WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

LOD is all about evolution

Querying and Managing evolving Linked Open Data



Javier D. Fernández

Drift-a-LOD'17

11TH SEPTEMBER 2017

Special thanks to Axel Polleres for his input

About me:







Pablo de la Fuente Miguel A. Martínez-Prieto Claudio Gutiérrez

Óscar Corcho

Maurizio Lenzerini

Axel Polleres

since 2015 @WU, Inst. for Information Business

Research interest: Semantic Web, Open Data, Big (Semantic) Data Management, Databases, Data Compression, Privacy and Security

https://www.wu.ac.at/en/infobiz/team/fernandez/

General agenda

- Monitoring Evolution and Archiving
- Archiving the Web of Data
- Representing and querying evolving semantic data
- Open Data evolution







VIENNA

Why evolution matters (Creationists: please ignore this slide...)



HUB VIENNA





Monitoring evolution is relevant



Evolution matters



HUB VIENNA

Changes tell us "something"

- Uncertain information
- Validity of the information

Donald Trump: Difference between revisions

		24	0
	no	20	h'
_			υ.

File:Donald	Trump by Gage	Skidmore 3.jpg left thumb	Trump speaking at the	2015 [[Conservative Politi	ical Action Confer	rence]] (CPAC) in [[I	National Harbor,
/laryland]]]]							

In April 2011, Trump questioned President [[Barack Obama]]'s [[Barack Obama citizenship conspiracy theories|proof of citizenship]], --cref name=WashPost4711> {{cite news |title=Trump goes after Obama on US citizenship, says citizenship questions remain unanswered |date=April 7, 2011 |agency=[[Associated Press]] |work=[[The Washington Post]]}}-/ref> alleging that, "His grandmother in Kenya said he was born in Kenya and she was there and witnessed the birth, okay?"<ref name=Farley>Farley,Robert. [http://www.politifact.com/truth-o-meter/statements/2011/apr/07/donald-trump/donald-trump-says-president-obamas-grandmothercau/ "Donald Trump Says President Obama's grandmother caught on tape saying she witnessed his birth in Kenya"], "[[Politifact]]" (April 7, 2011): "What we heard was a very rough translation in which an elderly woman agreed to the leading question that Obama was born in Kenya and that she was present. But it was immediately and clearly corrected -- repeatedly."</ref ></ref name=McGraw7Apr>{{cite news |url=http://www.today.com/id/42469703/ns/today-today_people/ title=Trump: I have 'real doubts' Obama was born in U.S. |work=Today |first=Seamus |last=McGraw |date=April 7, 2011 |accessdate=September 13, 2015}|</ref> Trump's claim derived from an incomplete transcript of a telephone interview with Obama's grandmother, provided by Obama's opponents.<ref name=Farley /><ref name=Factcheck2011Apr /> Trump also questioned whether Obama had good enough grades to warrant entry to Harvard Law School.<ref>Madison, Lisa (April 26 2011).[http://www.cbsnews.com/news/trump-how-did-obama-get-into-the-ivy-league/ Trump: How did Obama get into the lvy League?]. CBS News.</ref> 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} 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

		View logs for this page	je					
		- Browse history -						
		From year (and e	artier): 2016	From month (and earlier)	all • Tag filter:		Show	
		For any version list External tools: Revi	ed below, click on its da sion history statistics	ate to view it. For more help, see H · Revision history search@ · Edite	telp:Page history and He s by userg? - Number of w	p:Edit summary. atchers@ • Page view	statistics g?	
		(cur) = difference fr (newest oldest) Vi Compare selected	om current version, (pr ew (newer 50 older 50 revisions	ev) = difference from preceding ve 0) (20 50 100 250 500)	rsion, m = minor edit, →	= section edit, ← = au	tomatic edit summary	
		• (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 byles) (-27)	(per talk page, unclu	ustering huge clusters of hidden footnotes)	
		• (cur prev) 🗎	00:27, 25 May 2016	BarrelProof (talk contribs) (2	82,961 bytes) (+57) (-	Religious views: Youn	ger than what?)	
		 (cur prev) 721935275)) 	00:06, 25 May 2016	AnomieBOT (talk contribs) (282,904 bytes) (+549)	(Rescuing orphaned re	fs ("msnbc1" from rev 721935275; "CBC_August29_201	15" from rev
		• (cur prev) 🔍	23:36, 24 May 2016	Anythingyouwant (talk contribs)	(282,355 bytes) (+1)	. (→Controversial imm	nigration policies: spelling, merge two pargraphs)	
		 (cur prev) main article on 	23:35, 24 May 2016 political positions, para	Anythingyouwant (talk contribs) phrase accordingly)	(282,354 bytes) (-661)(→Controversial in	nmigration policies: swap hidden cluster of footnotes for	two footnotes from
	_	 (cur prev) section) 	23:19, 24 May 2016	Anythingyouwant (talk contribs)	(283,015 bytes) (-2) .	. (moce reference out	of huge hidden cluster of footnotes, to appropriate spot i	n 2016 campaign
		• (cur prev) 🔍	23:13, 24 May 2016	Anythingyouwant (talk contribs)	(283,017 bytes) (-1,0	26)(→Controversial	immigration policies: trim some footnotes from big hidde	on cluster of footnote)
1.1		• (cur prev) 🗎	23:07, 24 May 2016	Anythingyouwant (talk contribs)	(284,043 bytes) (-7) .	. (→Controversial imm	igration policies: move reference out of cluster)	
Line	246:	• (cur prev) 🔍	21:58, 24 May 2016	Anythingyouwant (talk contribs)	m (284,050 bytes) (0	(→Business career	putting period after quote mark per Wikipedia habits)	
		• (cur prev) 🔍	20:40, 24 May 2016	Anythingyouwant (talk contribs)	(284,050 bytes) (+66	5)(→Trump Tower:	add info from main article) (Tag: nowlkl added)	
		• (cur prev) 🔍	20:29, 24 May 2016	Anythingyouwant (talk contribs)	(283,385 bytes) (+92	5)(→Trump Tower:	add re. appeal)	
	mes.	 (cur prev) 	19:52, 24 May 2016	Anythingyouwant (talk contribs)	(282,460 bytes) (-30)	(→2010–present: fb	<pre>x pic)</pre>	
	II.	• (cur prev) 🔍	19:50, 24 May 2016	Anythingyouwant (talk contribs)	(282,490 bytes) (+10	l)(→Legal affairs: in	esert pic)	
	Ma	 (cur prev) 	19:39, 24 May 2016	Anythingyouwant (talk contribs)	m (282,383 bytes) (+	4)(→Taxes)		
1	1- 1	• (cur prev) 🔍	19:38, 24 May 2016	Anythingyouwant (talk contribs)	(282,379 bytes) (+85	(→Taxes: add re. d	lisclosures)	744-
	In A	• (cur prev)	19:35, 24 May 2016	Anythingyouwant (talk contribs)	(282,294 bytes) (+62	(→Taxes: pipe link	re. tax returns)	111>
	{{cite	news title=1rui	mp goes after OI	pama on US citizenship,	says citizenship q	uestions remain	unanswered (date=April 7, 2011 (agency	=[[Associated Press]]
	work	=[[The Washing	ton Post]]}} <td>> alleging that, "His grand</td> <td>dmother in Kenya</td> <td>said he was born</td> <td>i in Kenya and she was there and witnes</td> <td>sed the birth, okay?"<ref< td=""></ref<></td>	> alleging that, "His grand	dmother in Kenya	said he was born	i in Kenya and she was there and witnes	sed the birth, okay?" <ref< td=""></ref<>
	name	=Farley>Farley	, Robert. [http://	www.politifact.com/truth-	o-meter/statement	s/2011/apr/07/do	onald-trump/donald-trump-says-presiden	it-obamas-grandmother-
	cau/"	Donald Trump	Says President (Obama's grandmother ca	aught on tape sayi	ng she witnesse	d his birth in Kenya"], "[[Politifact]]" (April	7, 2011): "What we heard
	was a	very rough tra	nslation in which	an elderly woman agree	d to the leading qu	estion that Obar	ma was born in Kenya and that she was	present. But it was
	imme	diately and clea	rly corrected i	epeatedly." <ref na<="" td=""><td>me=McGraw7Apr</td><td><pre>{{cite news lurl=</pre></td><td>http://www.today.com/id/42469703/ns/to</td><td>day-today people/</td></ref>	me=McGraw7Apr	<pre>{{cite news lurl=</pre>	http://www.today.com/id/42469703/ns/to	day-today people/
	Ititle=	Trump: I have 'r	eal doubts' Oba	ma was born in U.S. Iwo	k=Today lfirst=Se	amue llast=McG	raw Idate=April 7, 2011 Jaccessdate=Ser	ntember 13, 2015\}
	Trues	ala alaim dariya	d from on incom	elate transportet of a talan	hana interview wit	h Ohomolo orong	mather provided by Obema's encount	a staf name-Earley /s staf
	moning	5 Claim derive		piete transcript of a telep		i Obalila s grand	aniourier, provided by Obarna's opponent	s. ter hame-raney iz ter
	name	=Factcheck20	rrapr /> rrump a	iso questioned whether	Ubama nad good e	enough grades to	5 warrant entry to Harvard Law School.<	rer>madison, Lisa (April 26,
	2011)	.[http://www.cbs	snews.com/news	s/trump-how-did-obama-	get-into-the-ivy-lea	ague/ Trump: Ho	w did Obama get into the Ivy League?]. (CBS News. On April
	25, 20)11, Trump calle	ed for Obama to	end the citizenship issue	by releasing the l	ong form of his b	irth certificate. <ref>{{cite news</ref>	
	url=h	ttp://ac360.blog	s.cnn.com/2011.	/04/25/trump-claims-oba	ma-birth-certificate	-missing/ title=T	rump claims Obama birth certificate 'mis	sing' date=April 25, 2011
	publis	sher=CNN acc	essdate=May 14	4, 2011}} <ref>{{cite</ref>	news title=Birther	ism: Where It All	Began	
+	lurl=h	ttp://www.politic	o.com/news/sto	ries/0411/53563 Page3.	html Iwork=[[Politic	oll Idate=April 22	2. 2011 accessdate=April 25. 2011}} <td>f> Two days later. Obama</td>	f> Two days later. Obama
	made	a formal stater	nent in efforts by	the [[White House]] to p	it the matter to res	t with the release	e of the long form <ref></ref>	
	Intte /	Annual buffin eter	mont com/2011#	1/27/ohama hirth a artifi	ate r n 954040 k	tral Ohama Pirth	Certificate Released By White House (F	
	inttp:/	www.ndningtor	postcom/2011/		ate-1_11_034240.1	uni Obalfia bilui	Gerundate released by White House (F	no roj, _E rne Hullington
	Post]]	April 27, 2011	. Retrieved May	o, 2011. trump exp	pressed pride at h	s role in the cert	moate's release in a press conference fo	now-up, saying ne noped it

Donald Trump: Revision history

Preservation matters



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

UAUDACKINACHIVE www.nytimes.com ×

Explore more than 305 billion web pages saved over time







In Hiroshima, Obama Issues Call for 'Moral Revolution'	
 President Ohama, the first sitting U.S. president to visit Hiroshima, Appan, said the bombing of the sity showed that "manining possessed the means to destroy itself." 	
After laying a weath at a memorial, Mr. Obama	12 New Books We're Reading This Summer (and 6 Not So New)



The Opinion Passes

When the Barbarous Brits First Quit Europe ³y TON HOLLARD Sizteen handred years before the Dimpen

rial: Hillary Cli

On Ed: Obama's Palet

Deliverative Why Do So Many Studies Fail to Replicate? Pratrat Rarts Decauer it is hard to recrute the exact conditions of the original research. - The Repl Trange Effect for When the detain

6

Time-based access matters



VIENNA

The Memento protocol



RFC 7089

Follow your nose (HTTP content negotiation with datetime)





Challenges (Web archives)



VIENNA

- Poor granularity ("some" snapshots)
- Aggregated data, only, rather than raw data access
 - (e.g. in Google trends)
- 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?

WIRTSCHAFTS UNIVERSITÄT UNIVERSITY OF ECONOMICS AND BUSINESS

Arching the Web of Data

MARCHINE.





Linked Data is evolving





One of the first (and last?) LOD archives: The Dynamic Linked Data Observatory (evolving Linked Data since 2012)



C Swse.deri.org/dyldo/ 15 🗘 🔩 🔶 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 Datasets collect frequent, continuous snapshots of a subset of the Web of Data that is Team & Contact interesting for further study and experimentation, with an aim to capture raw Tobias Käfer (KIT, Karlsruhe) data about the dynamics of Linked Data. The resulting corpora will be made Publications Jürgen Umbrich (DERI, Galway) openly and continuously available to the Linked Data research community. Aidan Hogan (DERI, Galway) Patrick O'Byrne (NUIG, DERI, Galway, About Datasets Ireland) Ahmed Abdelrahman (NUIG, DERI, Results Good news! We started with the weekly crawls. Galway, Ireland) Google group 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.

⊢	→ C	swse.deri.org/dyldo/data
	Parent Di	irectory
	2012-05-	06/ 08-Dec-2012 12:33
	2012-05-	13/ 08-Dec-2012 12:40
	2012-05-	<u>21/</u> 08-Dec-2012 11:57
	2012-05-	<u>27/</u> 08-Dec-2012 11:57
	2012-06-	03/ 05-Dec-2012 17:25
	2012-06-	10/ 08-Dec-2012 11:58
	2012-06-	<u>17/</u> 08-Dec-2012 11:59
	2012-06-	<u>24/</u> 08-Dec-2012 11:59
	2012-07-	01/ 05-Dec-2012 17:25
	2012-07-	08-Dec-2012 12:00
	2012-07-	16/ 08-Dec-2012 12:00
	2012-07-	<u>22/</u> 08-Dec-2012 12:01
	2012-07-	<u>29/</u> 08-Dec-2012 12:01
	2012-08-	05/ 04-Dec-2012 10:54

Weekly dumps of crawl snapshots...

Granularity? Queries? Crawl failures?

Linked Data Archives: The missing link in the RDF evolution



VIENNA



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

RDF Archiving. Example







Research challenges on evolving structured interlinked data



- How can we represent archives of continuously evolving linked datasets?
- 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?

... in the last few years:



VIENNA





Representing and querying evolving semantic data



COMPLEXITY

SCIENCE HUB

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

The cold-start problem



How we can get archives of RDF data

© 2014 BOANET.AT

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



 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...



- Retrieve historical versions of a DBpedia resource
 - What was the version of "Donald Trump" on dd/mm/yyyy?
- Re-apply DBpedia mappings on the Wikipedia revision history



DBpedia Wayback Machine



- How can one represent revisions while respecting DBpedia?
 - a) quads \rightarrow <dbpediaSubject> <pred> <obj> <Revision> .
 - b) proprietary triples \rightarrow <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



DBpedia Wayback Machine



- 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...



HUB VIENNA



However: Only one version per "irregular" dbpedia dump

Acknowledgments: Lyudmila Balakireva, Harihar Shankar, Ruben Verborgh

LOD Laundromat



HUB VIENNA

Lodlaundromat.org: a central repository of LD



- Problems?
 - Still you need to access/query 650K datasets
 - Of course the solution is not complete, but "a good approximation"

LOD-a-lot: LOD-a-lot: Low cost archiving of LOD



Kudos Javier D. Fernandez, Wouter Beek, Miguel A. Martínez-Prieto, Mario Arias, Ruben Verbogh



#Triples	#Subjects	#Predicates	#Objects	#Common SO	#Literals
28,362,198,927	3,214,347,198	1,168,932	$3,\!178,\!409,\!386$	1,298,808,567	1,302,285,394

- Disk size:
 - HDT: 304 GB
 - HDT-FoQ (additional indexes): 133 GB



- Memory footprint (to query):
 - 15.7 GB of RAM (3% of the size)
 - 144 seconds loading time
 - 8 cores (2.6 GHz), RAM 32 GB, SATA HDD on Ubuntu 14.04.5 LTS
- LDF page resolution in milliseconds.

(LOD-a-lot creation took 64 h & 170GB RAM. HDT-FoQ took 8 h & 250GB RAM)

LOD-a-lot

http://purl.org/HDT/lod-a-lot





https://datahub.io/dataset/lod-a-lot

LOD-a-lot (some use cases)



HUB

VIENNA

Archiving

27

- We plan to have quarterly releases
- Query resolution at Web scale
- Evaluation and Benchmarking
 - No excuse ☺
- RDF metrics and analytics



Room 6+7 Session 2.6 Data Science

Querying and Visualization

Laura Hollink

LOD-a-lot: A Single-File Enabler for Data Science (Wouter Beek, Javier D. Fernández and Ruben Verborgh) *



umec

UNIVERSIDAD DE CHILE

h

d

28

ACKs LOD-a-lot





DataWebResearch

http://dataweb.infor.uva.es

VU UNIVERSITY AMSTERDAM

Knowledge Representation and Reasoning



IT Systems Engineering | Universität Potsdam

Insight











The archiving problem

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS



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

© 2014 BOANET.AT

RDF Archiving. Archiving policies



a) Independent Copies/Snapshots (IC)



BEAR



HUB VIENNA





OVERVIEW BEAR-A BEAR-B BEAR-C BENCHMARK RESULTS PUBLICATIONS CONTACT

BEAR

BEnchmark of RDF ARchives



1. Overview

There is an emerging demand on efficiently archiving and (temporal) querying different versions of evolving semantic Web data. As novel archiving systems are starting to address this challenge, foundations/standards for benchmarking RDF archives are needed to evaluate its storage space efficiency and the performance of different retrieval operations.

To this end, we have developed a BEnchmark of RDF ARchives (BEAR), a test suite composed of three real-world datasets together with queries with varying complexity, covering a broad range of archiving use cases.

BEAR-B

The DATA:

BEAR comprises three main datasets, namely BEAR-A, BEAR-B, and BEAR-C, each having different characteristics.

Dynamic Linked Data

BEAR-A is composed of 58 weekly snapshots from the Dynamic Linked Data Observatory. BEAR-A provides triple pattern queries to test atomic queries such as Materialization, Diff, Version, etc.

DBpedia Live

The BEAR-B dataset has been compiled from DBpedia Live changesets over the course of three months and contains the 100 most volatile resources along with their updates and real-world triple pattern queries from user logs.

Open Data portals BEAR-C

QADLOD PROJECT

BEAR-C used the Open Data Portal Watch project to take the datasets descriptions of the European Open Data portal for 32 weeks. With the help of Open Data experts, we created 10 complex queries that retrieve different information from datasets and files.



- 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





SCIENCE HUB VIENNA



- 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 gueries: Materialize, get Version...
 - Initial evaluation on archiving policies











VIENNA

Benchmarking: 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

$$\delta_{i,j} = \frac{|\Delta_{i,j}^+ \cup \Delta_{i,j}^-|}{|V_i \cup V_j|}$$
$$growth(V_i, V_j) = \frac{|V_j|}{|V_i|}.$$

Data static core

- $\mathcal{C}_{\mathcal{A}} = \{(s, p, o) | \forall i \in \mathcal{N}, (s, p, o) : [i] \in \mathcal{A}\}$
- Total triples (version-oblivious)
- ♀ Others
 - ☑ RDF vocabulary
 - ♀ Per version / evolution

$$\mathcal{O}_{\mathcal{A}} = \{(s, p, o) | \exists i \in \mathcal{N}, (s, p, o) : [i] \in \mathcal{A}\}$$



SCIENCE HUB VIENNA

Benchmarking: Define the queries

- 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



HUB VIENNA

- Design of benchmark queries
 - ☑ Archive-driven Cardinality + Selectivity (disregard versions)
 - ▷ Version-driven Cardinality + Selectivity + dynamicity

$$dyn(Q, V_i, V_j) = \frac{|(\Omega_i \setminus \Omega_j) \cup (\Omega_j \setminus \Omega_i)|}{|\Omega_i \cup \Omega_j|}$$

- Basic temporal retrieval features of queries
 - \heartsuit Mat (Q, V_i): version materialization
 - \heartsuit Diff (Q, V_i,V_j): delta materialization
 - ♀ Version(Q): results of Q annotated with the version
 - □ Join(Q1, V_i , Q2, V_j)
 - ▷ Change(Q): Returns versions in which Diff(Q, V_i, V_{i-1}) $!=\emptyset$



- 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)



- 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)
- SELECT * WHERE { {Q : [v1] } {Q : [v2] }

Change(Q)



VIENNA

Benchmarking: Define the queries

- Instantiation of archive queries in AnQL [1]
 - Antoine Zimmermann, Nuno Lopes, Axel Polleres, and Un general framework for representing, reasoning and query annotated Semantic Web data. Journal of Web Semantics 95, March 2012.

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

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

SELECT ?V1 ?V2 WHERE
{ {{Q :?V1 } MINUS {Q :?V2}} UNION
 {{Q :?V2 } MINUS {Q :?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







VIENNA

https://aic.ai.wu.ac.at/qadlod/bear.html

BEAR: Benchmarking the Efficiency of RDF Archiving



VIENNA

- Queries and systems
 - We implemented and evaluate archiving systems on Jena-TDB and HDT, based on IC, CB and TB policies.
 - Serve as an initial baseline to compare archiving systems
 - More info: <u>https://aic.ai.wu.ac.at/qadlod/bear.html</u>

BEAR datasets



HUB VIENNA

Dynamic Linked Data

BEAR-A is composed of 58 weekly snapshots from the Dynamic Linked Data Observatory. BEAR-A provides triple pattern queries to test atomic queries such as Materialization, Diff, Version. etc.

DBpedia Live

logs.

The BEAR-B dataset has been compiled from DBpedia Live changesets over the course of three months and contains the 100 most volatile resources along with their updates and real-world triple pattern queries from user

BEAR-B

Open Data portals

BEAR-C

BEAR-C used the Open Data Portal Watch project to take the datasets descriptions of the European Open Data portal for 32 weeks. With the help of Open Data experts, we created 10 complex gueries that retrieve different information from datasets and files.

Open Data Portal Watch

Quality Assessment and Monitoring of 260 Open Data Portals

i ODPW	■ System	III Portals	API	Quality Measures	Q License Search	Q SPARQL
					Open D	ata Portal Wa
	The Ope Web cat	n Data Portal alogues and pe	Watch fram rform a qual	ework is a scalable quality lity assessment of the met	assessment and evolut adata along 6 dimensio	ion monitoring framework ns and 19 metrics.

	Quality Metrics 6 Dimensions 19 Metrics	Portals List 261 registered Portals
	More introamtion about our quality dimensions and metrics.	Basic information such as the country, softwa the number of datasets.
Q3	<pre>PREFIX dcat: <http: dcc<br="" ns="" www.w3.org="">PREFIX dc: <http: 2006="" <br="" www.w3.org="">PREFIX vcard: <http: 2006="" <br="" www.w3.org="">PREFIX rdf: <http: 02<br="" 1909="" www.w3.org="">{ ?dataset rdf:type dcat:Dataset . ?dataset dcat:contactPoint ?contact ?contact vcard:fn ?name. OPTIONAL{ ?contact vcard:hasEmail ?email } }</http:></http:></http:></http:></pre>	t#> syl1.1/> JPW API /vcard/ns#> /22-rdf-syntax-ns#> :t . ↓ .

The Dynamic Linked Data Observatory

Datasets **Publications** About Results 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

Good news! We started with the weekly crawls.





COMPLEXITY

SCIENCE HUB VIENNA

RDF Archiving. Archiving policies







RDF Archiving. Archiving policies

COMPLEXITY SCIENCE HUB VIENNA

a) Independent Copies/Snapshots (IC)



Time-based access. Queries





Time-based access. Queries





Time-based access. Queries



HDT-IC

50

HB₁₆

diff(?,?,o ; version0 ; version t) HB₄ HB8 CB IC 48 GB 28 GB 34 GB 31 GB 29 GB Jena-IC HDT-IC Jena-TB ·····* Jena-CB HDT-CB HDT-HB₄ — HDT-HB₁₆*.... 10000 HDT-HB₈ ----×----HDT-CB 10 query time in ms (logscale) 1000 query time in ms (logscale) 100 1 10 0.1 1 0.1 0.01 10 20 30 40 0.01 0 10 50 20 30 40 0 diff(q,0,t)diff(q,0,t)

Hybrid approach

Self-Indexing RDF Archives: v-RDFCSA

COMPLEXITY SCIENCE HUB VIEWA

- RDFCSA: RDF index based on a Compressed Suffix Array
- v-RDFCSA[2] is designed as a lightweight TB approach
- Version information encoding
 - Any triple can be identified by the position of its subject within SA
 - Let be N the number of different versions and n the set of version-oblivious triples
 - Two alternative encoding strategies
 - tpv: N bitsequences, each position i encodes if the triple i appears in the version
 - vpt: n bitsequences, each position i encodes if the version i includes the triple



Performs more than **one order of magnitude** faster than Jena-TDB

[2] Ana Cerdeira-Pena, Antonio Fariña, Javier D. Fernández, and Miguel A. Martínez-Prieto. Self-Indexing RDF Archives. Data Compression Conference (DCC), 2016.





Open Data evolution





OPEN DATA PORTAL WATCH ... a first step.

JNIVERSITY C ECONOMICS AND BUSINESS COMPLEXITY SCIENCE HUB VIENNA

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

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



ECDA: Evolving CSV Data Analyzer



UNIVERSITY OF

AND BUSINESS COMPLEXITY SCIENCE HUB VIENNA

ECDA: Evolving CSV Data Analyzer



- Analysis of 726 datasets
 - Mean of 18 versions per file
 - Mean of 430 rows and 5.4 columns
- Increasing nature (x1.85 number of rows)
 - Small value modifications (0.85 jaccard)
 - Mostly string types (80% of 8-25 characters)

Type of changes	Number of sets
STATIC: No changes in number of rows or	10 (1.4%)
columns	
INCREASING: No deletions of rows or columns,	71 (9.8%)
just additions.	
DECREASING: No deletions of rows or columns,	2(0.3%)
just deletions.	
SAME SCHEMA: Only changes (additions or	117 (16.12%)
deletions) of rows	
SAME ROWS: Only changes (additions or dele-	9(1.24%)
tions) of columns	
IRREGULAR: Changes (additions or deletions)	600 (82.64%)
of rows and columns	

Variable	Information
Number of versions	mean = 17.83, std = 6.18
Original rows	mean = 431.23, std = 1596.23
Original columns	mean = 5.40 , std = 8.48
Added rows	mean= 1.85 , std= 21.37
Deleted rows	mean = 0.79, std = 2.02
Modified rows	mean = 0.09 , std = 0.80
Added columns	mean = 0.23 , std = 0.28
Deleted columns	mean = 0.22 , std = 0.29
Modified columns	mean = 0, std = 0
Changed cells	mean = 1663.96 , std = 29564.60

Data domain	Number of versions
www.landesdatenbank.nrw.de	493
ourairports.com	83
data.gov.au	23
Other domains	127

ECDA: Evolving CSV Data Analyzer

- Analysis of 726 datasets
 - Entities (Babelfy)
 - On average there are around 0.07 entities per cell
 - Entities are static in the header (a mean of 3)
 - 1/3 of the entities change across time
 - Number of entities slightly decrease in time

Variable	Information
Number of cells	mean = 2597.09, std = 9890.53
Original entities	mean= 180.86 , std= 184.41
Actual entities	mean= 149.40 , std= 202.90
New entities	mean= 86.65 , std= 170.97
Entities maintained	mean= 62.75 , std= 131.23
Original header entities	mean= 2.68 , std= 11.09
Actual header entities	mean = 2.74, std = 11.14
Header entities maintained	mean= 2.64 , std= 11.07



ECONOMICS AND BUSINESS

COMPLEXITY SCIENCE HUB VIENNA

Finally, many open questions remain still!



VIENNA

Archiving and querying evolving semantic Web data

Objective	Research Question
Representation of archives	 minimize the redundant information respect the original modeling and provenance information (e.g. LOD-a-lot)
Query language	 design a query language satisfying these requirements for evolving interlinked data our BEAR operations are meant to be an extensible starting point
Indexing	 index archives at large scale keep up with evolution rate (streaming vs. archiving) to process the queries efficiently
Analysis/Optimization	 use evolution patterns to optimize representations and queries Querying archives of structures and non-structured sources? E.g. Open Data!
Application	 LOD-a-lot is a good examples but modularity can be improved Any low-cost but functional archiving at LOD scale can be a major milestone for the community





Thank you!

INTERNATIONAL

The measure of intelligence is the ability to change" Albert Einstein