabteilung 23 - Wirtschaft
Lizenz	Creative Commons Namensnennung 3.0 Österreich
Zeitliche Ausdehnung (Anfang)	2011 -
Veröffentlichende Stelle	Stadt Wien

Zusätzliche Informationen [↑ ausblenden](#)

Connection to  
Challenge 1  
(Synchronization):  
Adequate meta-data  
could help us to steer  
crawling, and more  
efficient storage, we  
are experimenting  
with this...

# Now: How do data quality and archiving connect?

- Idea: if we know how data changed, we could assess “bogus” changes...
  - Look at **time series**: Detect outliers over historical data
  - Example: wikipedia change history!
    - Wrong/disputed data changes often!

e.g. obvious Idea:  
Once data is semantically integrated in an archive, it is easy to include time-series analysis to filter out implausible/inconsistent data automatically...



citydata.wu.ac.at

**WU** WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

**SIEMENS**

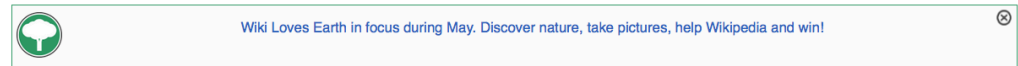
Vienna 

Population

- > **1991**: 1539848 persons (from <http://epp.eurostat.ec.europa.eu/>)
- > **1997**: 1609631 persons (from <http://epp.eurostat.ec.europa.eu/>)
- > **1998**: 1606843 persons (from <http://epp.eurostat.ec.europa.eu/>)
- > **1999**: 1608144 persons (from <http://epp.eurostat.ec.europa.eu/>)
- > **2000**: 1615438 persons (from <http://epp.eurostat.ec.europa.eu/>)
- > **2001**: 1829876 persons (from <http://data.un.org/>)
- > **2001**: 1550123 persons (from <http://data.un.org/>)
- > **2001**: 1550123 persons (from <http://epp.eurostat.ec.europa.eu/>)
- > **2004**: 1598626 persons (from <http://epp.eurostat.ec.europa.eu/>)
- > **2005**: 1626440 persons (from <http://data.un.org/>)

# Now: How do data quality and archiving connect?

- An idea only, so far: if we know how data changed, we could assess “bogus” changes...
  - Look at time series: Detect outliers over historical data
  - Example: wikipedia change history!
    - Assumption: Wrong/disputed data changes often!



## Maurice Jarre: Difference between revisions

From Wikipedia, the free encyclopedia

**Student hoaxes world's media on Wikipedia**  
Phony quote appears in obituaries for French composer Maurice Jarre

By **Shawn Pogatchnik**  
Associated Press  
updated 6:12 2009-10-09 26 AM ET

**DUBLIN** — When Dublin university student Shane Fitzgerald posted a poetic but phony quote on Wikipedia, he said he was testing how our globalized, increasingly Internet-dependent media was upholding accuracy and accountability in an age of instant news.

His report card: Wikipedia passed. Journalism flunked.

The sociology major's made-up quote — which he added to the Wikipedia page of Maurice Jarre hours after the French composer's death March 28 — flew straight on to dozens of U.S. blogs and newspaper Web sites in Britain, Australia and India.

They used the fabricated material, Fitzgerald said, even though administrators at the free online encyclopedia quickly caught the quote's lack of attribution and removed it, but not quickly enough to keep some journalists from cutting and pasting it.

A full month went by and nobody noticed the editorial fraud. So Fitzgerald told several media outlets in an e-mail and the corrections began.

"I was really shocked at the results from the experiment," Fitzgerald, 22, said Monday in an interview a week after one newspaper at fault, The Guardian of Britain, became the first to admit its obituarist lifted material straight from Wikipedia.

Shane Fitzgerald's obituary-friendly quote — which he added to the Wikipedia page of Maurice Jarre hours after the French composer's death March 28 — flew straight on to dozens of U.S. blogs and newspaper Web sites in Britain, Australia and India.

Revision as of 23:39, 29 March 2009 (edit)  
70.92.177.157 (talk)  
← Previous edit

Revision as of 02:29, 30 March 2009 (edit) (undo)  
86.42.227.123 (talk)  
Next edit →

**Line 54:**

==Music style==

Jarre wrote mainly for [[orchestra]]s, but began to favor [[synthesizer|synthesized]] music in the 1980s, mostly for practical rather than aesthetic motivations, many critics feel.{{Factdate=February 2007}} Jarre denies this and has pointed out that his electronic score for "[[Witness (1985 film)|Witness]]" was actually more laborious, time-consuming and expensive to produce than an orchestral score. Jarre's electronic scores from the 80s also include "[[Fatal Attraction]]", "[[The Year of Living Dangerously]]" and "[[No Way Out (1987 film)|No Way Out]]". A number of his scores from that era also feature electronic/acoustic blends, such as "[[Gorillas in the Mist]]", "[[Dead Poets Society]]", "[[The Mosquito Coast]]" and "[[Jacob's Ladder (movie)|Jacob's Ladder]]".

==Quotes==

Nowadays, if a studio assumes that his film is bad, there is always an executive that gets more nervous than usual and thinks that if they change the music, the film will become a masterpiece.

One could say my life itself has been one long soundtrack. Music was my life, music brought me to life, and music is how I will be remembered long after I leave this life. When I die there will be a final waltz playing in my head and that only I can hear.

When I was 15, I did not know nothing about what concerned the world of music

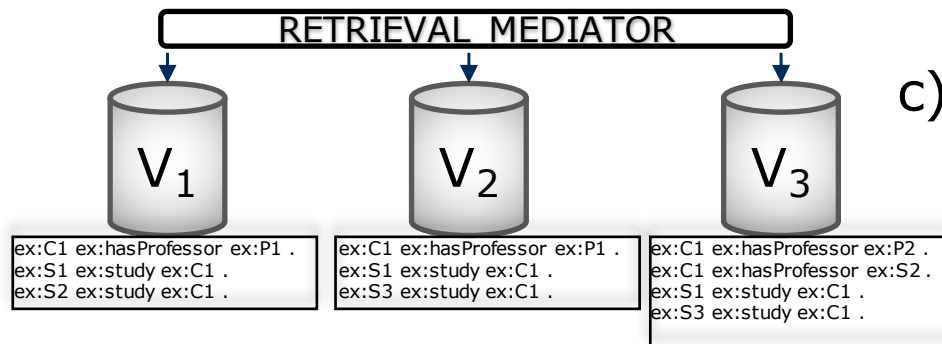


## **Finally: The *archiving problem***

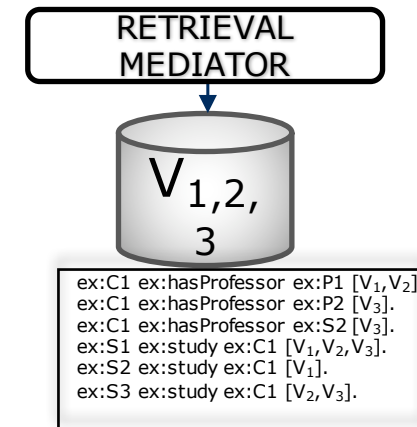
Now, how can we efficiently archive and perform time-based retrieval queries of a dataset?

# RDF Archiving. Archiving policies

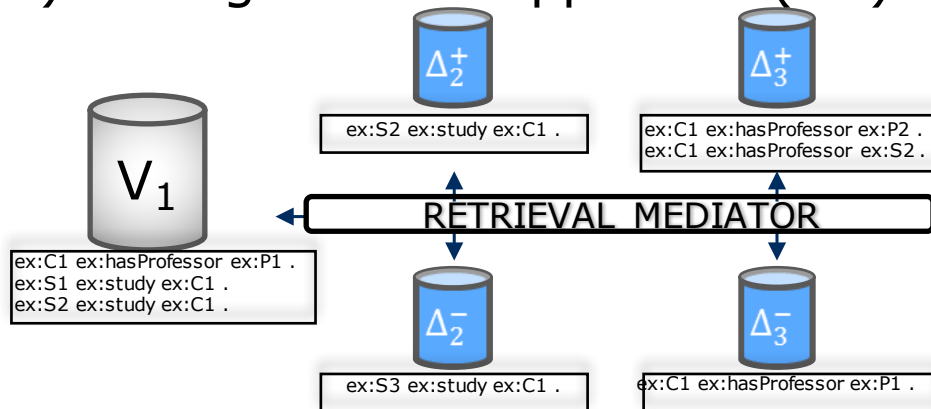
## a) Independent Copies/Snapshots (IC)



## c) Timestamp-based approach (TB)



## b) Change-based approach (CB)



# RDF Archiving. Querying

- Structured query languages managing time.
  - Temporal databases (T-Quel, TSQL2)
    - Overlapping, meeting, before, equal, during, finish
  - RDF/Linked Data
    - SPARQL extensions
      - T-SPARQL, SPARQL-ST
      - AnQL
    - DIACHRON Query Language
      - SPARQL with specific constructors such as DATASET (similar to a named graph), VERSION, or CHANGES

# BEAR:

## Benchmarking the Efficiency of RDF Archives

- Blueprint on benchmarking archives of semantic data
  - How can one define the corpus?
  - How can one design benchmark queries? Which queries?
- BEAR: concrete basic benchmark
  - Data: Crawl from Linked Data Observatory
  - Basic queries: Materialize, get Version...
  - Initial evaluation on archiving policies



# BEAR: Benchmarking the Efficiency of RDF Archiving

- **Blueprint on benchmarking archives of semantic data**
  - **How can one define the corpus?**
  - **How can one design benchmark queries? Which queries?**
- BEAR: concrete basic benchmark
  - Data: Crawl from Linked Data Observatory
  - Basic queries: Materialize, get Version...
  - Initial evaluation on archiving policies



# BEAR: Benchmarking the Efficiency of RDF Archiving

- Define the corpus
  - ☑ Number of versions / size

*Definition 1 (RDF Archive). A version-annotated triple is an RDF triple  $(s, p, o)$  with a label  $i \in \mathcal{N}$  representing the version in which this triple holds, denoted by the notation  $(s, p, o) : [i]$ . An RDF archive graph  $\mathcal{A}$  is a set of version-annotated triples.*

- ☑ Data dynamicity
  - ☑ Version change ratio
  - ☑ Version data growth
- ☑ Data static core
- ☑ Total triples (version-oblivious)
- ☑ RDF vocabulary
  - ☑ Per version / evolution

# BEAR: Benchmarking the Efficiency of RDF Archiving

- Design of benchmark queries
  - Cardinality / Selectivity + dynamicity
  - ☑ Archive-driven C/S/D
  - ☑ Version-driven C/S/D
  - ☑ Basic temporal retrieval features of queries
    - ☑ Version/Delta Materialization ( $V_i$ )
    - ☑ Version(Q): in which version Q is not empty
    - ☑ Change( $V_i, V_j$ ): true if delta  $\neq$  null
    - ☑ Join( $Q_1, V_i, Q_2, V_j$ )
    - ☑ Change(Q): Returns versions in which  $\text{Diff}(Q, V_i, V_{i-1}) \neq \emptyset$

# BEAR: instantiation of basic query features

- Instantiation of basic archive queries, e.g. in AnQL [1]

- *Antoine Zimmermann, Nuno Lopes, Axel Polleres, & Straccia. A general framework for representing, reasoning and querying with annotated Semantic Web data. Journal of Semantic Web (JWS), 12:72--95, March 2012.*

- *Mat(Q, V)*
- *Diff(Q, V1, V2)*
- *Ver(Q)*
- *join(Q1, vi, Q2, vj)*
- **Change(Q)**

```
SELECT ?V1 ?V2 WHERE
{ {{P :?V1 } MINUS {P :?V2}} UNION
  {{P :?V2 } MINUS {P :?V1}}
  FILTER( abs(?V1-?V2) = 1 ) }
```

Open question  
remains:  
What is the right  
query syntax for  
archive queries?



# BEAR: Benchmarking the Efficiency of RDF Archiving

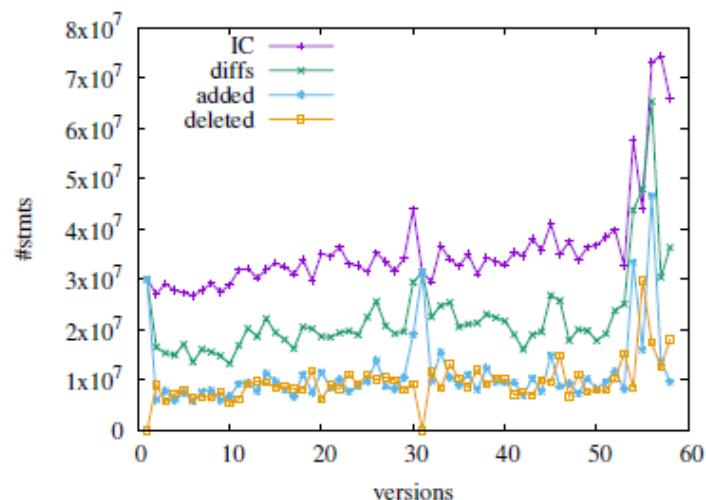
- blueprint on benchmarking archives of semantic data
  - How can one define the corpus?
  - How can one design benchmark queries? Which queries?
- **BEAR: concrete basic benchmark**
  - **Data: Crawl from Linked Data Observatory**
  - **Basic queries: Materialize, get Version...**
  - **Initial evaluation of archiving policies (IC,CB,TB)**



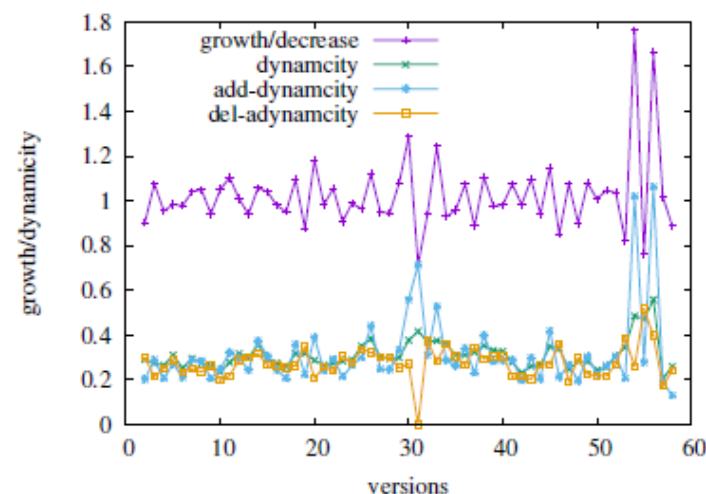
# BEAR: Benchmarking the Efficiency of RDF Archiving

- Corpus

versions	$ V_0 $	$ V_{57} $	$\overline{growth}$	$\overline{\delta}$	$\overline{\delta^-}$	$\overline{\delta^+}$	$\mathcal{C}_A$	$\mathcal{O}_A$
58	30m	66m	101%	31%	32%	27%	3.5m	376m



(a) Number of statements



(b) Relative growth and dynamicity

# BEAR: Benchmarking the Efficiency of RDF Archiving

- Queries and systems
  - We implemented and evaluate archiving systems on **Jena-TDB** and **HDT**, based on IC, CB and TB policies.
    - Confirm the initial premises of the archiving policies:
      - In space, IC is the worst, CB improves the space and TB increases the size as it has to index a new dimension
      - In time, CB is bad at getting a particular version because it has to reapply the changes ... (but good e.g. for `Change(SELECT * {?S ?P ?O})`)
    - Serve as an initial baseline to compare archiving systems
    - More info: <https://github.com/webdata/BEAR>

# Finally, many open questions remain still!

## Archiving and querying evolving semantic Web data

Objective	Research Question
Representation	<ul style="list-style-type: none"><li>❖ minimize the redundant information</li><li>❖ respect the original modeling and provenance information</li></ul>
Query language	<ul style="list-style-type: none"><li>❖ capture the expressiveness of emerging retrieval demands in archiving</li><li>❖ our base operations are meant to be an <b>extensible</b> starting point</li><li>❖ design a query language satisfying these requirements for evolving interlinked data</li></ul>
Indexing	<ul style="list-style-type: none"><li>❖ index archives at large scale (and keeping up with evolution rate – streaming vs. archiving) to process the queries efficiently</li></ul>
Query optimization	<ul style="list-style-type: none"><li>❖ optimizing query resolution plans for archives</li><li>❖ enabling the integration of other sources (federated infrastructure)</li><li>❖ Query rewriting for querying archives of structured non-RDF sources? Open Data!</li></ul>
Application	<ul style="list-style-type: none"><li>❖ Is there a actual and urgent need in the community?</li><li>❖ We believe yes, but where's the killer-app?</li></ul>

# Thanks!

**Dept. of Information Systems &  
Operations**

Institute for Information Business  
Welthandelsplatz 1, 1020 Vienna, Austria

**Univ.Prof. Dr. Axel Polleres**

T +43-1-31336/5297  
[axel.polleres@wu.ac.at](mailto:axel.polleres@wu.ac.at)  
[polleres.net](http://polleres.net)

Big (Semantic) Data  
Versions  
Evolving Data  
Streaming  
Compression

- Instantiation of archive queries in AnQL [1]
  - *Antoine Zimmermann, Nuno Lopes, Axel Polleres, and Umberto Straccia. A general framework for representing, reasoning and querying with annotated Semantic Web data. *Journal of Web Semantics (JWS)*, 12:72--95, March 2012.*
  - ***Mat(Q,V)***
  - *Diff(Q,V1,V2)*
  - *Ver(Q)*
  - *join(Q1,vi,Q2,vj)*
  - *Change(Q)*

```
SELECT * WHERE { Q :[v] }
```

- Instantiation of archive queries in AnQL [1]
  - *Antoine Zimmermann, Nuno Lopes, Axel Polleres, and Umberto Straccia. A general framework for representing, reasoning and querying with annotated Semantic Web data. Journal of Web Semantics (JWS), 12:72--95, March 2012.*
  - *Mat(Q,V)*
  - ***Diff(Q,V1,V2)***
  - *Ver(Q)*
  - *join(Q1,vi,Q2,vj)*
  - *Change(Q)*

```
SELECT * WHERE {  
  { { {Q :[v1]} MINUS {Q :[v2]} } BIND (v1 AS ?V )  
  }  
  UNION  
  { { {Q :[v2]} MINUS {Q :[v1]} } BIND (v2 AS ?V )  
  }  
}
```

- Instantiation of archive queries in AnQL [1]
  - *Antoine Zimmermann, Nuno Lopes, Axel Polleres, and Umberto Straccia. A general framework for representing, reasoning and querying with annotated Semantic Web data. *Journal of Web Semantics (JWS)*, 12:72--95, March 2012.*
  - *Mat(Q, V)*
  - *Diff(Q, V1, V2)*
  - **Ver(Q)**
  - *join(Q1, vi, Q2, vj)*
  - *Change(Q)*

```
SELECT * WHERE { P :?V }
```



- Instantiation of archive queries in AnQL [1]
  - *Antoine Zimmermann, Nuno Lopes, Axel Polleres, and Umberto Straccia. A general framework for representing, reasoning and querying with annotated Semantic Web data. *Journal of Web Semantics (JWS)*, 12:72--95, March 2012.*
  - *Mat(Q, V)*
  - *Diff(Q, V1, V2)*
  - *Ver(Q)*
  - **join(Q1, v1, Q2, v2)**
  - *Change(Q)*

```
SELECT * WHERE { {Q :[v1]} {Q :[v2]} }
```

- Instantiation of archive queries in AnQL [1]
  - *Antoine Zimmermann, Nuno Lopes, Axel Polleres, and Umberto Straccia. A general framework for representing, reasoning and querying with annotated Semantic Web data. Journal of Web Semantics (JWS), 12:72--95, March 2012.*
  - *Mat(Q, V)*
  - *Diff(Q, V1, V2)*
  - *Ver(Q)*
  - *join(Q1, vi, Q2, vj)*
  - **Change(Q)**

```
SELECT ?V1 ?V2 WHERE
{ {{P :?V1 } MINUS {P :?V2}} UNION
  {{P :?V2 } MINUS {P :?V1}}
  FILTER( abs(?V1-?V2) = 1 ) }
```