

ALEXANDRIA

Temporal Retrieval, Exploration and Analytics in Web Archives

Wolfgang Nejdl

L3S Research Center Hannover, Germany





Web Science @ L3S

Computer Science and interdisciplinary research on all aspects of the Web

- Internet: Communication and Networks
- Information: Accessing information and knowledge on and through the Web
- Community: Supporting communities and groups on the Web, for research, education, production and entertainment
- Society: Requirements (technological, social, legal) for the Web

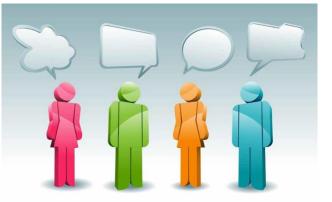


Real-time data processing for finance predictions



Cross-media analysis and interpretation

ForgetIT



LivingKnowledge:
Diversity, opinion and
bias on the Web



ForgetIT: Concise Preservation via Managed Forgetting

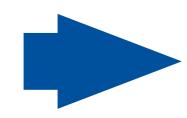


CUbRIK: Searching by computers and humans



Privacy, Property and Internet Governance

Selected projects





Web Science - Investigatir

News Sport Comment Culture Business Money Life & style

Conservative party deletes archive of

Decade's worth of records is erased, including PM's speech

praising internet for making more information available

News Politics Conservatives

speeches from internet

theguardian

Are we loosing the past of the web?



EgyTweets

Gun running from

RT @RSSEGYPTcom: #Egypt #Jan25 كام عسر 10 لـ 10 سنوات ضد البلطجة واسرقة بالإكراه وخرق حظر التجوال http://dlvr.it/KCHz8 #tahrir

about 1 hour ago via EgyTweets

, التحريلات RT @RSSEGYPTcom: #Egypt #Jan25 http://dlvr.it/KCHym #tahrir الدستورية

about 1 hour ago via EgyTweets

علجل. حل انحلا كرة الغم RT @RSSEGYPTcom: #Egypt #Jan25 http://dlvr.it/KCHyJ #tahrin

about 1 hour ago via EgyTweets

RT @SoulfunkLA: comes rough, tough like an elephant tusk. Ya head rush, fly like Egyptian musk...

about 1 hour ago via EgyTweets

Attack on Copts

RT @flavianoflavian: DailyNewsEgypt: Lgypt-s arms from Sudan http://tinyurl.com/4rxo8dx #fb

about 1 hour ago via EgyTweets

RT @AlMasryAlYoum_E: Armed forces attacked sleeping #Copts, say Coptic leaders http://ow.ly/4e0V4 #Atfeeh #Egypt

about 1 hour ago via EgyTweets

RT @techsynd: Intel Buys Egypt-Based SySDS ro Boost Its 4G LTE Efforts: http://tinyurl.com/4brmh4n

about 1 hour ago via EgyTweets

Randeep Ramesh and Alex Hern

The Guardian, Wednesday 13 November 2013 15.40 GMT

Jump to comments (1284)





Spam

A speech in which David Cameron said the internet would help people hold politicians to account was among those deleted. Photograph: Barcroft Media



Are we loosing the past of the web?

Library of Congress

- In April 2010 LoC and Twitter signed an agreement to archive all tweets since 2006
- January 2013: It is clear that technology to allow for scholarship access to large data sets is lagging behind technology for creating and distributing such data. The Library is pursuing partnerships to allow some limited access capability in reading rooms.

German National Library

 Based on a law of June 22, 2006, the GNL should collect, enrich, catalog, archive Web publications

Internet Archive

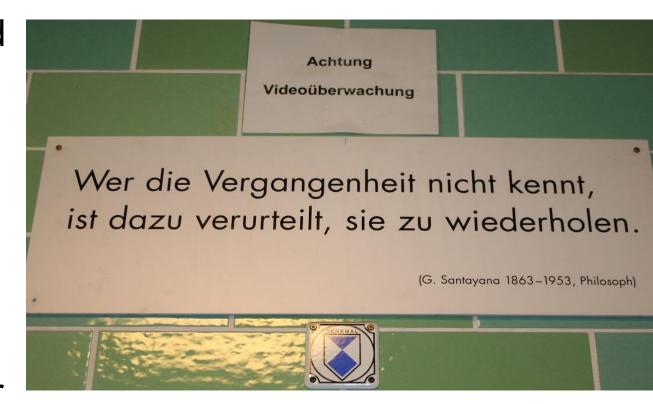
- Archiving the Web (10 Petabyte) since 1996
- Access possible through the URL

Relevant Projects @ L3S

- Web Archiving: LiWA, ARCOMEM, ForgetIT
- Web Search: PHAROS, CUBRIK
- Web and Stream Analytics: EUMSSI, Qualimaster
- ERC Advanced Grant: ALEXANDRIA (2014 2018, 2.5 Mill. Euro)

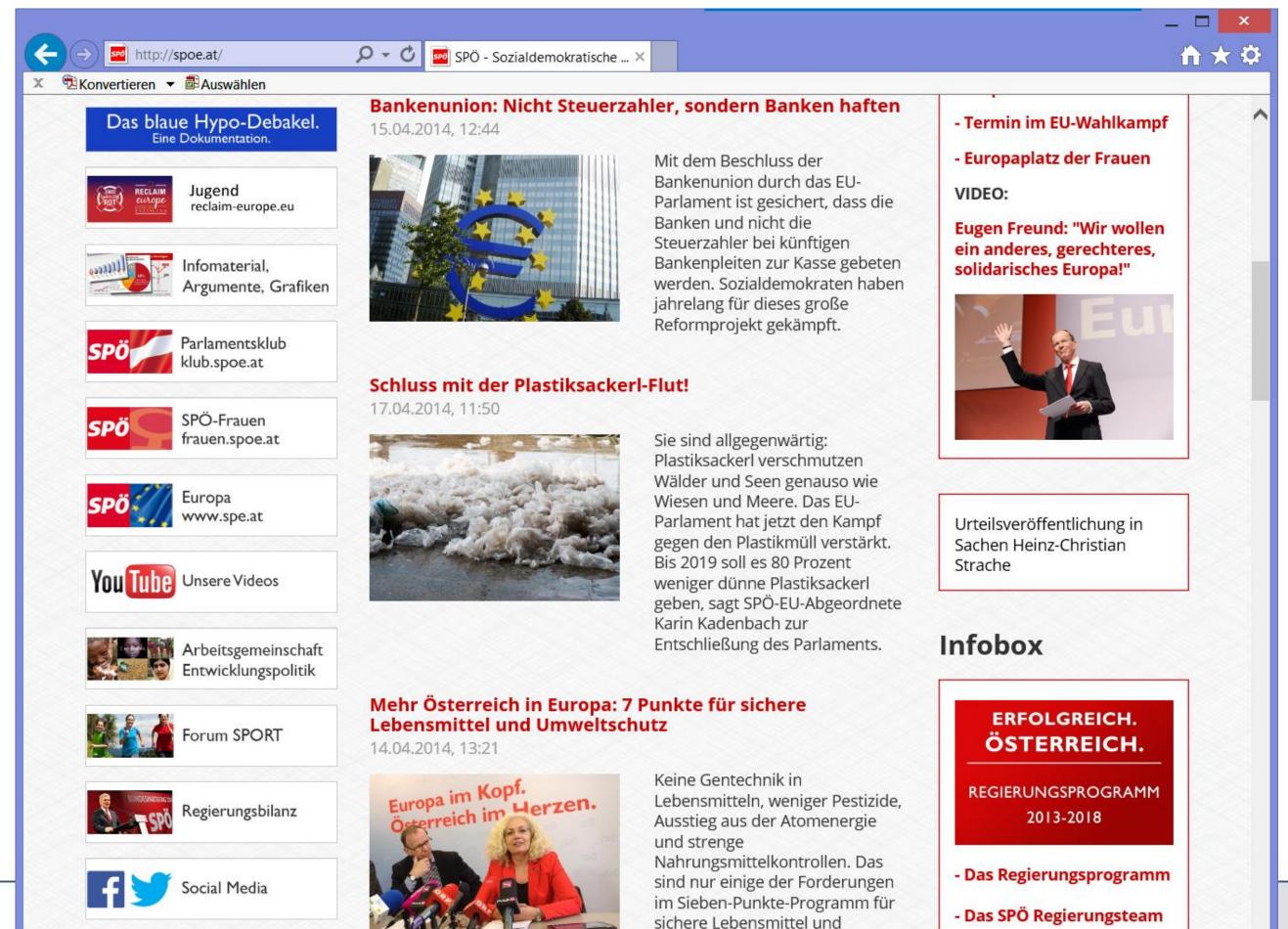
Cooperations

■ German National Library, British Library, Internet Archive, Rutgers University, et al





Looking back: The Austrian Socialist Party and Europe





Wayback Machine

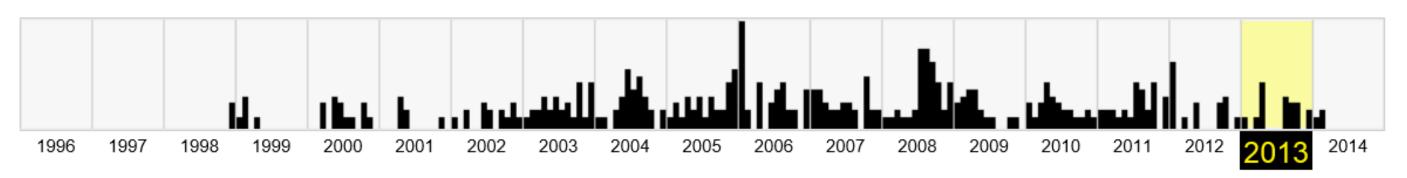
http://spoe.at/

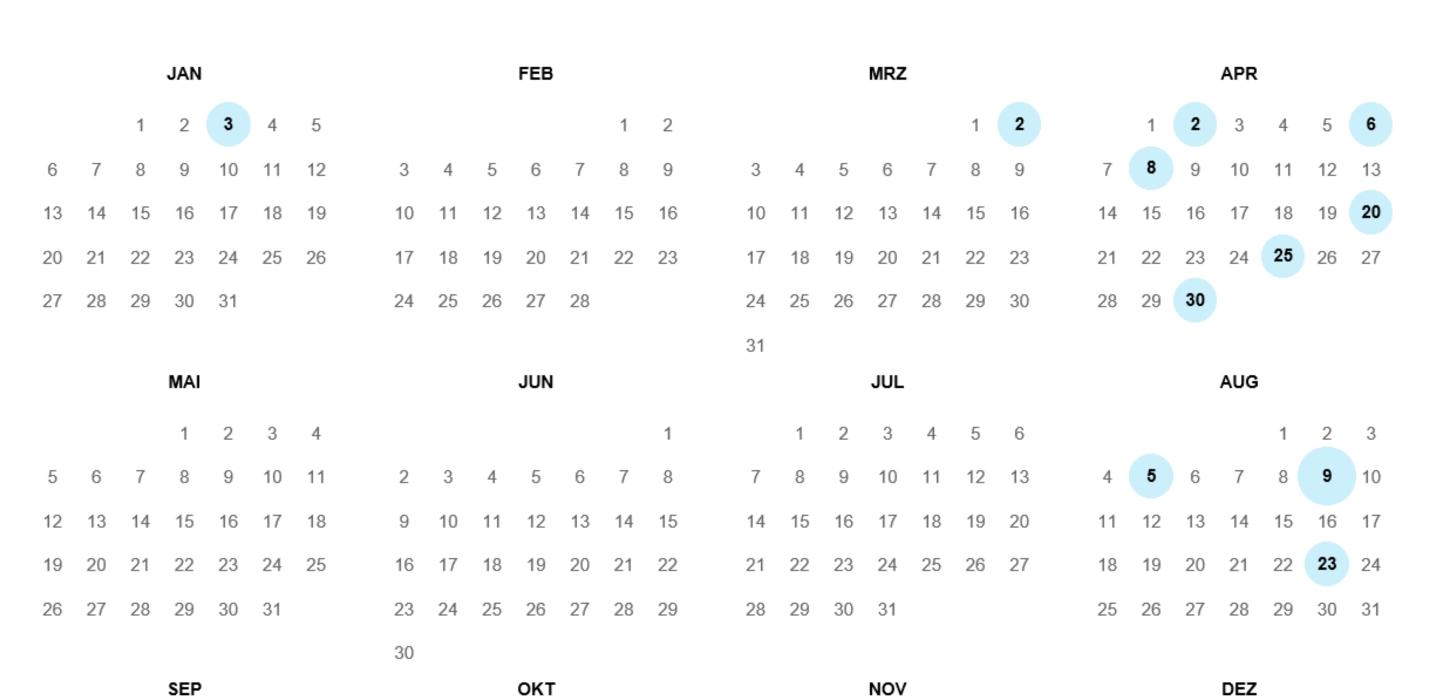
BROWSE HISTORY

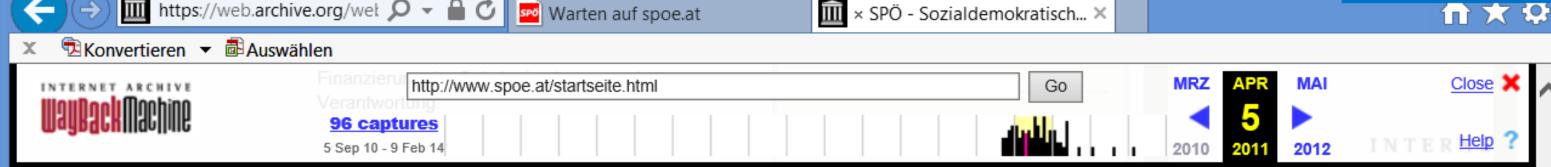
http://spoe.at/

Saved 519 times between Dezember 5, 1998 and Februar 12, 2014.

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







31.03.2011 [12:00]

Arbeitsmarktöffnung: Gesetz gegen Lohn- und Sozialdumping beschlossen



Am 1. Mai öffnet Österreich seinen Arbeitsmarkt für die neuen EU-Staaten. Um darauf vorbereitet zu sein, die kollektivvertraglich festgelegten Löhne zu schützen und die hohen österreichischen Standards aufrecht zu erhalten, hat die SPÖ-geführte Bundesregierung gesetzliche Maßnahmen ergriffen, die pünktlich in Kraft treten werden.

30.03.2011 [13:00]

Cold Calling: Neues Gesetz gegen unseriöse Werbeanrufe



Die Strafen für unzulässige Werbeanrufe wurden empfindlich erhöht. Infrastrukturministerin Doris Bures knöpft sich nun überhöhte Handy-Rechnungen vor.

30.03.2011 [12:00]

Der Euro-Schutzschirm ist wichtig für Krisenbewältigung und Stabilität



Nicht am Euro-Schutzschirm mitzuwirken würde eine Schwächung der Abwehr von Spekulation und eine Schwächung Österreichs bedeuten, machte Bundeskanzler Werner Faymann in einer Aktuellen Stunde des Nationalrats deutlich. Die FPÖ-Ideen eines Euro-Ausstiegs kritisiert Faymann als "Abenteuer".

Bundesländer



Aus den Ländern

- Wien | Sima l\u00e4dt zum gro\u00dden Wiener Anti-Atom-Gipfel
- OÖ | "Lascher Umgang mit rechtsextremen Straftaten in Oberösterreich"
- NÖ | Arbeitsklausur des SPNÖ-Landtagsklubs in Gmünd
- Bgld | Landtag: Schulterschluss gegen Atomkraft und für Energiewende
- Stmk | Gratis-Urlaub f\u00fcr bed\u00fcrftige SeniorInnen
- Sbg | Heimische Ressourcen zur Energieerzeugung nutzen
- Ktn | FPK startet
 Ablenkungsmanöver von Connect-Affäre
- Tirol | Junge Menschen wollen freiwilliges soziales Jahr
- Vbg | BürgerInnen sparen clever eine Kampagne der SPÖ Vorarlberg

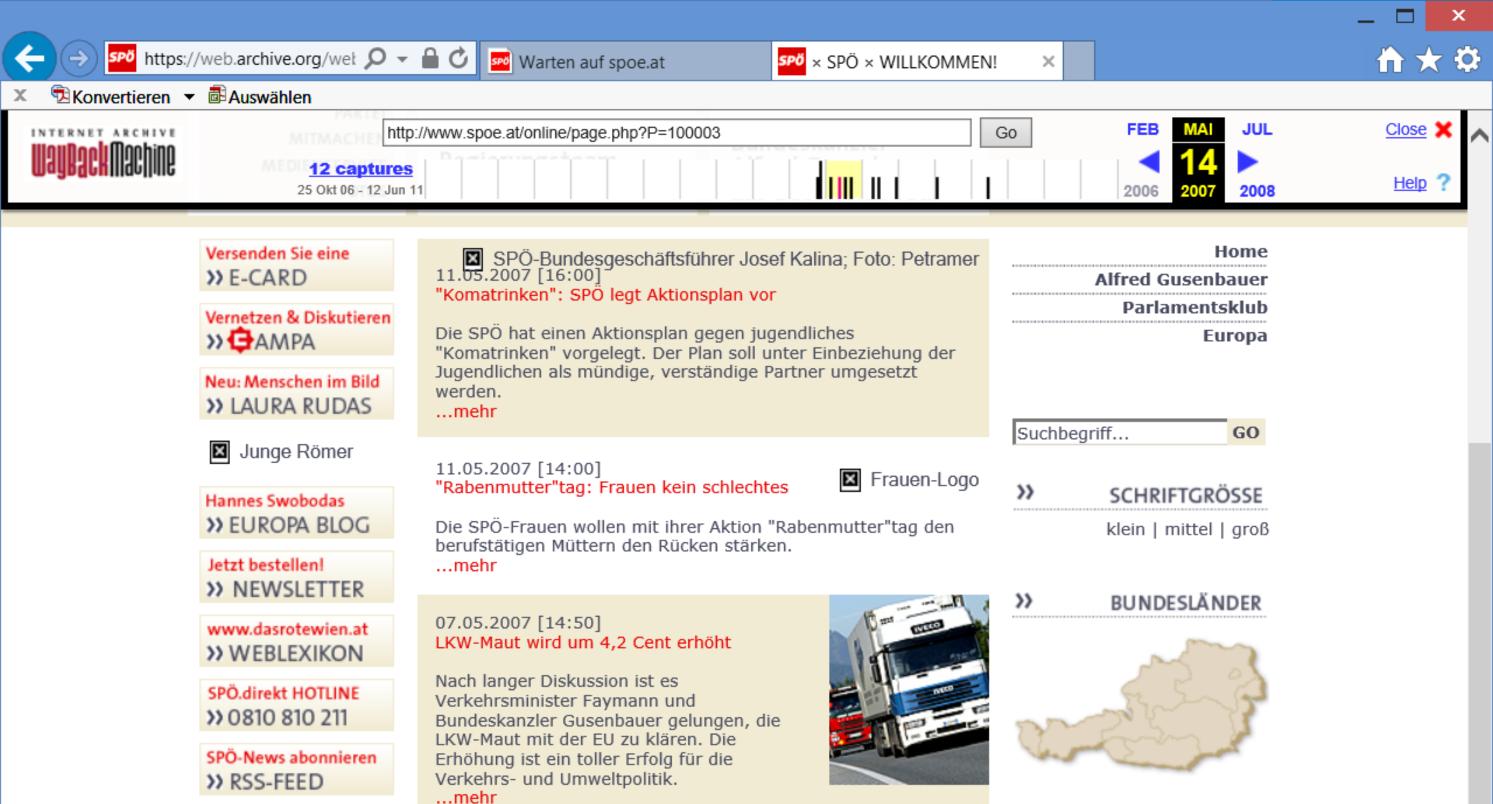


http://www.facebook.com

Lá



Weitere Meldungen ...



07.05.2007 [12:00]

Jetzt bestellen!

Imageshop

>> SPÖ SHOP

>> SPÖ-AKTUELL

SPÖ-Buchhandlung

>> LÖWELSTRAßE

Mädchenförderung: Reformbedarf bei Blum-Bonus

Frauenministerin Doris Bures will, dass Lehrstellen in Zukunftsbranchen stärker gefördert werden, als in klassischen Frauenberufen ohne Perspektiven. ...mehr



>> ORGANISATIONEN

SPÖ Frauen
Jugend
Kinderfreunde
Pensionistenverband
Wirtschaftsverband
SPÖ Bauern
Renner Institut
Zukunftswerkstätte
Alle Organisationen

>> PROGRAMME

Regierungsprogramm SPÖ-Umweltprogramm Parteiprogramm Bildungsprogramm Wirtschaftsprogramm











FÜR GERECHTE PENSIONEN

Talen

■ Konvertieren

■ Auswählen





Wirtschaftskammerwahlen, die von 12. bis 15. März 2005 stattfinden, zu erreichen, sagte SPÖ-Budgetsprecher Christoph Matznetter am Freitag in einer Pressekonferenz, bei der er als Spitzenkandidat des WV präsentiert wurde.

mehr 2



08.10.2004 [11:45 Uhr]

Katastrophale Zustände an Uni Wien

An der Uni Wien herrscht Chaos. Vor allem an der Publizistik und der Pädagogik sind die Betreuungsverhältnisse katastrophal. Die SPÖ fordert eine sofortige Finanzspritze für die Unis.

mehr 🖸



Frauenstudie 2004:

SEP

Berater-Gesellschaft

SPÖ hat seit

2000 bereits

gewonnen

Europäische

Rifkin: Der

Traum

years?

Volkshilfe bittet um

Darfur

HINTERGRUND

Spenden für

Flüchtlinge aus

USA: Four more

bei 12 Wahlen

Junge Frauen wollen berufstätig sein

□ OECD-Studie: Bildung auf einen Blick



Wendewörterbuch Standardwerk zur K.O.lition



Programm Das Grundsatz-Programm



Kommentare, Diskussionen & Hintergrundinformationen zu brisanten Themen

07.10.2004 [13.45 Uhr]

Gusenbauer gratuliert Jelinek zu Literatur-**Nobelpreis**

Auf das Herzlichste gratuliert SPÖ-Vorsitzender Alfred Gusenbauer Elfriede Jelinek, die mit dem Literatur-Nobelpreis ausgezeichnet wird.

mehr 🛂



06.10.2004 [15:15 Uhr]

Türkei: EU-Kommissionsbericht ungenügend

Für SPÖ-Vorsitzenden Alfred Gusenbauer beleuchtet der EU-Kommissionsbericht die Auswirkungen eines Türkei-Beitritts auf die EU ungenügend.

mehr D



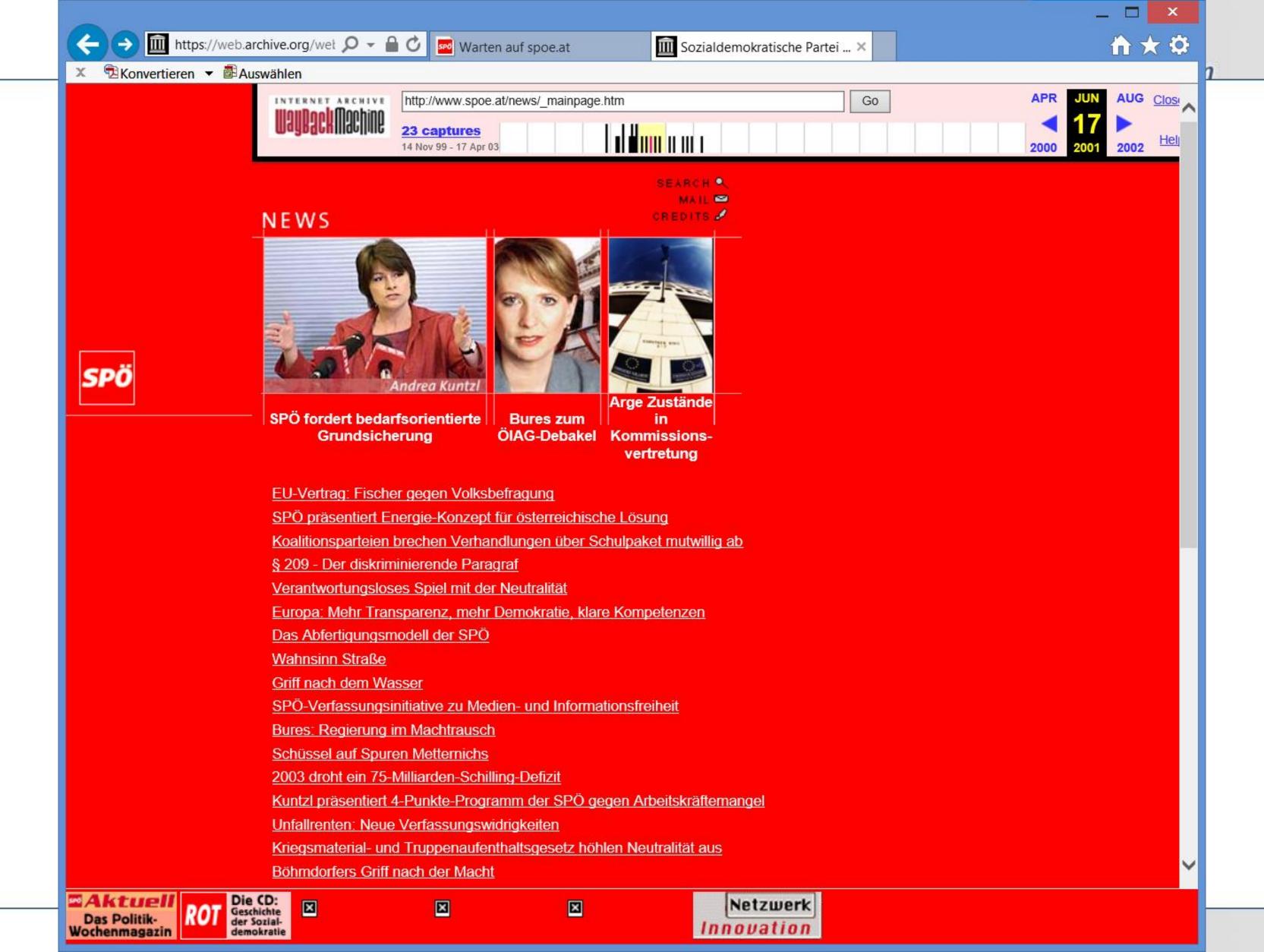
05.10.2004 [12:30 Uhr]

Arbeitslosigkeit: Desaströse Bilanz der Regierung Schüssel

Seit dem Jahr 2000 ist die Arbeitslosigkeit in Österreich um 35,8 Prozent gestiegen, es herrscht Rekordarbeitslosigkeit vor. "Die Bilanz der Regierung Schüssel ist desaströs", kommt SPÖ-Bundesgeschäftsführerin Doris Bures zum Schluss.

mehr 2

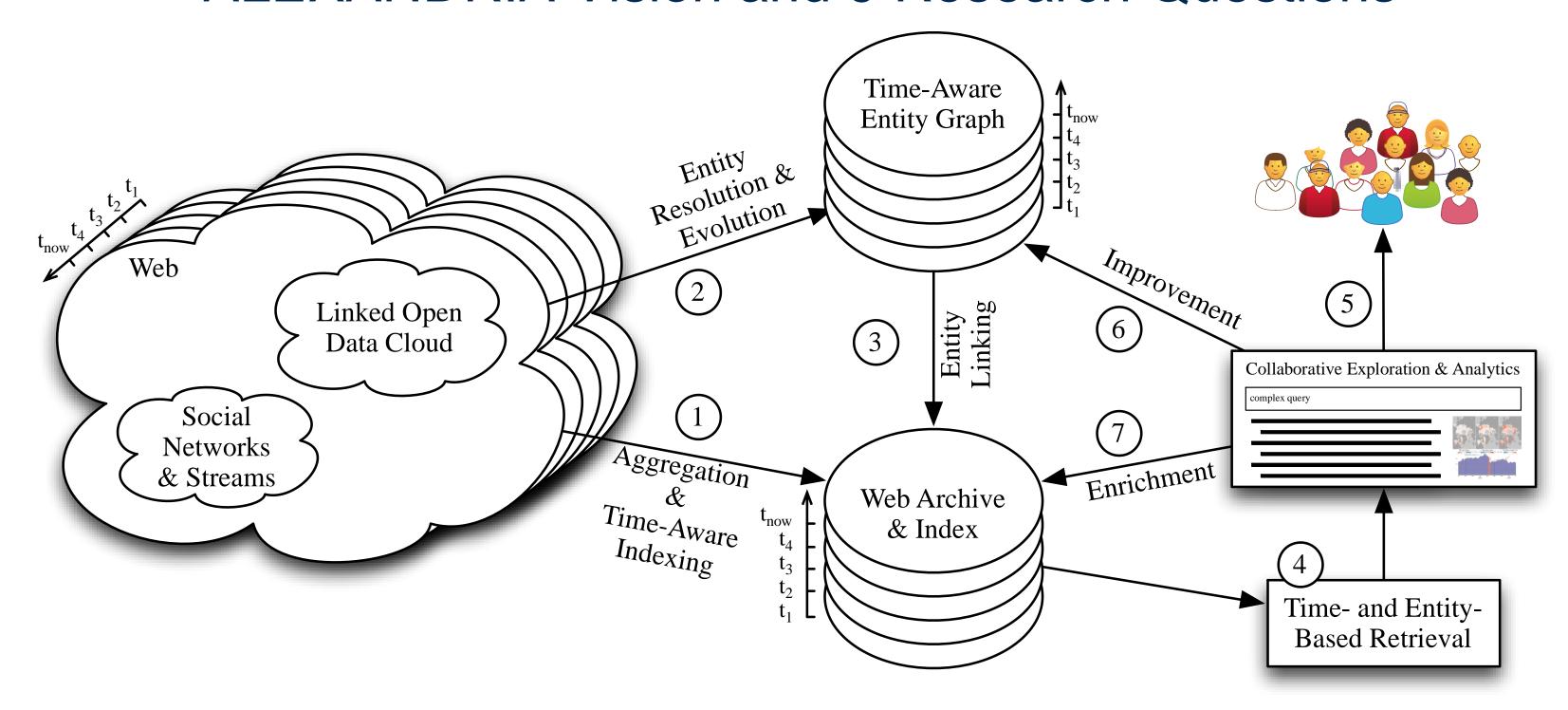






What is missing?

ALEXANDRIA Vision and 9 Research Questions





Evolution-Aware Entity-Based Enrichment and Indexing

Q1: How to link web archive content against multiple entity and event collections evolving over time?

Ioannou, E., Nejdl, W., Niederée, C. and Velegrakis, Y. 2011. LinkDB: A Probabilistic Linkage Database System. *SIGMOD* (New York, New York, USA, Jun. 2011)

Q2: How to maintain entity and event information and indexes for webscale archives?

Papadakis, G., Ioannou, E., Niederée, C., Palpanas, T. and Nejdl, W. 2012. Beyond 100 million entities: large-scale blocking-based resolution for heterogeneous data. WSDM (New York, NY, USA, 2012), 53–62.

Papadakis, G., Ioannou, E., Palpanas, T., Niederée, C. and Nejdl, W. 2012. A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces. TKDE. (2012).



Huge and Heterogeneous Information Spaces

Voluminous, (semi-)structured datasets.

- DBPedia 3.4: 36,5 million triples and 2,1 million entities
- *BTC09*: 1,15 billion triples and 182 million entities.

Users are free to insert not only attribute values but also attribute names → high levels of heterogeneity.

- DBPedia 3.4: 50,000 attribute names
- Google Base: 100,000 schemata and 10,000 entity types.

Large portion of data stemming from automatic information extraction → noise, tag-style values

and this does neither involve time nor entity evolution ...



Aggregating Social Networks and Streams

Q3: How to archive complex and dynamic network structures from social media?

Siersdorfer, S., Chelaru, S., Nejdl, W. and San Pedro, J. 2010. How useful are your comments? Analyzing and Predicting YouTube Comments and Comment Ratings. *WWW* (New York, New York, USA, Apr. 2010), extended for TWEB (2014)

Risse, T., Dietze, S., Peters, W., Doka, K., Stavrakas, Y. and Senellart, P. 2012. Exploiting the Social and Semantic Web for guided Web Archiving. *TPDL* (Sep. 2012)

Q4: How to aggregate social media streams for archiving?

Minack, E., Siberski, W. and Nejdl, W. 2011. Incremental diversification for very large sets: a streaming-based approach. *SIGIR* (New York, New York, USA, Jul. 2011)

Diaz-Aviles, E., Drumond, L., Schmidt-Thieme, L. and Nejdl, W. 2012. Real-time top-n recommendation in social streams. *RecSys* (New York, New York, USA, 2012)



Using comment analysis to find relevant resources













Figure 8: Videos with high (upper row) versus low variance (lower row) of comment ratings

Table 3: Top and Bottom-25 tags according to the variance of comment ratings for the corresponding videos

High comment rating variance							
presidential	nomination	muslim	shakira	islam nice iraq celebrities			
campaign	station	itunes	grassroots				
xbox	barack	efron	zac				
3g	kiss	obama	deals				
jew	space	shark	hamas	kiedis			
Low comment rating variance							
betting	turns	puckett	tmx	tropical			
skybus	peanut	defender	f-18	vlog			
butter	chanukah	form	savings	iditarod			
lent	daylight casserole	egan	snowboard	havanese			
menorah		1040a	1040ez	booklet			



Temporal Retrieval and Ranking

Q5: How to support time-sensitive and entity-based query formulation?

Kanhabua, N. and Nørvåg, K. 2010. Exploiting time-based synonyms in searching document archives. *JCDL* (New York, New York, USA, Jun. 2010)

Nguyen, T., and Kanhabua, N. 2014. Leveraging dynamic query subtopics for time-aware search result diversification. ECIR (Amsterdam, April 2014)

Q6: How to improve result ranking and clustering for time-sensitive and entity-based queries?

Kanhabua, N., Blanco, R. and Matthews, M. 2011. Ranking related news predictions. *SIGIR* (New York, New York, USA, Jul. 2011)

G. Demartini, C. Firan, T. Iofciu, R. Krestel, W. Nejdl: Why finding entities in Wikipedia is difficult, sometimes. Inf. Retr. 13(5): 534-567 (2010)



Dynamic subtopic mining for query extension and ranking

query: ncaa

March 14th	March 31th	April 07th	
0.0132 · march madness schedule	0.0100 · oakland raiders	0.0122 · ncaa women's basketball tournament	
$0.0117 \cdot \text{ncaa basketball tournament}$	0.0090 · ncaaw	0.0053 · ncaa basketball tournament	
0.0068 · nfl draft	0.0042 · tito francona	$0.0049 \cdot \text{cbs sports line}$	
$0.0048 \cdot \text{selection sunday}$	$0.0031 \cdot \text{ncaa brackets}$	0.0033 · ncaaw	
$0.0037 \cdot \text{oakland raiders}$	0.0029 · ncaa division ii	0.0031 · ncaa final four	
0.0032 · 2006 ncaa tournament bracket	$0.0024 \cdot \text{andy goram}$	0.0029 · ncaa wrestling	
0.0026 · brad hopkins released nfl	$0.0024 \cdot lakers$	$0.0028 \cdot \text{march madness bracket}$	
$0.0023 \cdot \text{roger clemens}$	0.0024 · ncaa women's basketball	0.0019 · ncaa basketball results	
	tournament bracket		
0.0021 · ncaa division ii	0.0021 · ncaa basketball brackets	$0.0009 \cdot \text{andy goram}$	
$0.0014 \cdot \text{college basketball}$	$0.0021 \cdot \text{nit brackets}$	0.0009 · ncaa division ii	

14/03/2006

18/03/2006

01/04/2006

march madness began

ncaa women tournament began

final four began



Collaborative Exploration and Analytics

Q7: How to support collaborative and complex search and analysis processes?

Ivana Marenzi and Sergej Zerr. Multiliteracies and Active Learning in CLIL - The Development of LearnWeb2.0 - IEEE Transactions on Learning Technologies (2012)

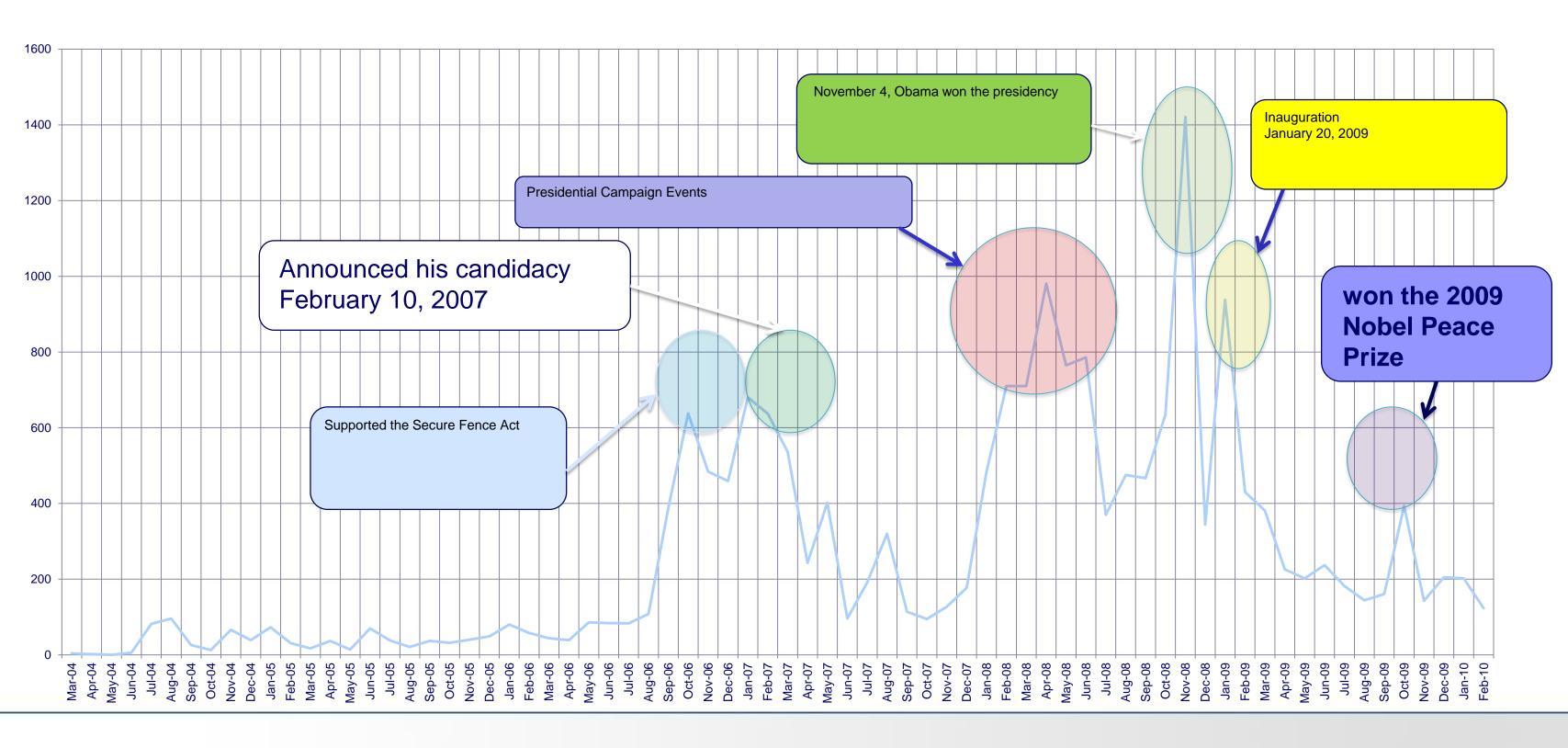
Q8: How to leverage (user) search and analysis processes to improve the web archive?

K. Bischoff, C. Firan, W.Nejdl, R. Paiu: Bridging the gap between tagging and querying vocabularies: Analyses and applications for enhancing multimedia IR. J. Web Sem. 8(2-3): 97-109 (2010)

M. Georgescu, N. Kanhabua, D. Krause, W. Nejdl, S. Siersdorfer: Extracting Event-Related Information from Article Updates in Wikipedia. ECIR 2013: 254-266

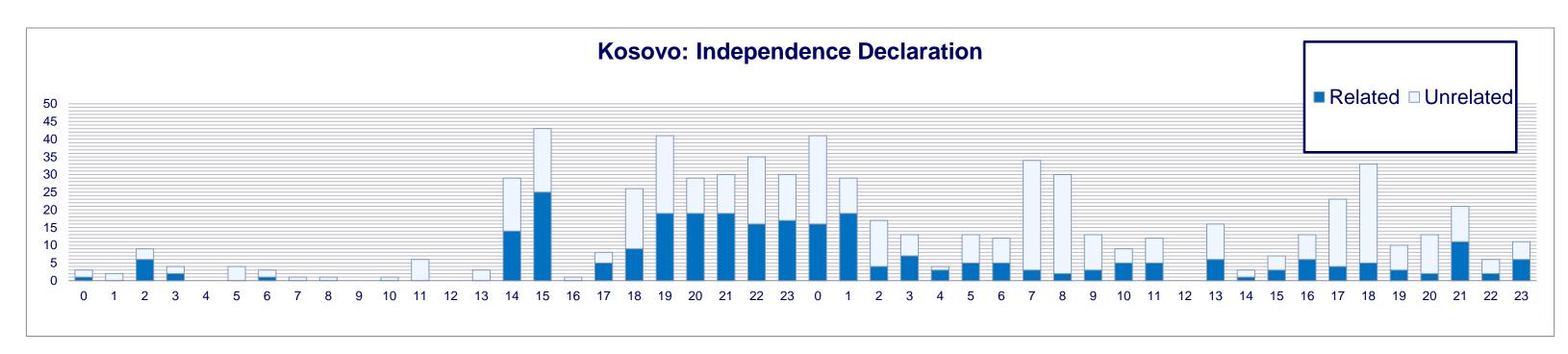


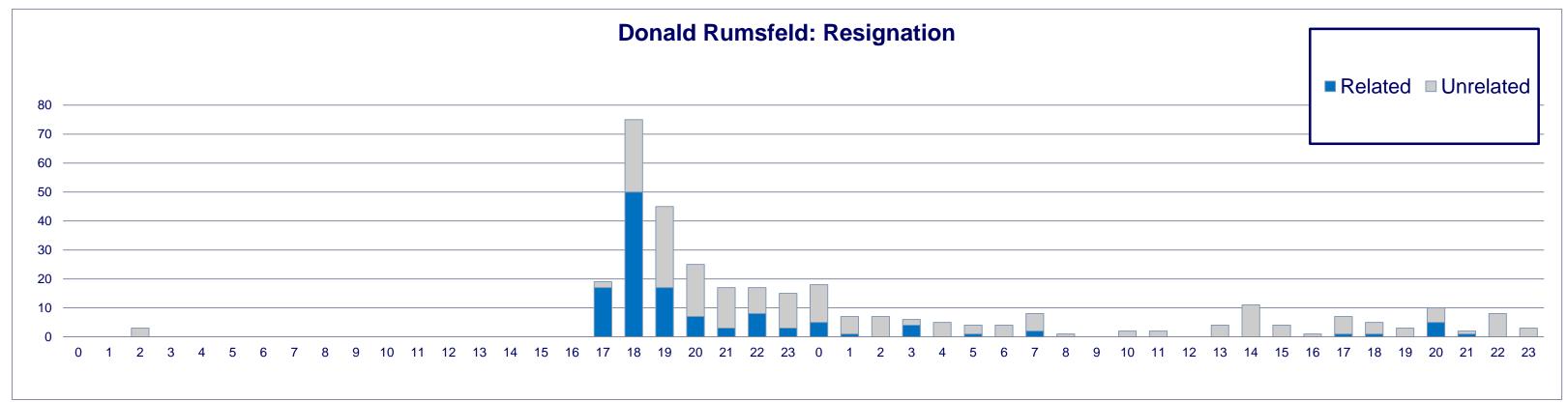
Peaks in Wikipedia update activity correlate with events Edit history for the Barack Obama article (monthly)





Update activity: controversy- or event-related?







Trust, privacy, and privacy preserving data mining

Q9: How to achieve privacy using privacy-preserving data publishing and data-mining?

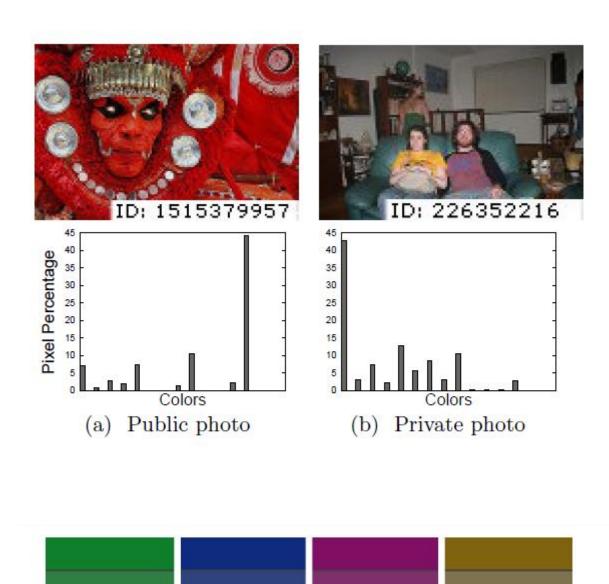
- W. Nejdl, D. Olmedilla, M. Winslett: Peertrust: Automated trust negotiation for peers on the semantic web. Secure Data Management 2004, 118-132.
- S. Zerr, D. Olmedilla, W. Nejdl, W. Siberski: Zerber+R: top-k retrieval from a confidential index. 12th Intl. Conference on Extending Database Technology, EDBT 2009, Saint Petersburg, Russia.
- S. Zerr, S. Siersdorfer, J. S. Hare, E. Demidova: Privacy-aware image classification and search. SIGIR 2012, 35-44
- N. Forgó, T. Krügel: Mit oder ohne Zustimmung? Soziale Netzwerke und der Datenschutz. FL 2011

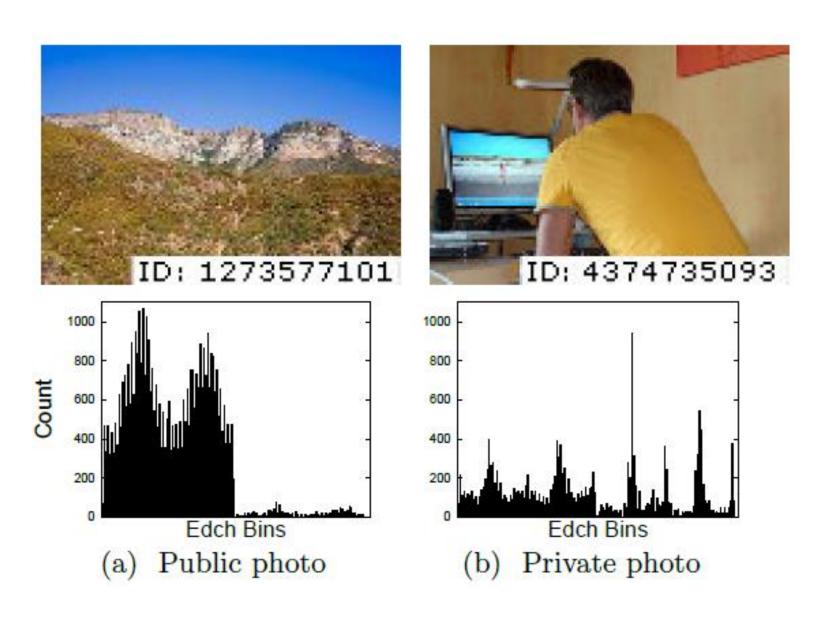


Public

Private

Public and private photos: colors and edges







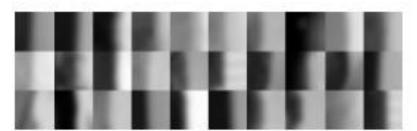
Public and private photos: SIFT and text



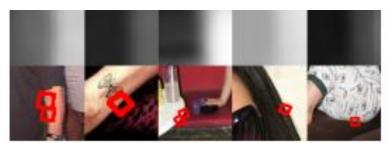
(a) Public features



(a) Public features



(b) Private features



(b) Private features

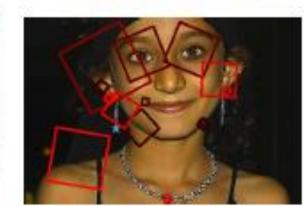






(a) Public photos





(b) Private photos

Terms for Private photos					
studio	photograph	sexi	strobist	februari	
blond	kid	hat	love	woman	
year	smile	black	eye	famili	
white	fun	portrait	$\operatorname{selfportr}$	peopl	
dress	valentin	day	friend	parti	
femal	boy	birthday	face	model	
fashion	daughter	children	coupl	self	
babi	cute	child	girl	beauti	
me	nikon	pretti	hair	man	
photographi	young	lip	happi	wed	

Terms for Public photos						
craft	mountain	natur	rock	draw		
sky	tree	snow	jan	flower		
januari	art	paint	lake	winter		
view	citi	car	moon	landscap		
anim	cat	road	north	build		
frozen	island	full	ice	bridg		
water	abstract	hdr	park	sea		
hous	boat	fire	bird	handmad		
architectur	sunset	sunris	cloud	macro		
night	dog	illustr	design	river		



Web Science - Investigating the Future of Information and Communication

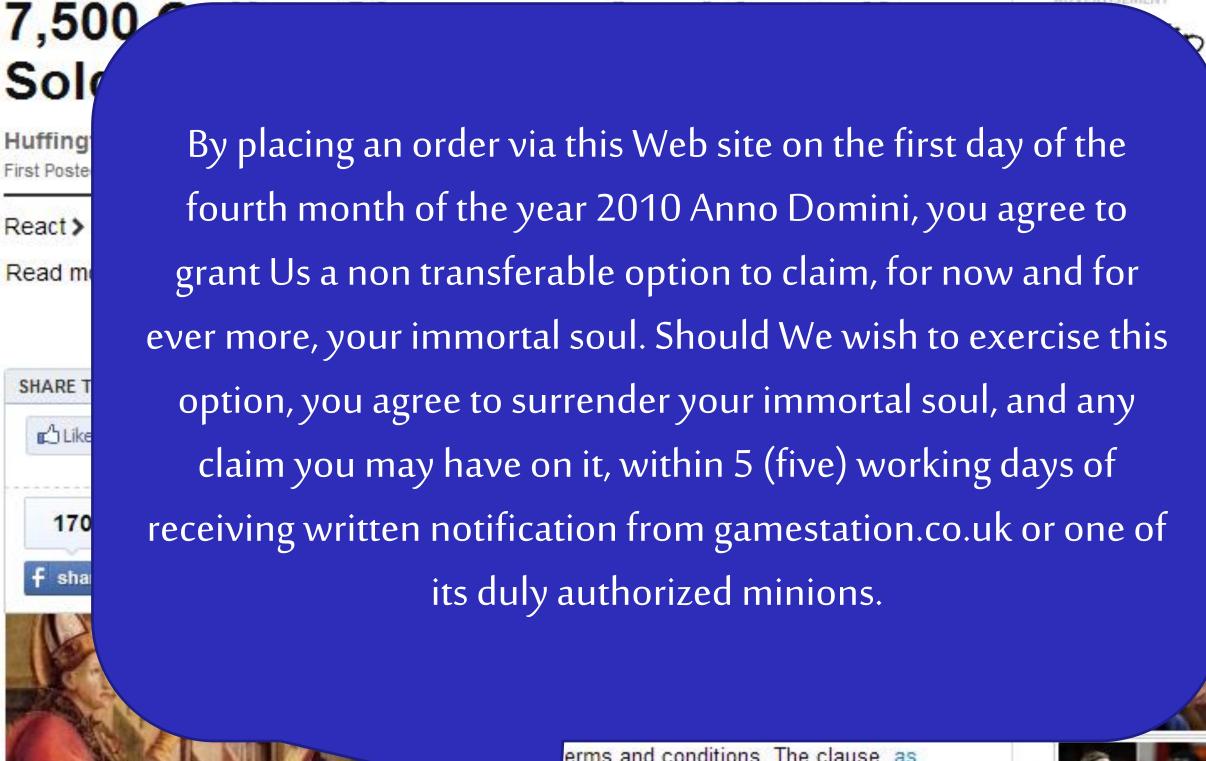


Facebook © 2011 · Deutsch Handy · Freunde finden · Banner · Personen · Seiten · Über uns · Werbung · Seite erstellen · Entwickler · Karrieren · Datenschutz · Impressum/Nutzungsbedingungen ·

Deutsch Italiano English (US) Español Português (Brasil) Français (France) العربية (中文(简体) 日本語 »

Get Technology Alerts

Sign Up



erms and conditions. The clause, as by GameStation, reads,

By placing an order via this Web site on the first

day of the fourth month of the year 2010 Anno Domini, you agree to grant Us a non transferable option to claim, for now and for ever more, your immortal soul. Should We wish to exercise this option, you agree to surrender your immortal soul, and any claim you may have on it, within 5



for a tir

ULAR

Cut

of

fo

Eli:

Ta

Ho

Wa

PR

NC

To

Bra





Alexandria Talks

Monday

- Creation of Focused Web Archives for Scientists (Elena Demidova)
- Temporal Web Dynamics and Implications for Information Retrieval (Nattiya Kanhabua)

Tuesday

- Studying Evolution of Temporal Collections (Avishek Anand)
- The Boon and Bane of Digital Forgetting (Claudia Niederee)
- Advanced Random Walk Techniques for Social Media Analysis (Xiaofei Zhu)
- WikiTimes: A Knowledge Base of News Events with Daily Summaries By the Crowd (Mohammad Alrifai)



Partner Talks

Monday

- Processing the National Mandate: Experiences and Ambitions in DNB (Elisabeth Niggemann)
- Observing the Web (Wendy Hall)
- Beyond 10 Blue Links: User-Oriented Design of Search Interfaces (Norbert Fuhr)
- Exploratory Entity Search over Time (Maarten de Rijke)
- Big Data & Big Theory: Utilizing Large Scale Data to Generate New Theories About Social Interaction (Matthew Weber)
- Multiple Media Analysis and Visualization with Large-Scale Temporal Web Archives (Masashi Toyoda)

Tuesday

- Collecting and Providing Access to Large Scale Archived Web Data (Helen Hocks-Yu)
- Enabling Analysis of Web Archives (Vinay Goel)