

Creation of Focused Web Archives for Scientists

Elena Demidova, Thomas Risse and Gerhard Gossen L3S Research Center, Hannover, Germany

ALEXANDRIA Workshop 15 / 16 September 2014 Hannover

Elena Demidova



Web Archiving

Web Archiving started with the Internet Archive 1996

- Meanwhile many archival organizations around the world are \bullet involved
- Aim: Long-term Web Preservation
- Method: Snapshots of domains (e.g. .de., .uk), rarely focused lacksquarecrawls (e.g. event-based at DNB)

Today

- Increasing scientific interest on Web content from the past
- Social Sciences, Historical Sciences, Law, ...
- What are the research questions?
- What are the aims of crawling?
- What are suitable approaches?
- Other requirements?

Initial Study with scientists at Leibniz University Hannover and GESIS – Leibniz Institute for the Social Sciences





Elena Demidova

15.09.2014



Historical Sciences

Study the past and how it relates to humans

- Events become interesting 25 years later
- Typically reconstruction of the history from fragments

Limited and diverse acceptance of the Web and Web Archives

- Web resources are anonymous and non-permanent
- Frequent changes with lack of proper archiving
- **Missing Provenance**
- Authenticity of content
- No unique identifier
- Lack of citation enabled Web archiving

@prefix key: <ht</pre> prefix ns: <htt ns:m.04jpl



Elena Demidova



Historical Sciences

Interesting Content

- Official Publications (e.g. Government)
- **Journalistic Resources**

What and When to collect?

- Important topics and events with a high media coverage
- Multi-cultural or controversial topics
- Optimal: continues observation of topics in the Web



Top-100 most controversial Wikipedia articles in English and German

Elena Demidova

15.09.2014



Social Sciences

Understand the society and the relationship among individuals within the society

What and When to collect?

- No indicators exists. Mainly defined by the researchers
- Observations of topics and events on major sites are good starting points
- Identified Topic
 - Official publications, journalistic and social media sources
 - Changes on the topic should be identified
- Metadata / Context (e.g. Author, Organizations and their interests, gender, location)
- Demographic information about social sites
- Provenance: Transparency and detailed documentation of content selection





Thomas Risse

15.09.2014



Law

- Research is based on official publications and protocols of parliaments or comments released by publishers
- Social media (especially blogs) are increasingly used
 - Only used as background information
 - Reason: missing citability and authenticity of resources

Genesis of laws

- Used to understand original intention of laws
- A democratic system requires a complete documentation of the law genesis.
- Currently different degrees of documentation
 - Official publications: parliament and committee meetings
 - Public discourse



Thomas Risse

15.09.2014



Derived Requirements (1/2)

Topical Dimension

- Crawl intention are mainly focused around events and rarely around entities
- What is the intention of the researcher?
- Easy monitoring by the researcher and possibility to correct
- **Time Dimension**
- Start, duration and termination of crawls
- Depends on research plan
- Stop criterion for crawler?
- Flexible Crawling Strategies
- Shallow observation crawls
- Focused crawls with prioritization (e.g. PageRank and/or semantics)



Thomas Risse





Derived Requirements (2/2)

Social Web Crawling

- General interest with different media focus
- Integrated with Web crawler





Öffnen

Stories of IREX

l ibraries in Ukraine Serv

Authenticity

See a web page as the user saw the page (e.g. including ads and tweets at that time point)

Context and Provenance

- **Demographics of sites**
- Documentation of crawl specification and history
- Verification and identification mechanisms



	by Arthur Berrian, IREX/Liberia	
	According to statistics from the Liberia Ministry of Health and the United	A Providence of
in Rural	Nations Missions in Liberia (UNMIL) the Ebola death toll in West Africa has	
	reached more than 2000 persons. Every day we hear of the death of a friend	
	or family member in our communities to Ebola.	
king with	Because of the rapid spread of the virus, the government of Liberia has declared a state of emergency restricting movement instituting a pationwide.	
g the	curfew, and quarantining comparison because the strong how the country. We are seeing the prices of food rise, while hones fall. Citizens are firstrated with a	A Bridger
	government who they feel has failed to protect them.	



Thomas Risse



iCrawl is:

User-friendly

- Support for crawl specification
- Interactive crawl monitoring
- Integrated analysis tools
- Wayback Machine for immediate QA

Integrated

Social media and web crawl

Focused

- Selects crawled documents based on topic relevance
- Current work: Adapting to evolving topics

Reproducible

Documentation of entire crawl process



welt.de: 17.1 %

http://www.I3s.de/~gossen/icrawl/guide/index.html





http://okkam.l3s.uni-hannover.de:8080/

Crawl Specification Wizard

Add crawl +		
ekkam.I3s. uni-hannover.de :8080/campaign/4/add	v C 🛿 vitis hannover	▶ ♣ ☆ 🖻 🔣 🐠 =
iCrawl Wizard		
ebola Search sources: 🖂 Web 🖂 Twitter	Search	Seeds (URLs) Add http://de.wikipedia.org/wiki/Ebolavirus
Web Scheme Schem	Twitter Science Scienc	http://www.tatafonaija.com/2014/09/research-e
Keywords: Seite, Video, Sprachen, Werkzeuge, Entwickler, Lize Entities: WHO, Centers for Disease Control and Prevention, CDC, Marburg, Taï Forest Ebolavirus, RKI, Zaire Ebolavirus, Sudan Ebolavirus, Zaire, J. H. Kuhn, WHO, Centers for Disease Control and Prevention, CDC, Marburg, Taï Forest Ebolavirus, RKI, Zaire Ebolavirus, Sudan Ebolavirus, RKI, Zaire, J. H. Kuhn	ebola-may-hit-15-more-countries.html Keywords: Entities: Research: Ebola May Hit 15 More Countries http://t.co/aQ9Sh4VPBw Research: Ebola May Hit 15 More Countries http://t.co/NY1CbL7j4m Research: Ebola May Hit 15 More Countries http://t.co/HsFfMGNILt	Keywords Add Ebola-Patienten
Vom Ebola Virus (EBOV) sind mehrere Varianten bekannt, z. B. Mayinga (abgekürzt als EBOV/May). Hier wird von der zuständigen Expertengruppe geraten,	http://bit.ly/WMPe2a http://bit.ly/WMPe2a Keywords: Entities:	Entities Sudan Ebolavirus
Ebola 1.1.2 - Ebola - D http://www.heise.de/download/ebola.html Keywords: Entities:	No Ebola cases spotted in China: official: No Ebola cases have been reported in China so far, a Chinese health http://t.co/8Mmpv6dLvU No Ebola cases spotted in China: official: No Ebola cases have been reported in China so far, a Chinese health http://t.co/JsCcCCFXP7	Zentralafrika
Nastmært time: Repe	tition: ONCE V End time:	Download resources

Elena Demidova



Monitoring of Ongoing Crawls

Statistics for: •Domains •Top level domains •File size Distance from seeds •Mime-types •Crawl status •etc.

http://okkam.l3s.uni-hannover.de:8080/campaign/1/1/status

Domains	=
Domain Count	zeit.de: 20.0 % doccheck.com: 20.0 %
doccheck.com 1	
sueddeutsche.de 1	
wikipedia.org 1	
onmeda.de 1	onmeda.de: 20.0 % sueddeutsche.de: 20.0 %
zeit.de 1	
	wikipedia.org: 20.0 %

Elena Demidova



Full Information about Crawled Pages

			iCrawl Campaigns Status		
			Ebola: Crawl 1 Edit History Statistics	Data	
iCrawl Campaigns	Status		Data for crawl "Ebola-	News"	
Ebola: Crawl 1 Edit	History Statistics	Data	Selection FETCHED Filter		
Resource http://alamandas.com/		ndas com/	Title	Status	Fetched
		http://alamandas.com/	fetched	15.10.2014 10:05:54	
g_dirty	java.nio.HeapByteBuffer[pos=0 lim=4 cap=4] http://alamandas.com/ 2 1413360354110 1410766562711 2592000 0 0 0 0 0 SUCCESS_args=[]	http://ajax.aspnetcdn.com/ajax/jQuery/jquery-1.6.1.mi	n.js fetched	15.10.2014 10:05:52	
baseUri status fetchTime prevFetchTime fetchInterval retriesSinceFetch		http://imworld.aufeminin.com/dossiers/acc1_29869 /accNoTexte686x400a488591.jpg	fetched	15.10.2014 09:21:39	
		http://imworld.aufeminin.com/manage/bloc/D2014091 /onmeda-085353_L.jpg	5 fetched	15.10.2014 09:21:49	
prevModifiedTime protocolStatus		http://admin.brightcove.com/js/BrightcoveExperiences	.js fetched	15.10.2014 09:21:39	
content contentType prevSignature signature title text parseStatus score reprUrl	<pre>is a point is point is point is a point is a point is a point is a point</pre>				
headers	Кеу	Value			
	Server	Apache/2.2.27 (Unix) mod_ssl/2.2.27 OpenSSL/1.0.1e-fips mod_bwlimited/1.4			
	Connection	close			
	Date	Mon, 15 Sep 2014 08:05:50 GMT			
	Link	<http: p466gb-78="" wp.me="">; rel</http:>	l=shortlink		

	Settings	About
URL		
http://alamandas.com/	Wayb	ack
http://ajax.aspnetcdn.com/ajax/jQuery/jquery-1.6.1.min	.js Wayb	ack
http://imworld.aufeminin.com/dossiers/acc1_29869 /accNoTexte686x400a488591.jpg	Wayb	ack
http://imworld.aufeminin.com/manage/bloc/D20140915 /onmeda-085353_L.jpg	Wayb	ack
http://admin.brightcove.com/js/BrightcoveExperiences.j	js Wayb	ack

Elena Demidova



Wayback Machine View of Crawled Pages

iCrawl capture: 15 September 2014 Updated: 13 August 2014

IA capture: 2 August 2014 Updated: 28 October 2013



http://okkam.l3s.unihannover.de:8080/wayback/1/20141015080552/ http://www.onmeda.de/krankheiten/ebola.html

http://web.archive.org/web/20140802112525/ http://www.onmeda.de/krankheiten/ebola.html



Future work: focused content selection for Web archives

Problem: Existing (unfocused) web archives are too large to process for typical research questions

Solution: Focused sub-collections State of the art selection: by host, by domain Actually required: by topic, for entities/events (for specific time span)

Challenges:

Temporal consistency Link graph consistency

Approach:

Port and extend the methods from focused Web crawlers (iCrawl)

image: http://comsys.informatik.uni-kiel.de/wp-content/uploads/2013/04/web-crawler.gif

Elena Demidova



Thank You!





Elena Demidova



15/09/14