On Archiving Linked and Open Data

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

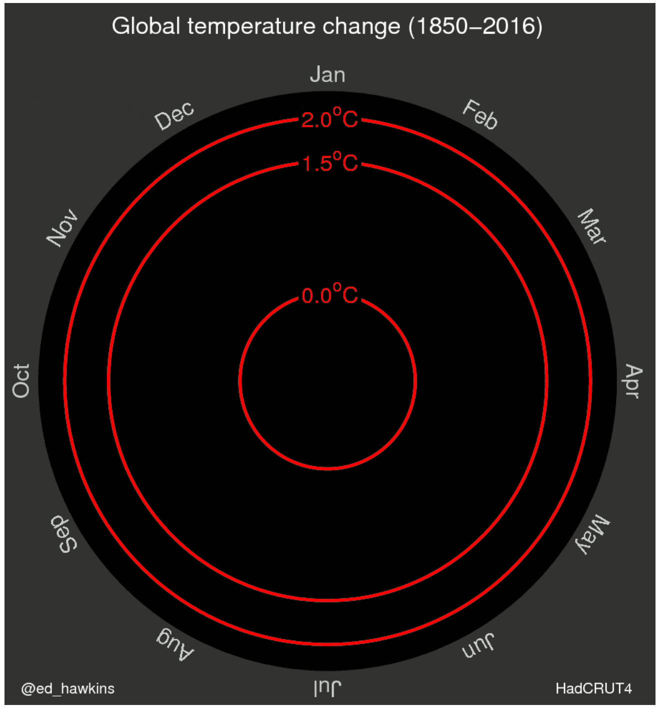
30.05.2016 - 2nd Workshop on Managing the Evolution and Preservation of the Data Web, MEPDaW@ESWC16

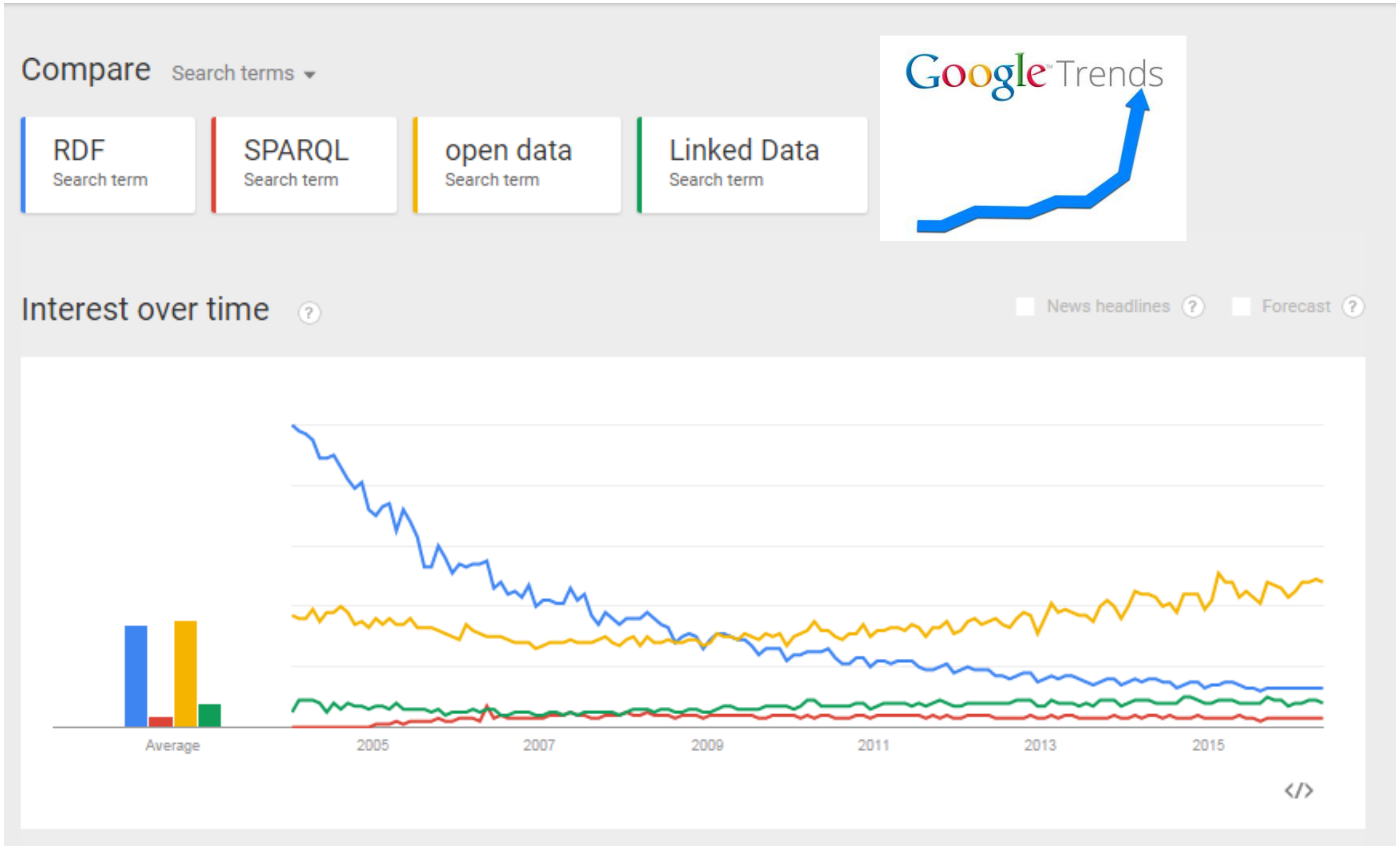
What we will talk about...

- ***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...)

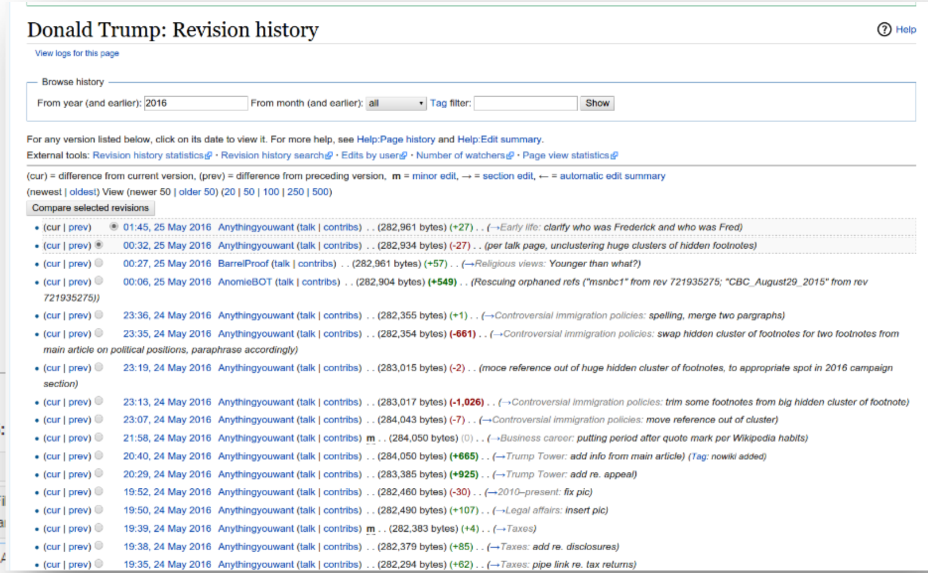
- Monitoring evolution is relevant

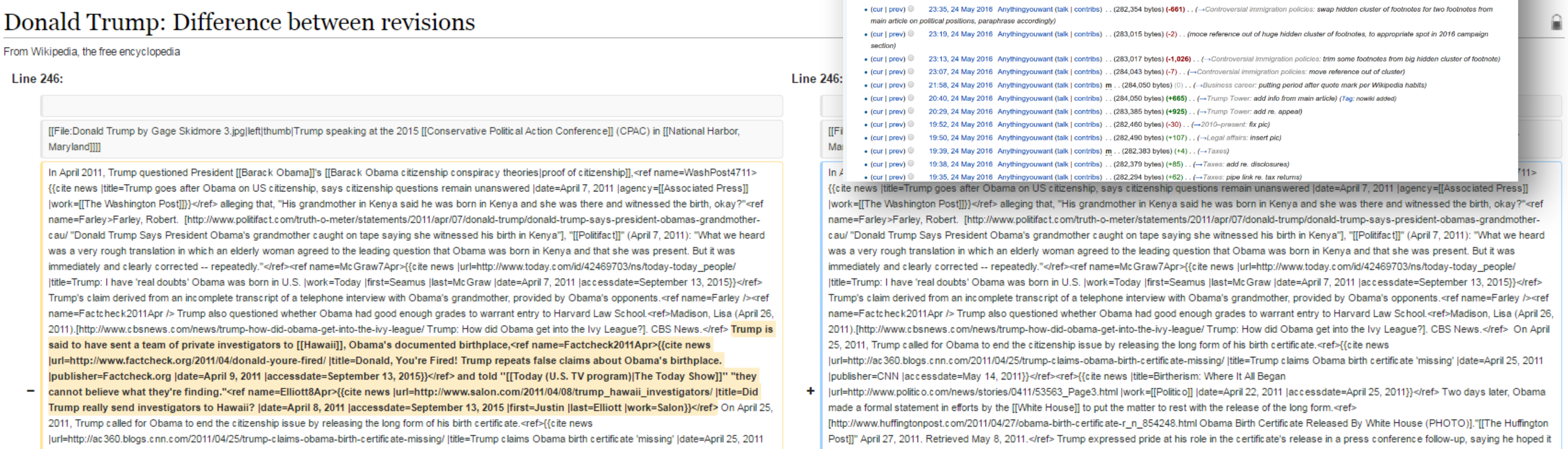




Evolution matters

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

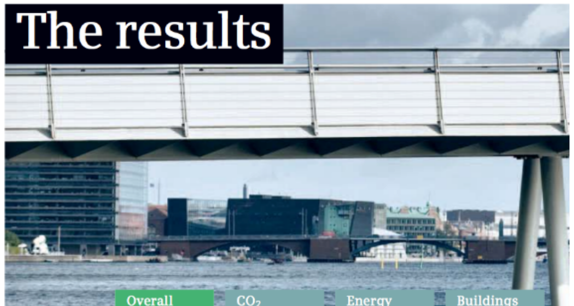




Evolution matters

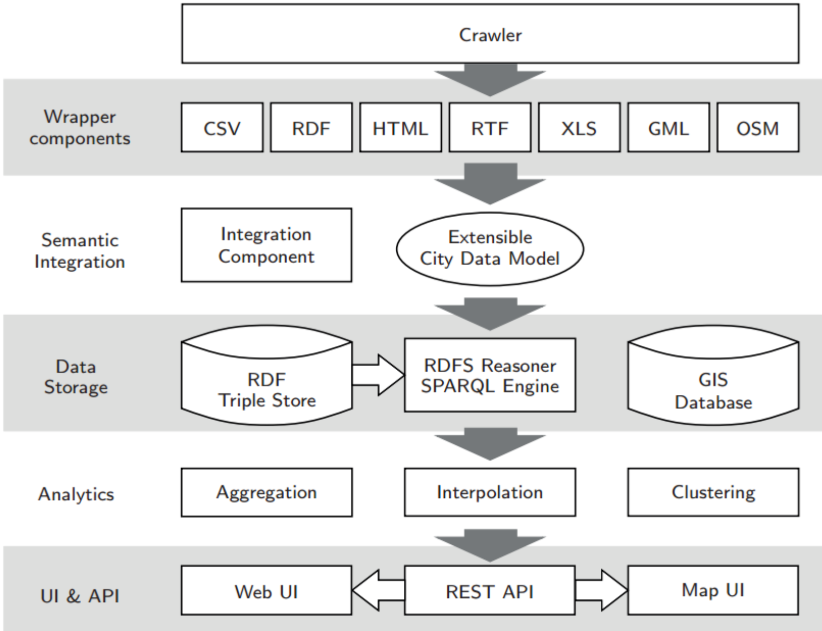
- Evolution may reveal actions vs. consequences
  - E.g. CityDataPipeline, http://citydata.wu.ac.at/
    - Collecting, Integrating and Predicting Open City Data

European Green City Index | The results



The complete results from the index, including the overall result of each city as well as the individual rankings within the eight categories.

Overall			CO2			Energy			Buildings		
	City	Score		City	Score		City	Score		City	Score
1	Copenhagen	87,31	1	Oslo	9,58	1	Oslo	8,71	=1	Berlin	9,44
2	Stockholm	86,65	2	Stockholm	8,99	2	Copenhagen	8,69	=1	Stockholm	9,44
3	Oslo	83,98	3	Zurich	8,48	3	Vienna	7,76	3	Oslo	9,22
4	Vienna	83,34	4	Copenhagen	8,35	4	Stockholm	7,61	4	Copenhagen	9,17
5	Amsterdam	83,03	5	Brussels	8,32	5	Amsterdam	7,08	5	Helsinki	9,11
6	Zurich	82,31	6	Paris	7,81	6	Zurich	6,92	6	Amsterdam	9,01
7	Helsinki	79,29	7	Rome	7,57	7	Rome	6,40	7	Paris	8,96
8	Berlin	79,01	8	Vienna	7,53	8	Brussels	6,19	8	Vienna	8,62
9	Brussels	78,01	9	Madrid	7,51	9	Lisbon	5,77	9	Zurich	8,43
10	Paris	73,21	10	London	7,34	10	London	5,64	10	London	7,96
11	London	71,56	11	Helsinki	7,30	11	Istanbul	5,55	11	Lisbon	7,34
12	Madrid	67,08	12	Amsterdam	7,10	12	Madrid	5,52	12	Brussels	7,14
13	Vilnius	62,77	13	Berlin	6,75	13	Berlin	5,48	13	Vilnius	6,91
14	Rome	62,58	14	Ljubljana	6,67	14	Warsaw	5,29	14	Sofia	6,25
15	Riga	59,57	15	Riga	5,55	15	Athens	4,94	15	Rome	6,16
16	Warsaw	59,04	16	Istanbul	4,86	16	Paris	4,66	16	Warsaw	5,99
17	Budapest	57,55	=17	Athens	4,85	17	Belgrade	4,65	17	Madrid	5,68
18	Lisbon	57,25	=17	Budapest	4,85	18	Dublin	4,55	18	Riga	5,43
19	Ljubljana	56,39	19	Dublin	4,77	19	Helsinki	4,49	19	Ljubljana	5,20
20	Bratislava	56,09	20	Warsaw	4,65	20	Zagreb	4,34	20	Budapest	5,01
21	Dublin	53,98	21	Bratislava	4,54	21	Bratislava	4,19	21	Bucharest	4,79
22	Athens	53,09	22	Lisbon	4,05	22	Riga	3,53	22	Athens	4,36
23	Tallinn	52,98	23	Vilnius	3,91	23	Bucharest	3,42	23	Bratislava	3,54
24	Prague	49,78	24	Bucharest	3,65	24	Prague	3,26	24	Dublin	3,39
25	Istanbul	45,20	25	Prague	3,44	25	Budapest	2,43	25	Zagreb	3,29
26	Zagreb	42,36	26	Tallinn	3,40	26	Vilnius	2,39	26	Prague	3,14
27	Belgrade	40,03	27	Zagreb	3,20	27	Ljubljana	2,23	27	Belgrade	2,89
28	Bucharest	39,14	28	Belgrade	3,15	28	Sofia	2,16	28	Istanbul	1,51
29	Sofia	36,85	29	Sofia	2,95	29	Tallinn	1,70	29	Tallinn	1,06
30	Kiev	32,33	30	Kiev	2,49	30	Kiev	1,50	30	Kiev	0,00



Vienna
Environment

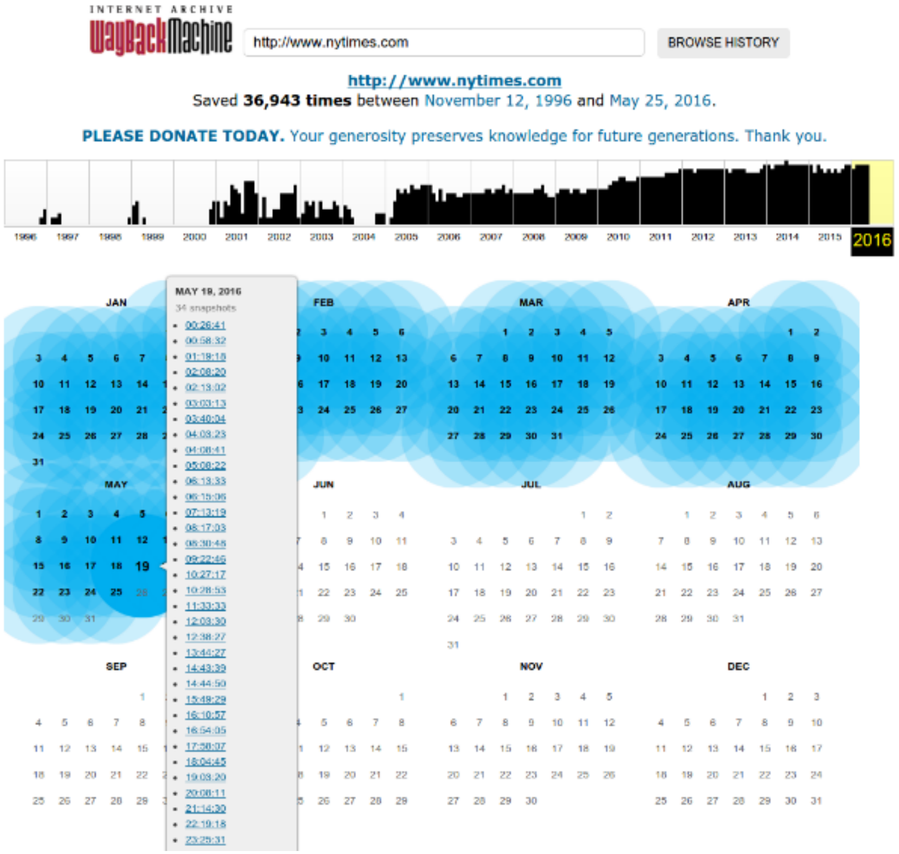
[Accumulated ozone concentration]

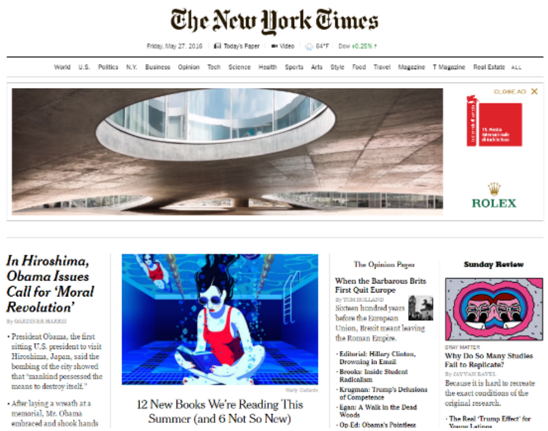
- **1992**: 6112.40 microgram per m3 (from http://epp.eurostat.ec.europa.eu/)
- **1994**: 5998.80 microgram per m3 (from http://epp.eurostat.ec.europa.eu/)
- **1995**: 5513.80 microgram per m3 (from http://epp.eurostat.ec.europa.eu/)
- **1997**: 5181.80 microgram per m3 (from http://epp.eurostat.ec.europa.eu/)
- **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

Preservation matters

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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

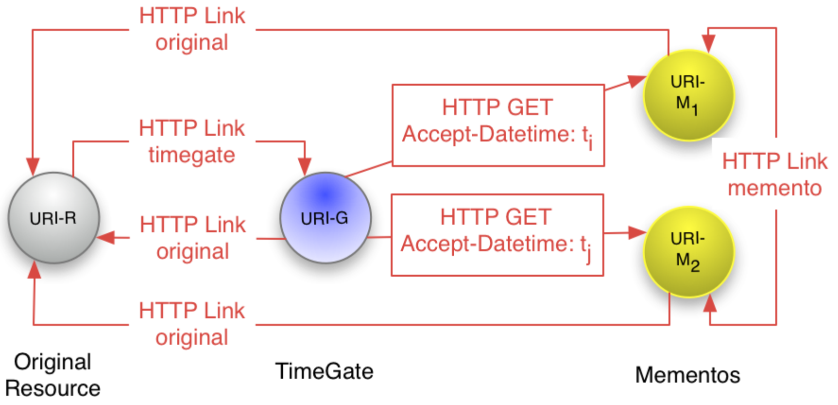




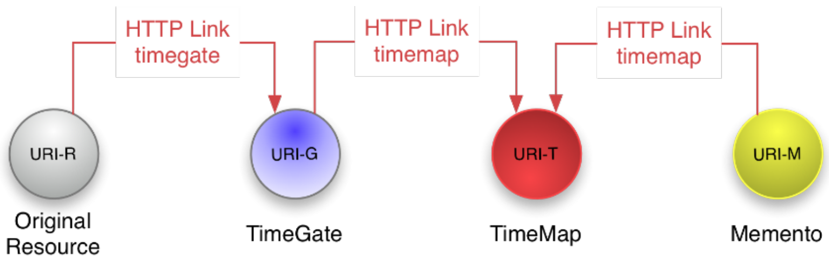
Time-based access matters

- The Memento protocol RFC 7089

Follow your nose
(HTTP content negotiation with datetime)



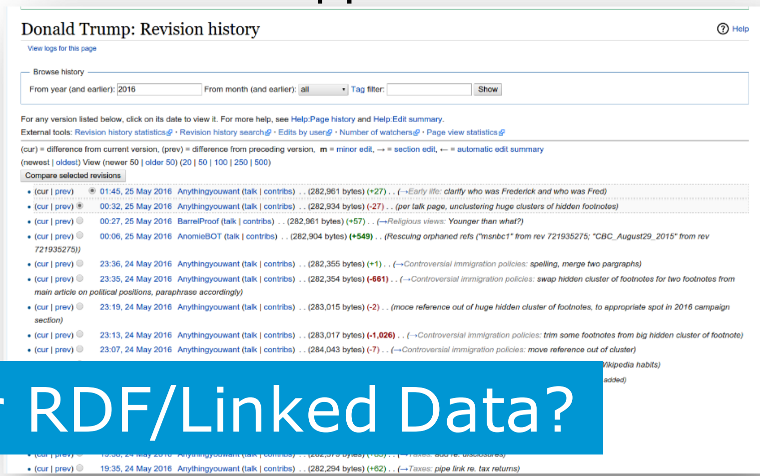
Batch discovery
(list of URIs of Mementos of the Original Resource)

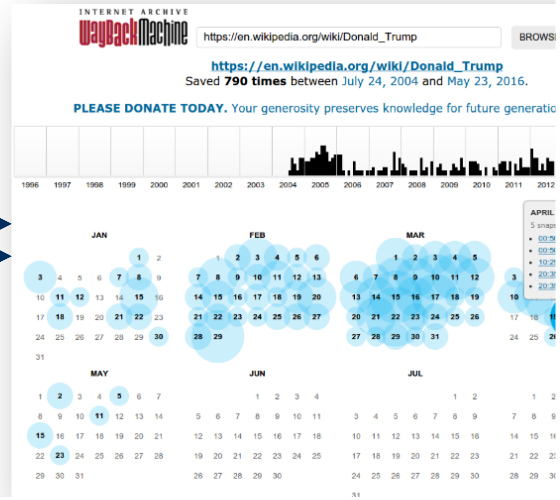


But…

Challenges

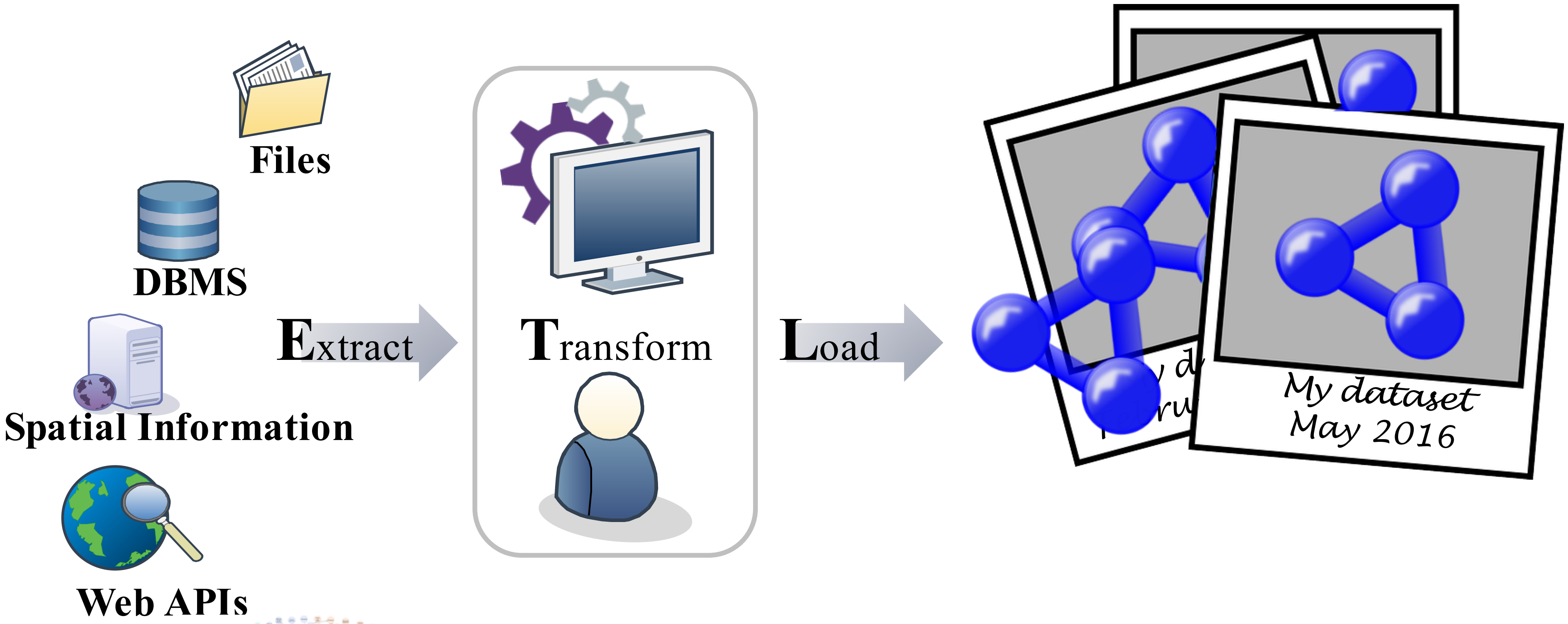
- Poor **granularity** ("some" snapshots)
- **Aggregated** data, only, rather than raw data access
  - (e.g. in Google trends)
- Few work on archiving the **Web of Data**, or on integrating archives
- What is the right **query language**?
  - basic retrieval features (get version at timestamp **t**)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

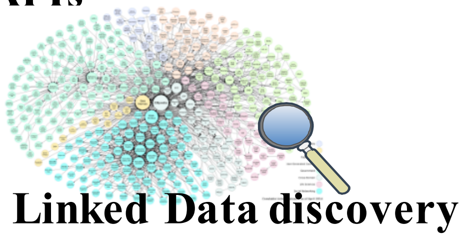




Is it easier/better for RDF/Linked Data?

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

RDF Graph V_1

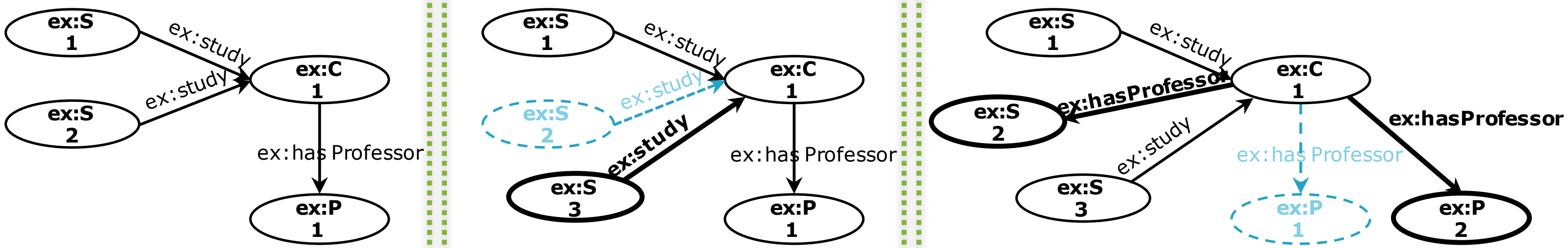
ex:C1 ex:hasProfessor ex:P1 .
ex:S1 ex:study ex:C1 .
ex:S2 ex:study ex:C1 .

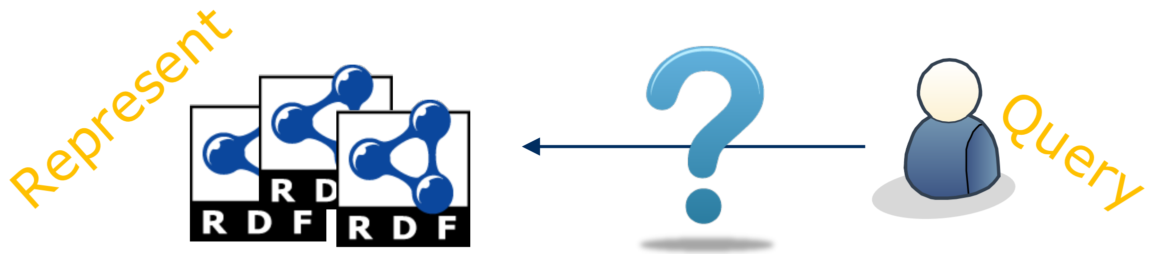
RDF Graph V_2

ex:C1 ex:hasProfessor ex:P1 .
ex:S1 ex:study ex:C1 .
~~ex:S2 ex:study ex:C1 .~~
ex:S3 ex:study ex:C1 .

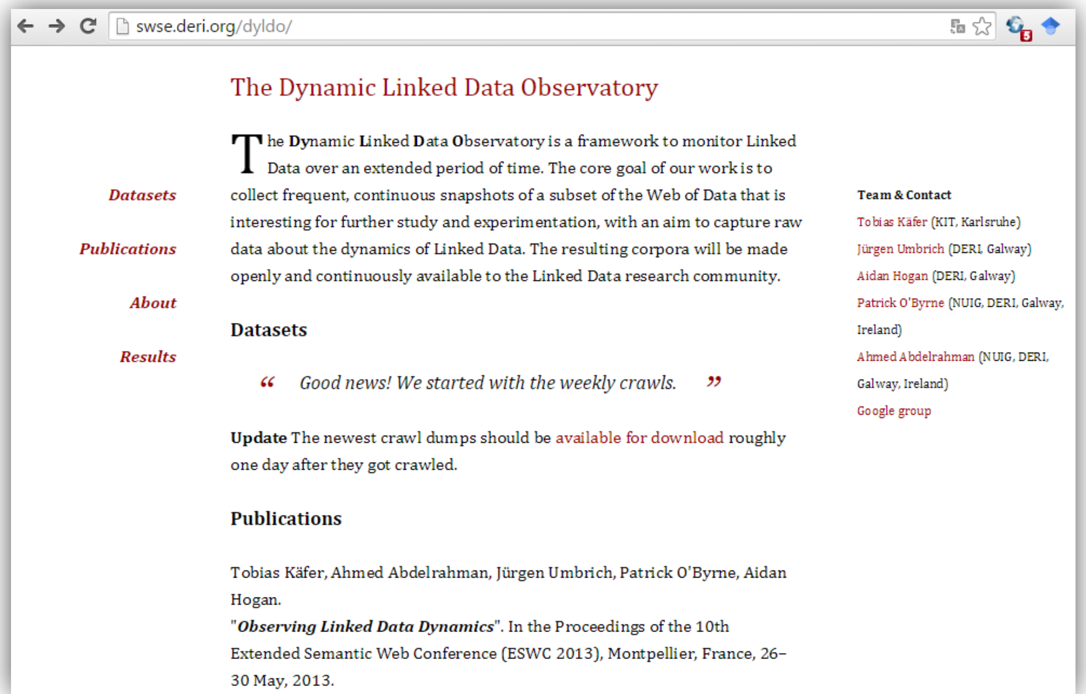
RDF Graph V_3

~~ex:C1 ex:hasProfessor ex:P1 .~~
ex:C1 ex:hasProfessor ex:P2 .
ex:C1 ex:hasProfessor ex:S2 .
ex:S1 ex:study ex:C1 .
ex:S3 ex:study ex:C1 .





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


swse.deri.org/dyldo/

The Dynamic Linked Data Observatory

Datasets
Publications
About
Results

The **D**ynamic **L**inked **D**ata **O**bservatory 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

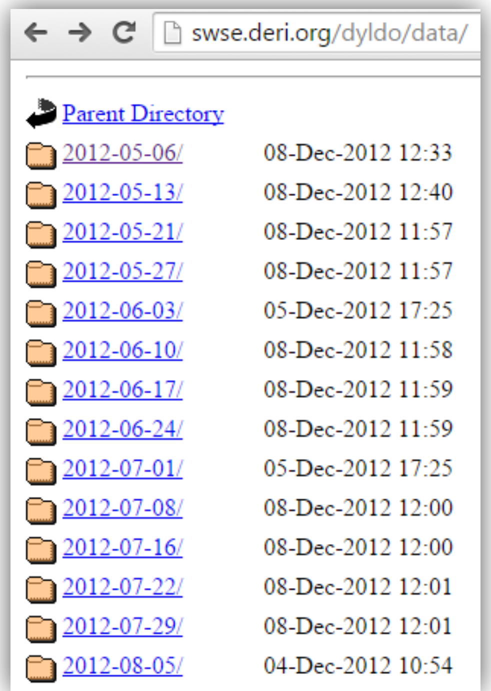
“ *Good news! We started with the weekly crawls.* ”

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.

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

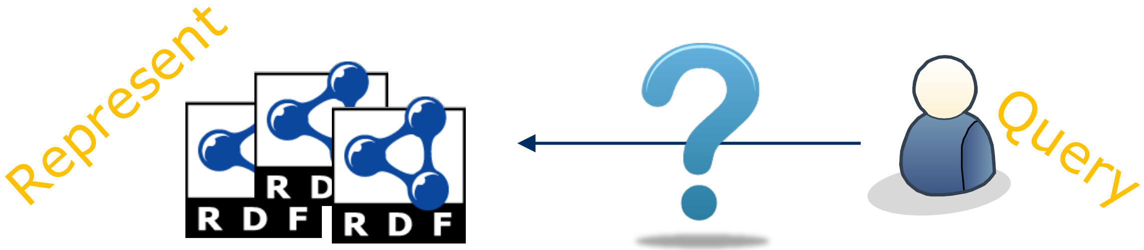

swse.deri.org/dyldo/data/

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?

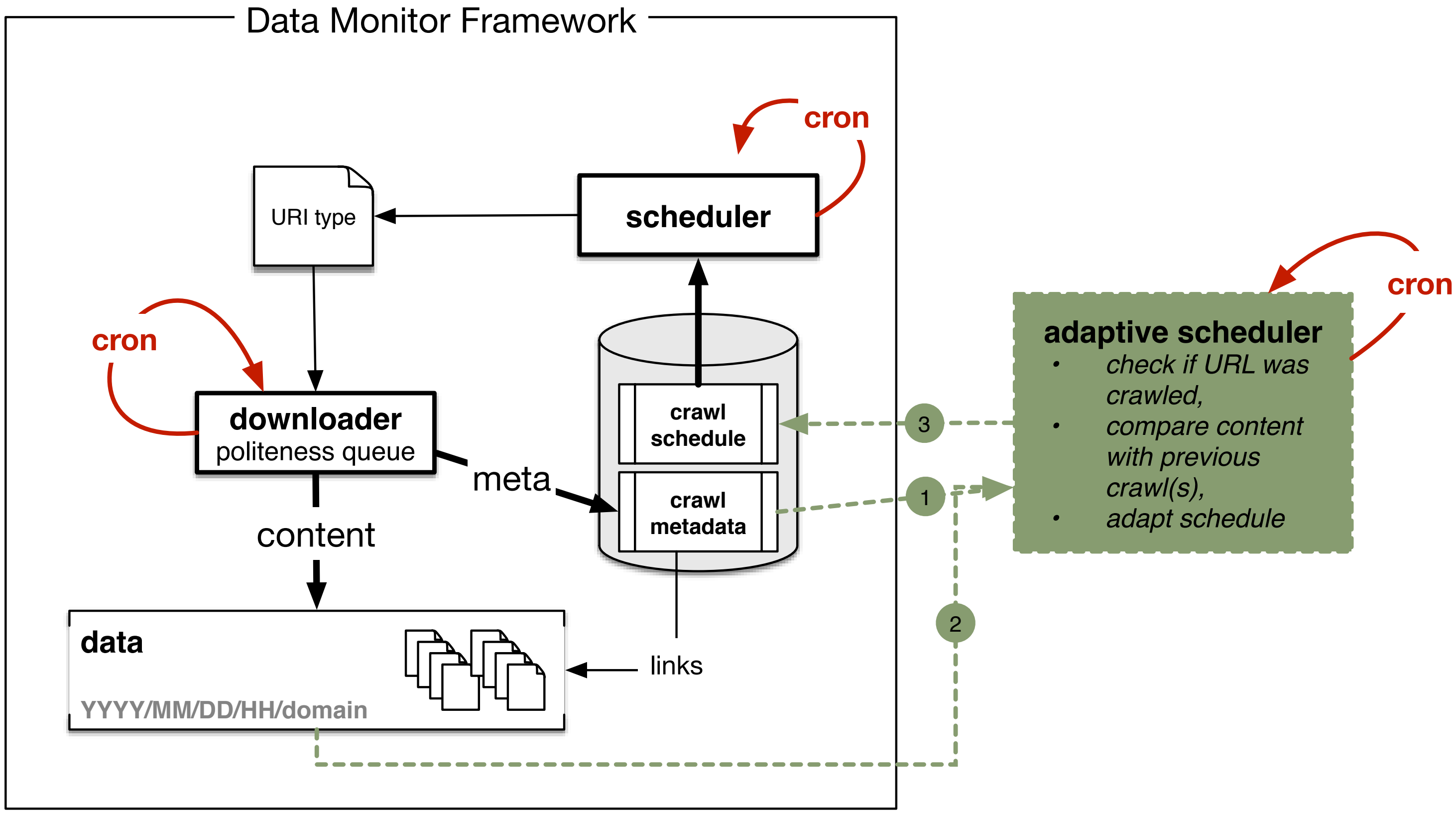
The synchronization problem

how can we monitor changes?

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

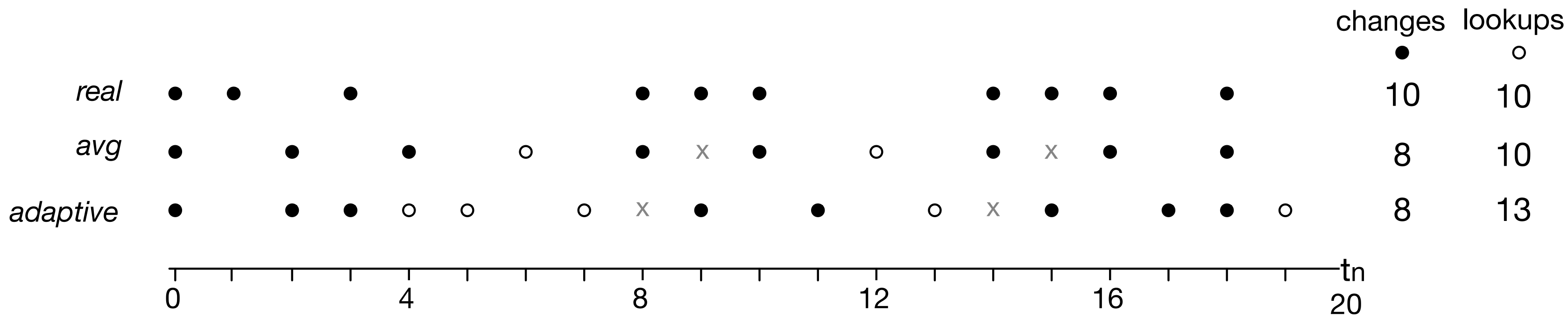
- Observations:
  - Some services that publish or are mapped to RDF change **regularly**, but we don't know the frequency upfront!
  - Some services mapped to RDF **announce/archive their changes already**, so they already keep an archive…

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



Experiment: Rescheduling

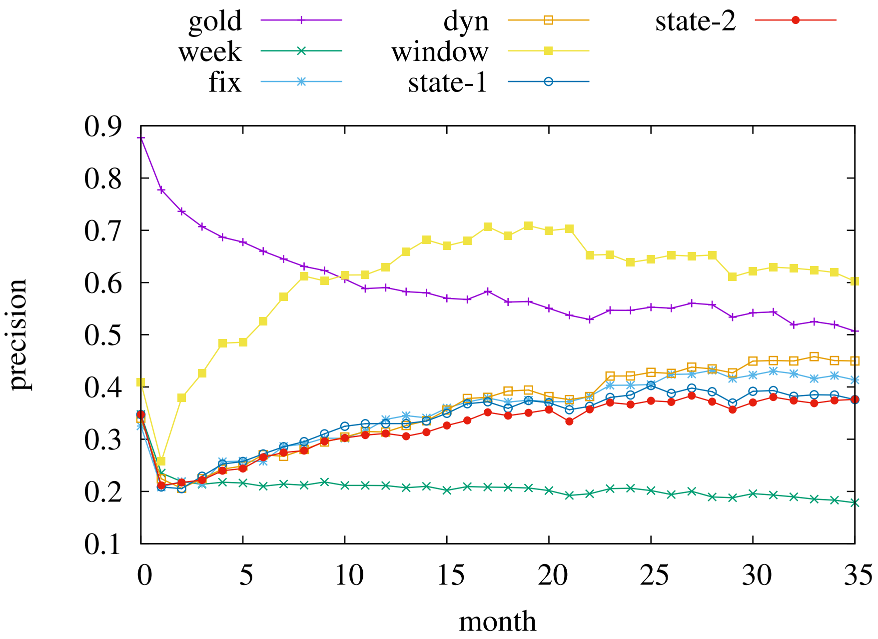
- Evaluating strategies to compute next crawl time for URLs to accurately capture content change

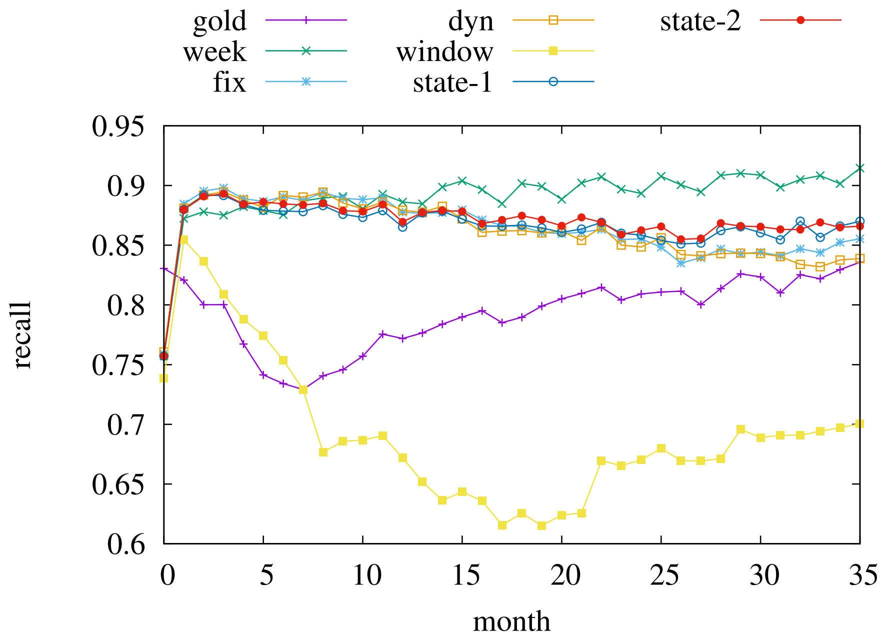


Test setup

- 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



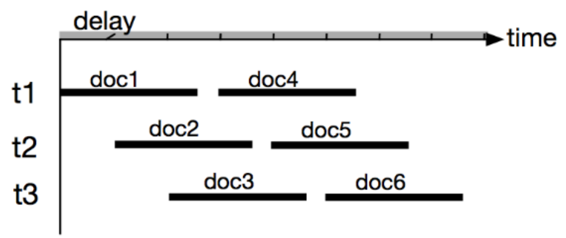


We observed the typical trade-off between recall and precision…
Strategies based on Markov models seems to provide best and most stable trade-off

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

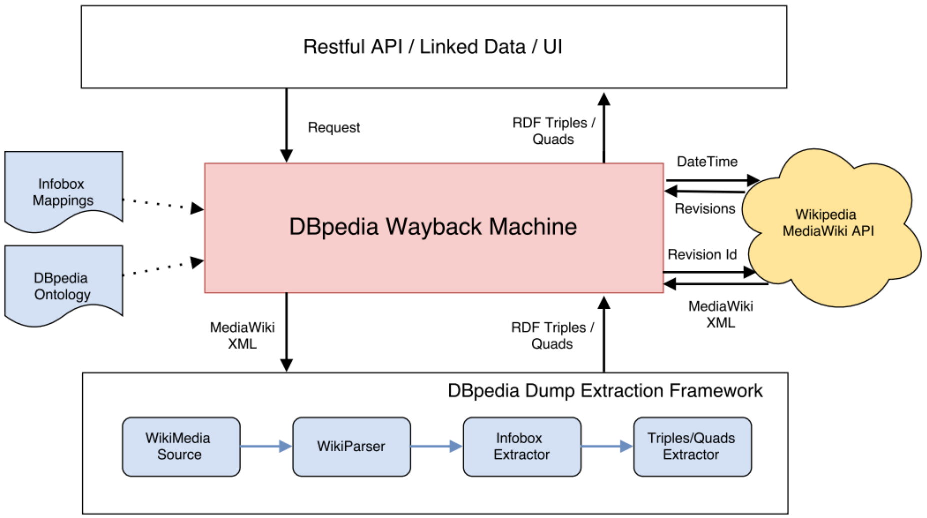
- Large over estimation (is acceptable)
- Small underestimation (e.g. 10mins for 60mins crawl)

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...)

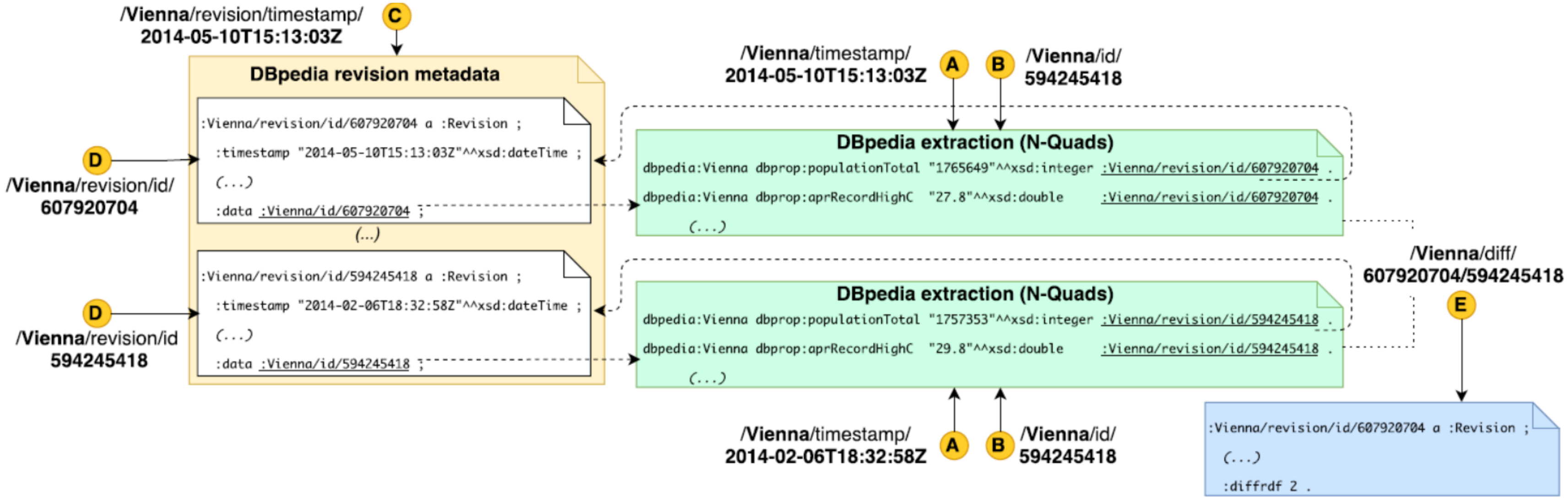
- If raw historical data on changes is is available...
- Aim: Fine grained access to previous versions, re-applying X2RDF transformations on the original source
- Example: **DBpedia Wayback machine**
  - Re-apply mappings on the Wikipedia revision history



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

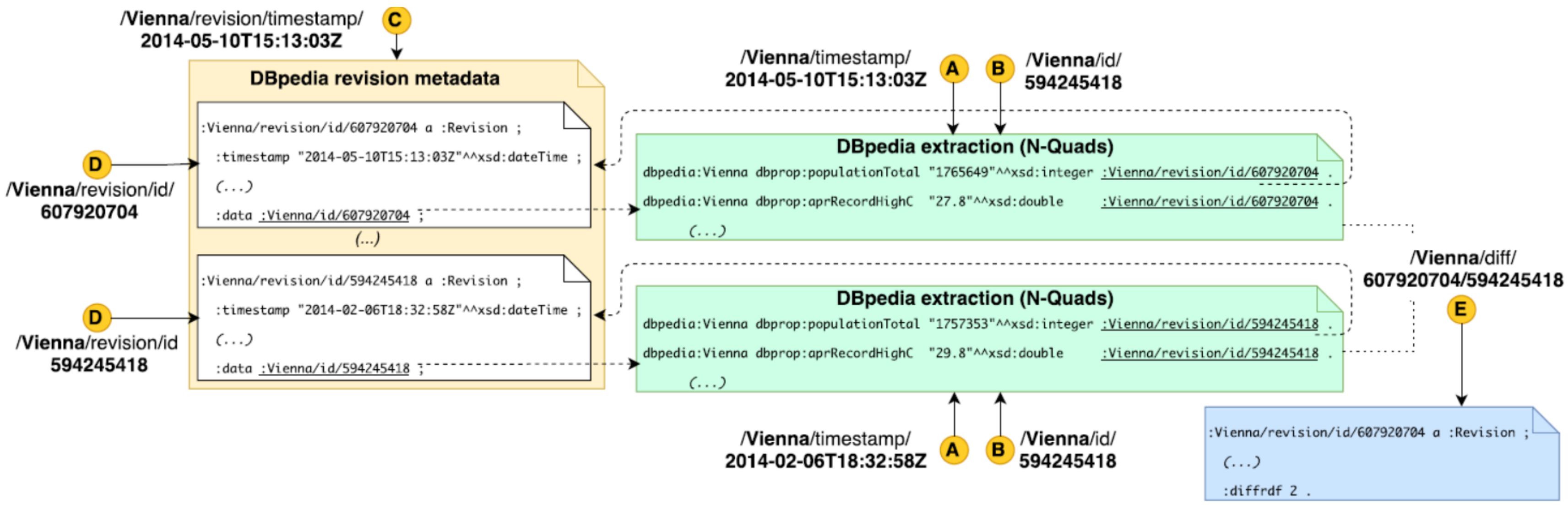
2- “Recreate” the versions

- 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



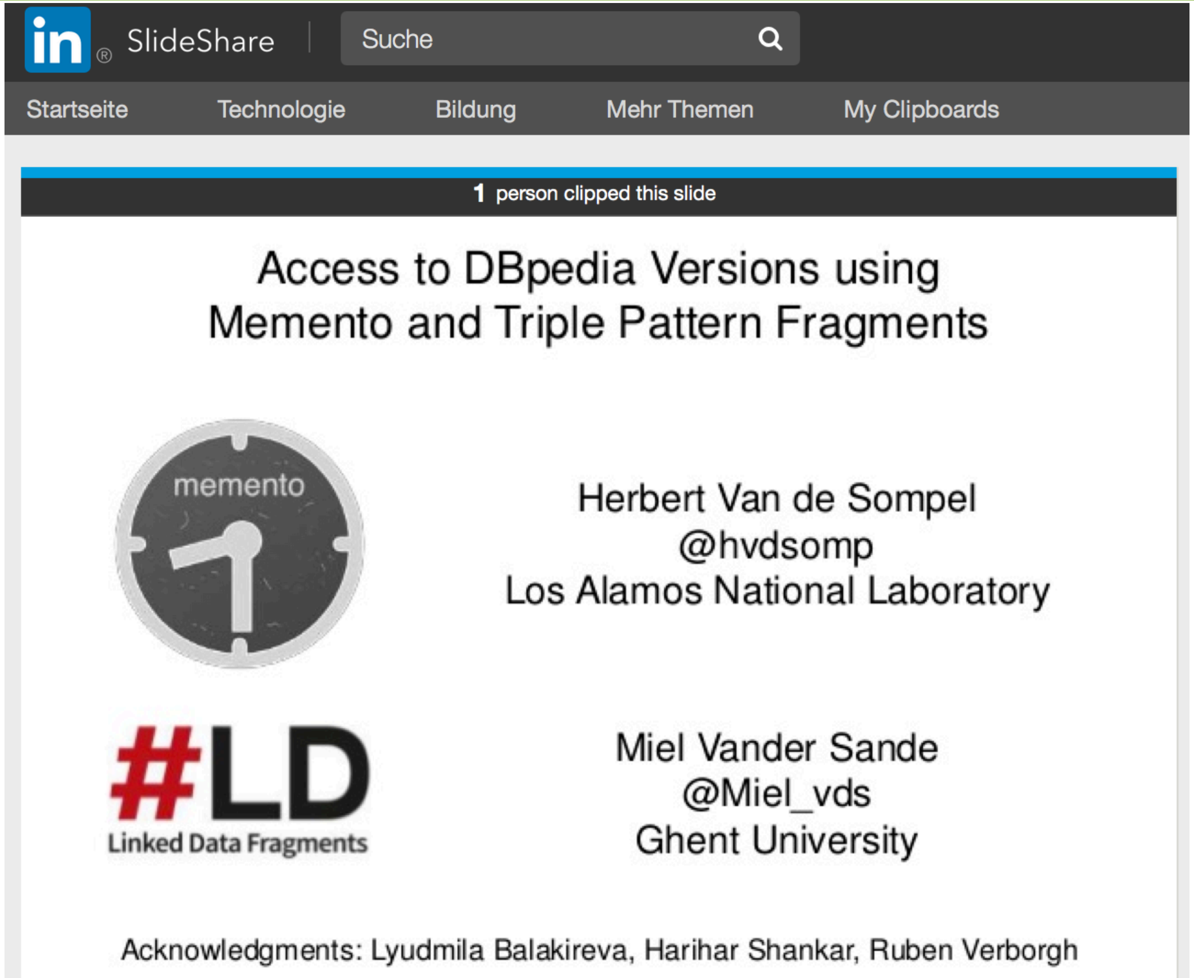
2- "Recreate" the versions

- 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…

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS



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/

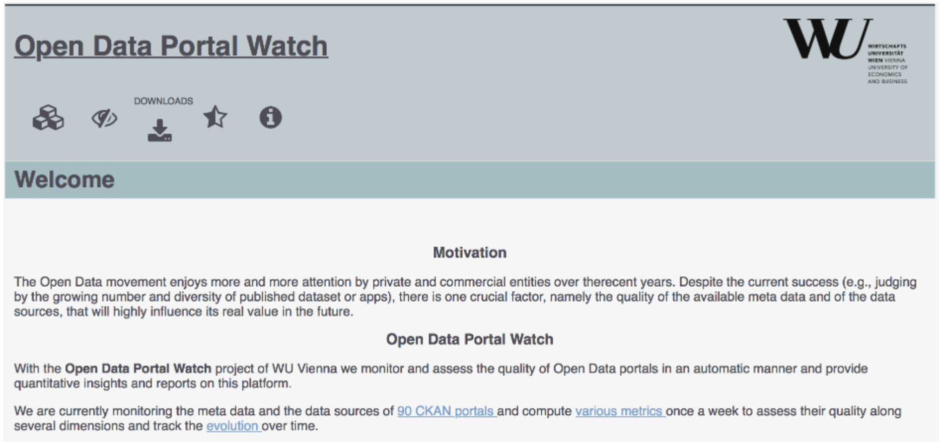
- machine-processable? …
  … **partially**

Open Data Portal Watch
... a first step.

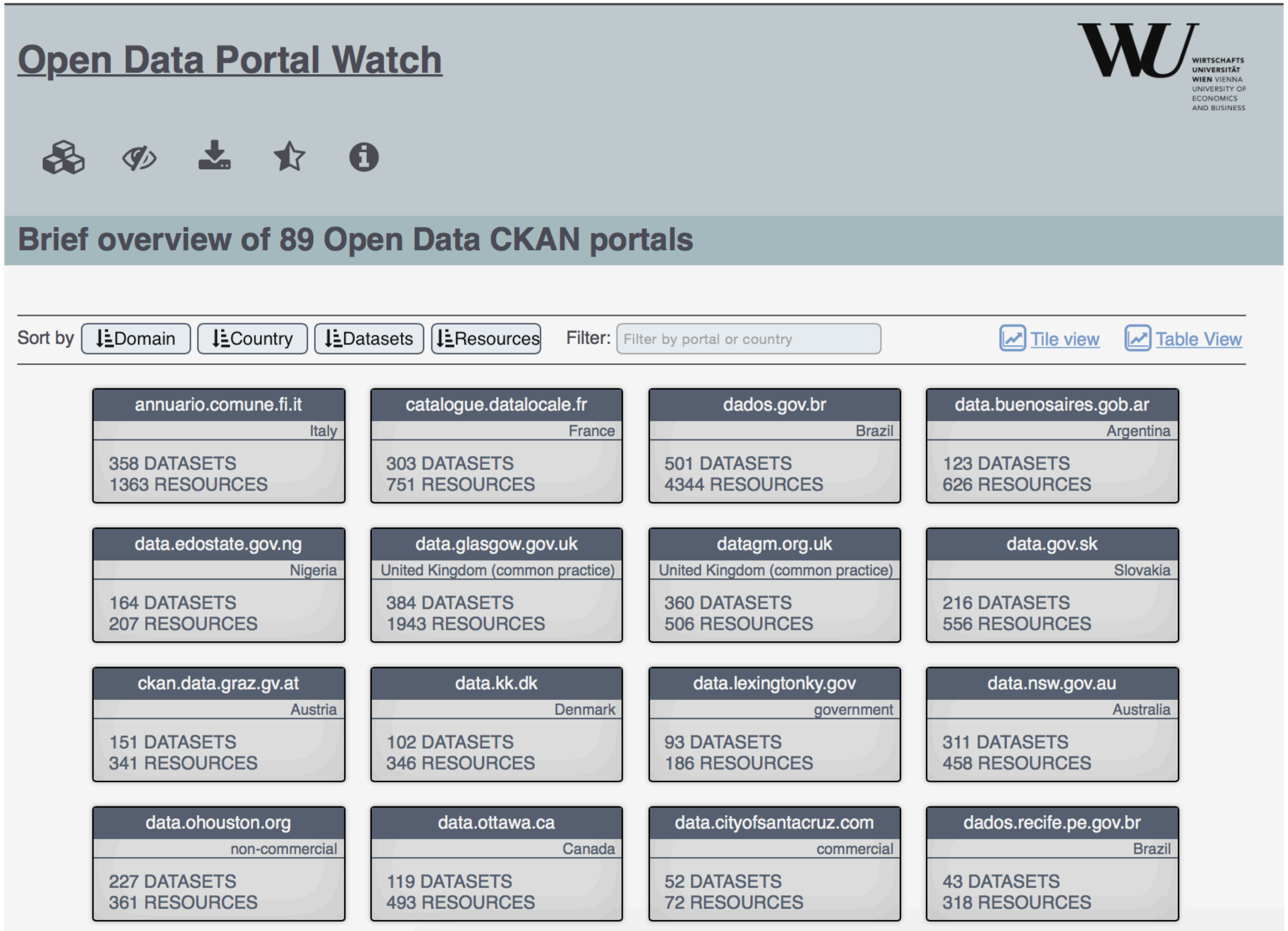
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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



Open Data Portal list

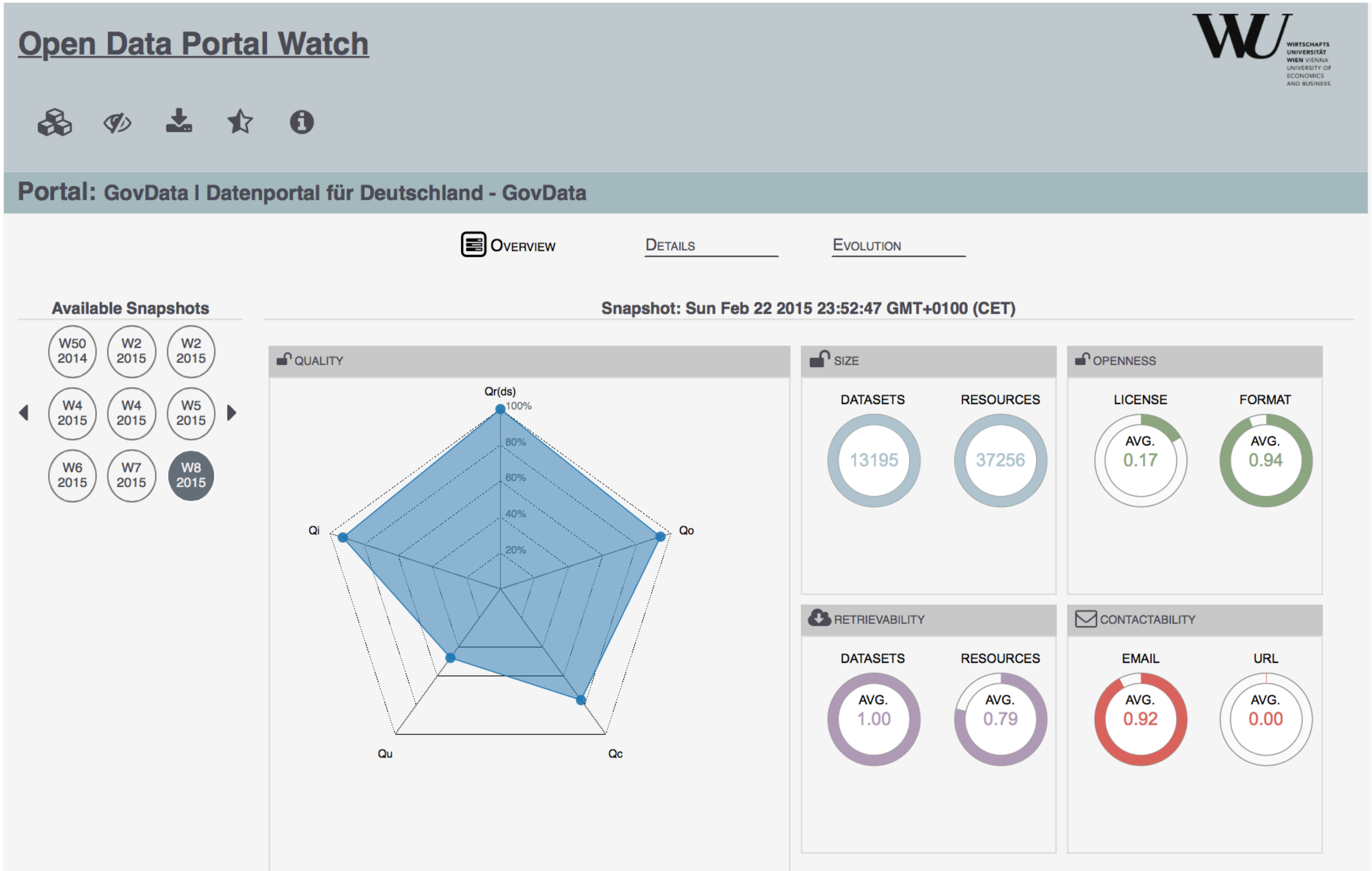


Quality Dimensions

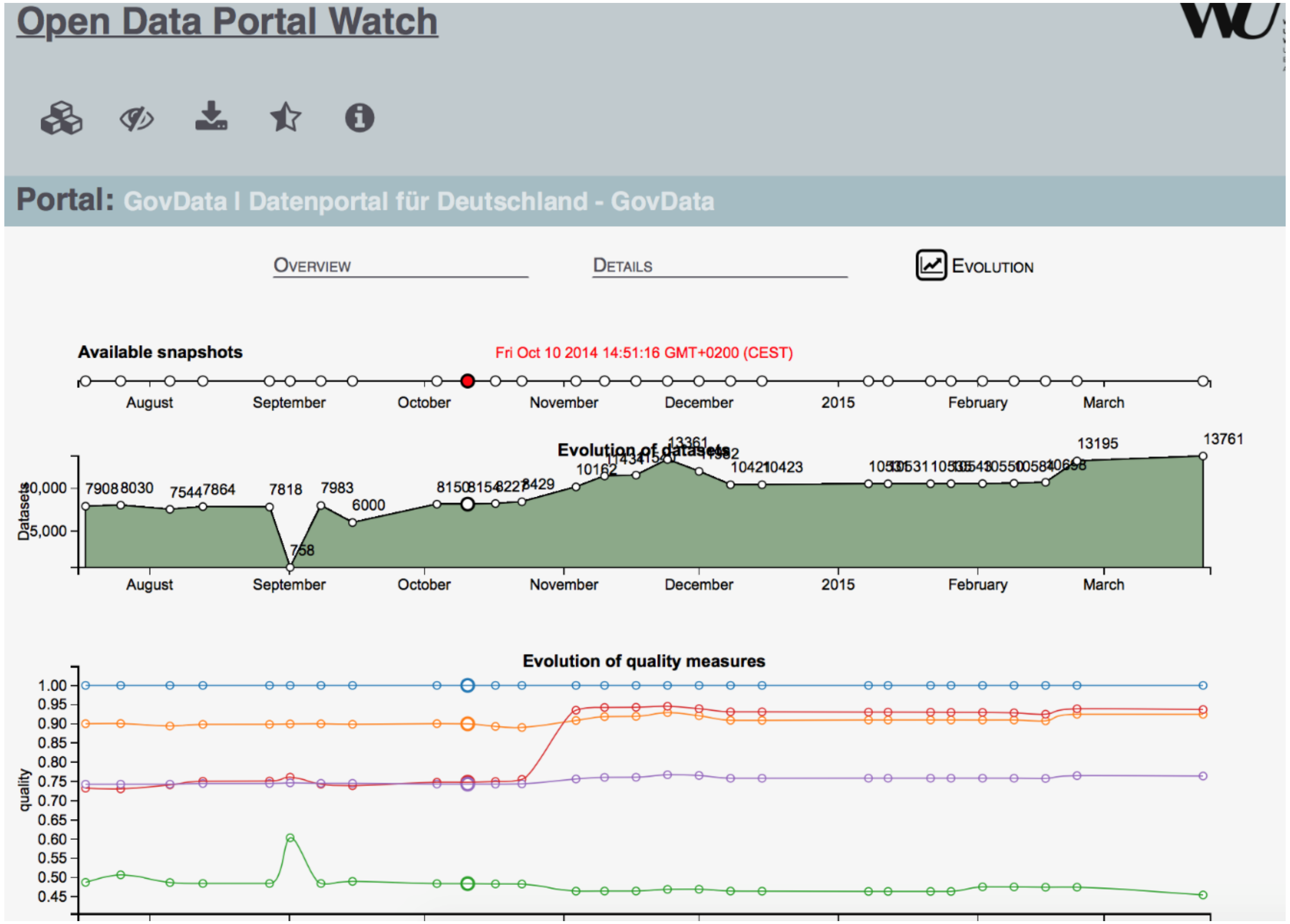
DIMENSION	DESCRIPTION
Retrievability	The extent to which meta data and resources can be retrieved.
Usage	The extent to which available meta data keys are used to describe a dataset.
Completeness	The extent to which the used meta data keys are non empty.
Accuracy	The extent to which certain meta data 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



ODP Evolution



ODP CHANGES

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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 10 entries

Search:

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

Open Data Portal Watch

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Portal Overview

Quality assessment & evolution of Open Data portals

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

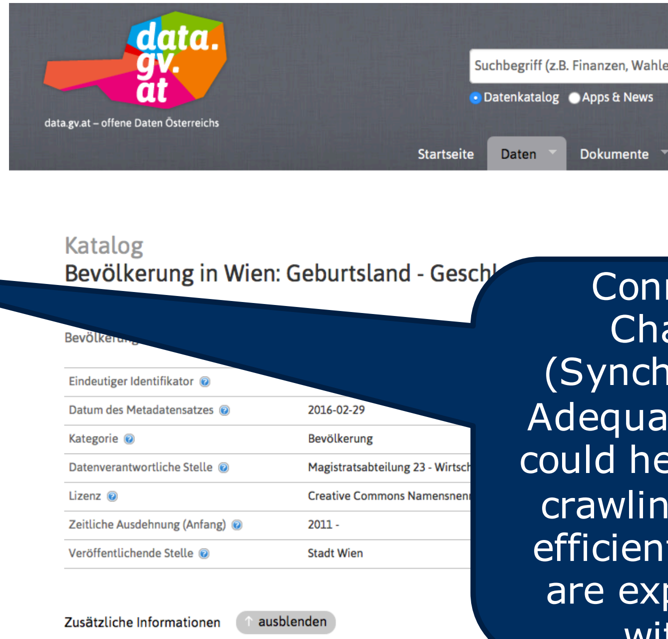
Historical vs. current-only data (monotonic changes vs. non-monotonic changes)

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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

Updated every 15min, only current data ←

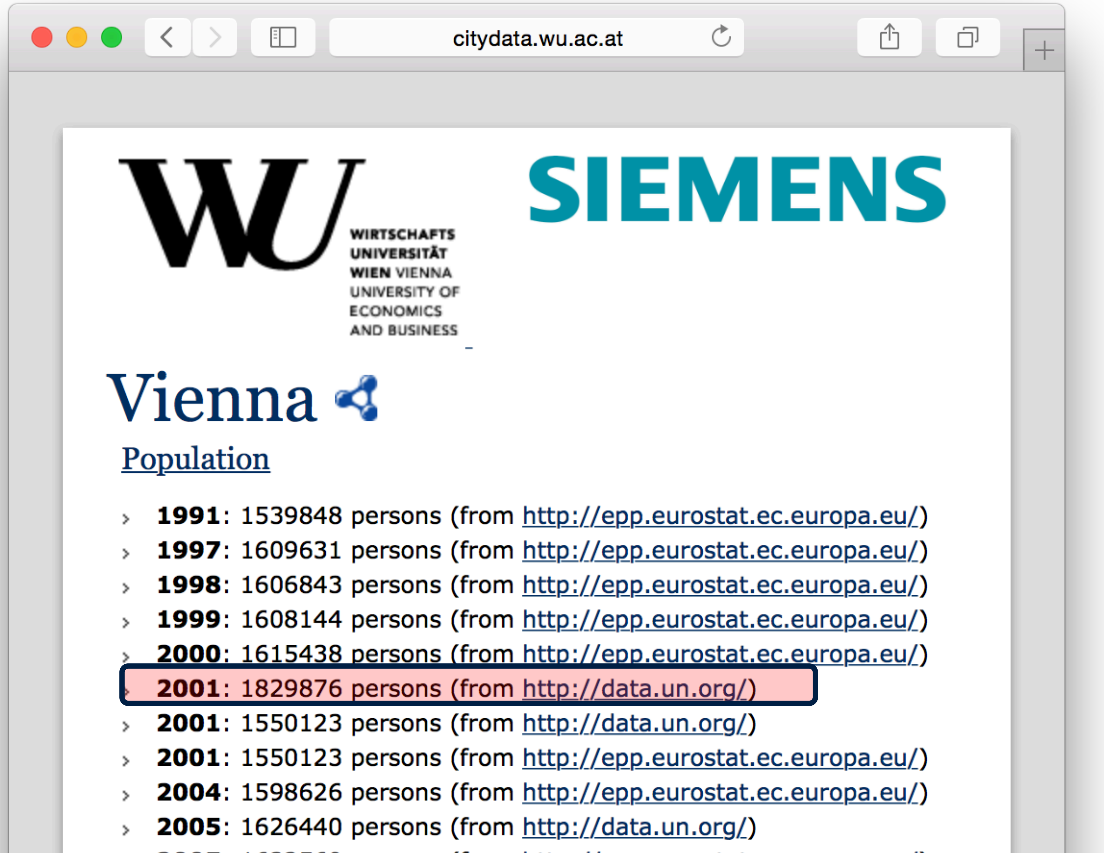
Updated annually, historical data since 2011 →



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…

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!

Student hoaxes world's media on Wikipedia

Phony quote appears in obituaries for French composer Maurice Jarre

By Shawn Pogatchnik
Associated Press
updated 5/12/2009 10:50:25 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.

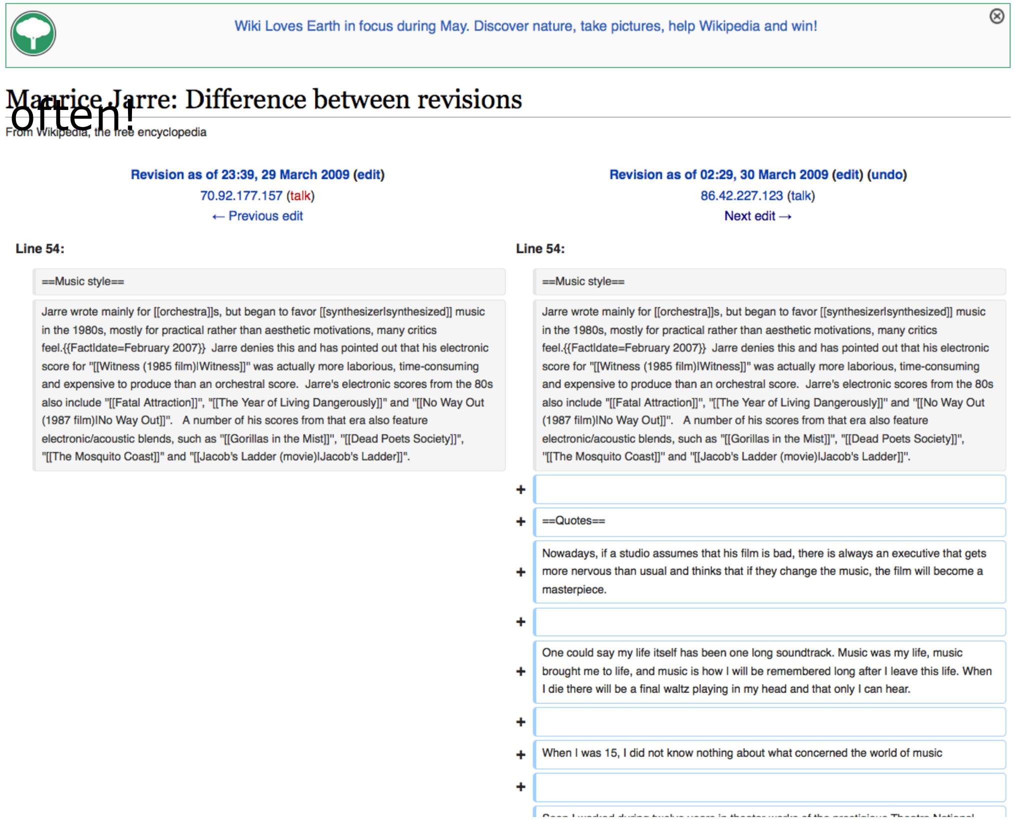
The sociology major's made-up quote — which he added to the Wikipedia page of Maurice Jarre hours after the French composer's death March 28 — flew straight on to dozens of U.S. blogs and newspaper Web sites in Britain, Australia and India.

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 first.

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.



Wiki Loves Earth in focus during May. Discover nature, take pictures, help Wikipedia and win!

Maurice Jarre: Difference between revisions

From Wikipedia, the free encyclopedia

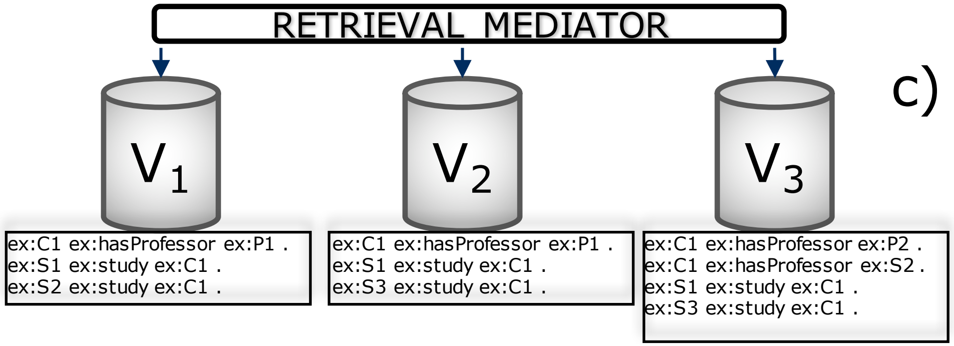
Revision as of 23:39, 29 March 2009 (edit) 70.92.177.157 (talk) ← Previous edit	Revision as of 02:29, 30 March 2009 (edit) (undo) 86.42.227.123 (talk) Next edit →
Line 54:	Line 54:
==Music style==	==Music style==
Jarre wrote mainly for [[orchestra]]s, but began to favor [[synthesizer|synthesized]] music in the 1980s, mostly for practical rather than aesthetic motivations, many critics feel.{{Fact|date=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]]".	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.{{Fact|date=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*

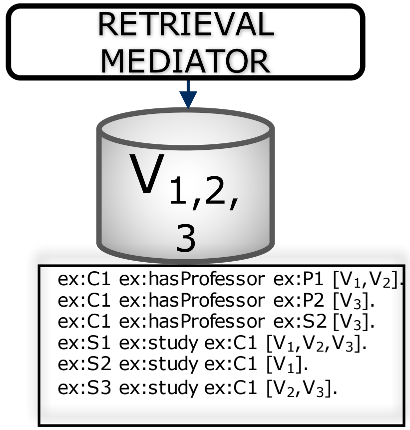
Now, how can we efficiently archive and perform time-based retrieval queries of a dataset?

RDF Archiving. Archiving policies

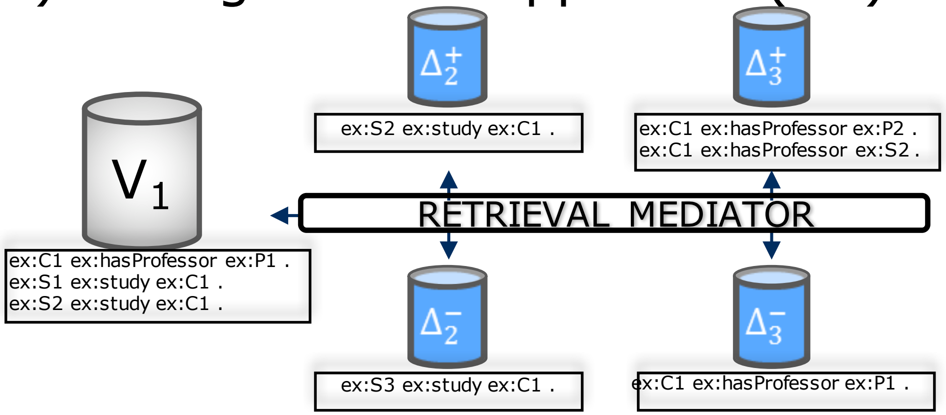
a) Independent Copies/Snapshots (IC)



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(Vi,Vj): true if delta!=null
    - Join(Q1,V_i, Q2,V_j)
    - Change(Q): Returns versions in which Diff(Q, V_i, V_{i-1}) $!=\varnothing$

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, re... querying with annotated Semantic Web data. *Journ... Semantics (JWS), 12:72--95, March 2012.*

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

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

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

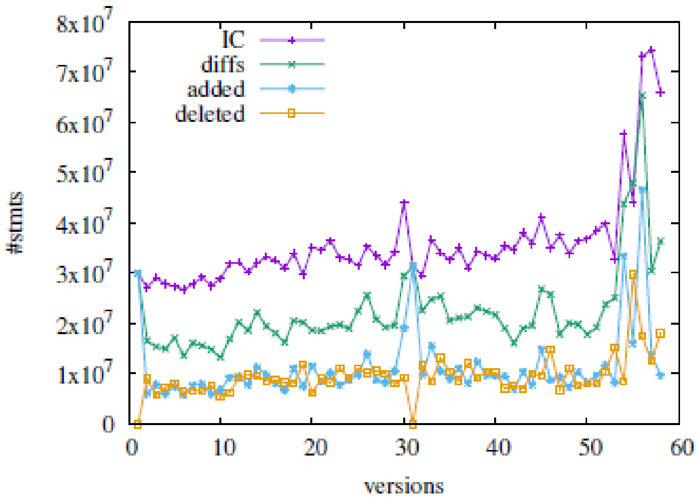
BEAR: Benchmarking the Efficiency of RDF Archiving

- 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)**

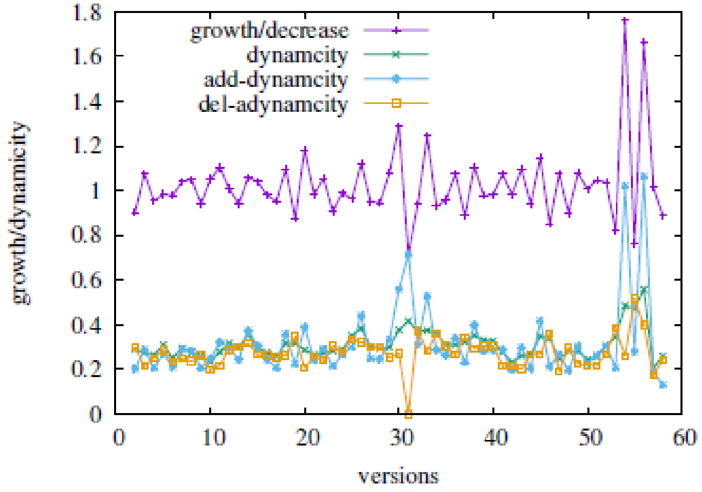
BEAR: Benchmarking the Efficiency of RDF Archiving

- Corpus

versions	$\|V_0\|$	$\|V_{57}\|$	$\overline{growth}$	$\overline{\delta}$	$\overline{\delta^-}$	$\overline{\delta^+}$	$\mathcal{C}_\mathcal{A}$	$\mathcal{O}_\mathcal{A}$
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!

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Archiving and **querying evolving** semantic Web data

Objective	Research Question
Representation	❖ minimize the redundant information ❖ respect the original modeling and provenance information
Query language	❖ capture the expressiveness of emerging retrieval demands in archiving ❖ our base operations are meant to be an **extensible** starting point ❖ design a query language satisfying these requirements for evolving interlinked data
Indexing	❖ index archives at large scale (and keeping up with evolution rate – streamingvs. archiving) to process the queries efficiently
Query optimization	❖ optimizing query resolution plans for archives ❖ enabling the integration of other sources (federated infrastructure) ❖ Query rewriting for querying archives of structured non-RDF sources? Open Data!
Application	❖ Is there a actual and urgent need in the community? ❖ We believe yes, but where’s the killer-app?

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
polleres.net

Big (Semantic) Data
Versions
Evolving Data
Streaming
Compression

Backup

- 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] }
```

Backup

- 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 )
}
```

Backup

- 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 }
```

Backup

- 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]} }
```

Backup

- 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 ) }
```