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

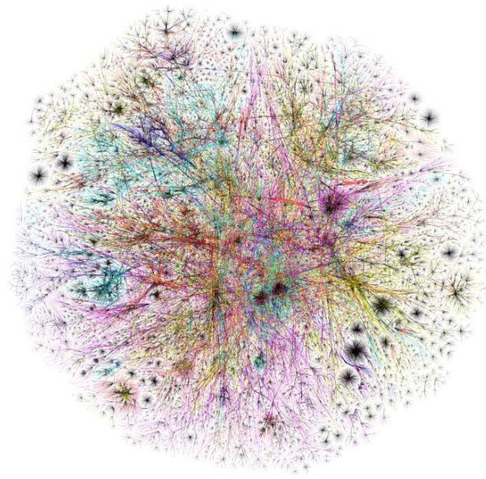
Web Archiving

Web Archiving started with the Internet Archive 1996

- Meanwhile many archival organizations around the world are involved
- Aim: Long-term Web Preservation
- Method: Snapshots of domains (e.g. .de., .uk), rarely focused crawls (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

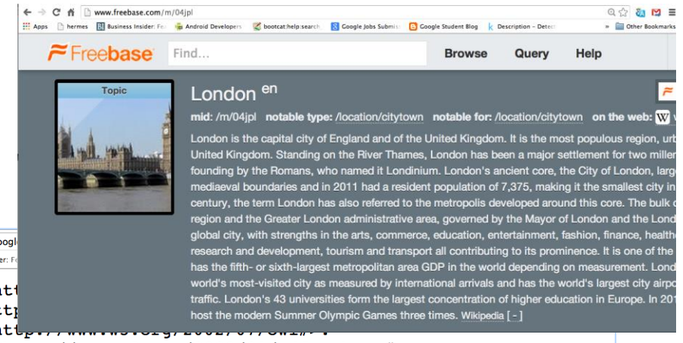
Historical Sciences

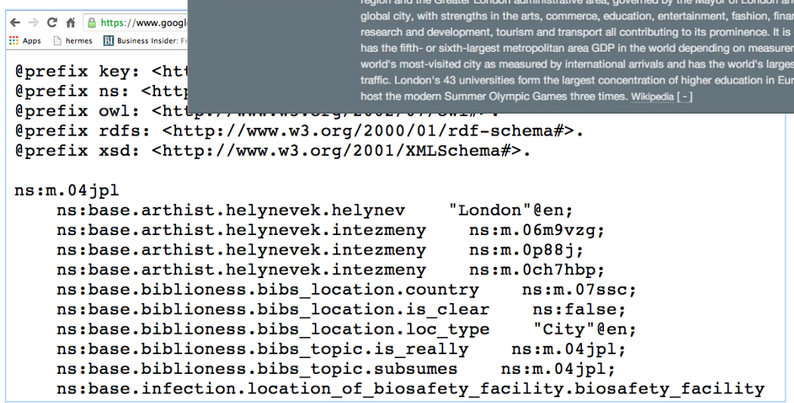
Study the past and how it relates to humans

- 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: <htt
@prefix ns: <http
@prefix owl: <htt
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.

ns:m.04jpl
    ns:base.arthist.helynevek.helynev    "London"@en;
    ns:base.arthist.helynevek.intezmeny    ns:m.06m9vzg;
    ns:base.arthist.helynevek.intezmeny    ns:m.0p88j;
    ns:base.arthist.helynevek.intezmeny    ns:m.0ch7hbp;
    ns:base.biblioness.bibs_location.country    ns:m.07ssc;
    ns:base.biblioness.bibs_location.is_clear    ns:false;
    ns:base.biblioness.bibs_location.loc_type    "City"@en;
    ns:base.biblioness.bibs_topic.is_really    ns:m.04jpl;
    ns:base.biblioness.bibs_topic.subsumes    ns:m.04jpl;
    ns:base.infection.location_of_biosafety_facility.biosafety_facility
```

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

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

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

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)

Derived Requirements (2/2)

Social Web Crawling

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

Authenticity

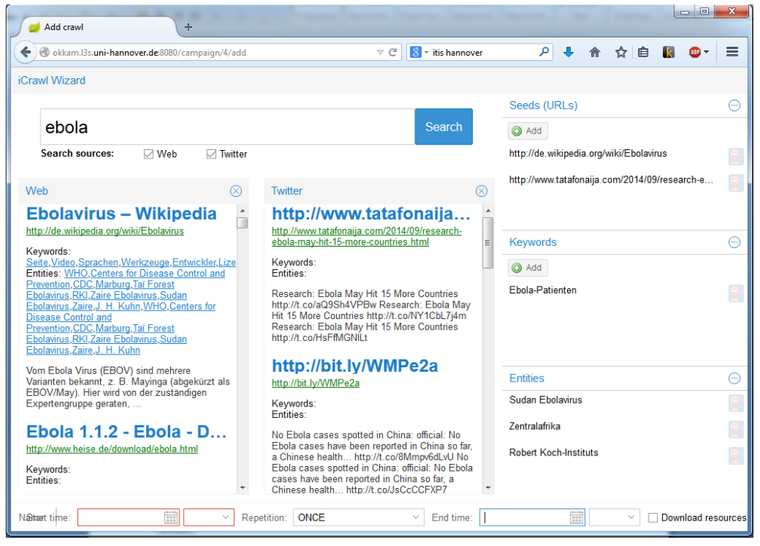
- See a web page as the user saw the page (e.g. including ads and tweets at that time point)

Context and Provenance

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

iCrawl is:

http://www.l3s.de/~gossen/icrawl/guide/index.html



User-friendly

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

Integrated

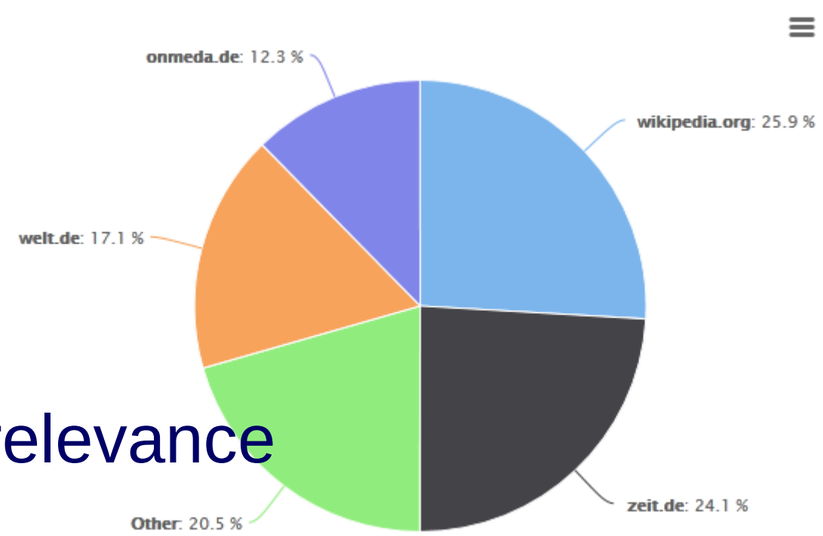
- Social media and web crawl

Focused

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

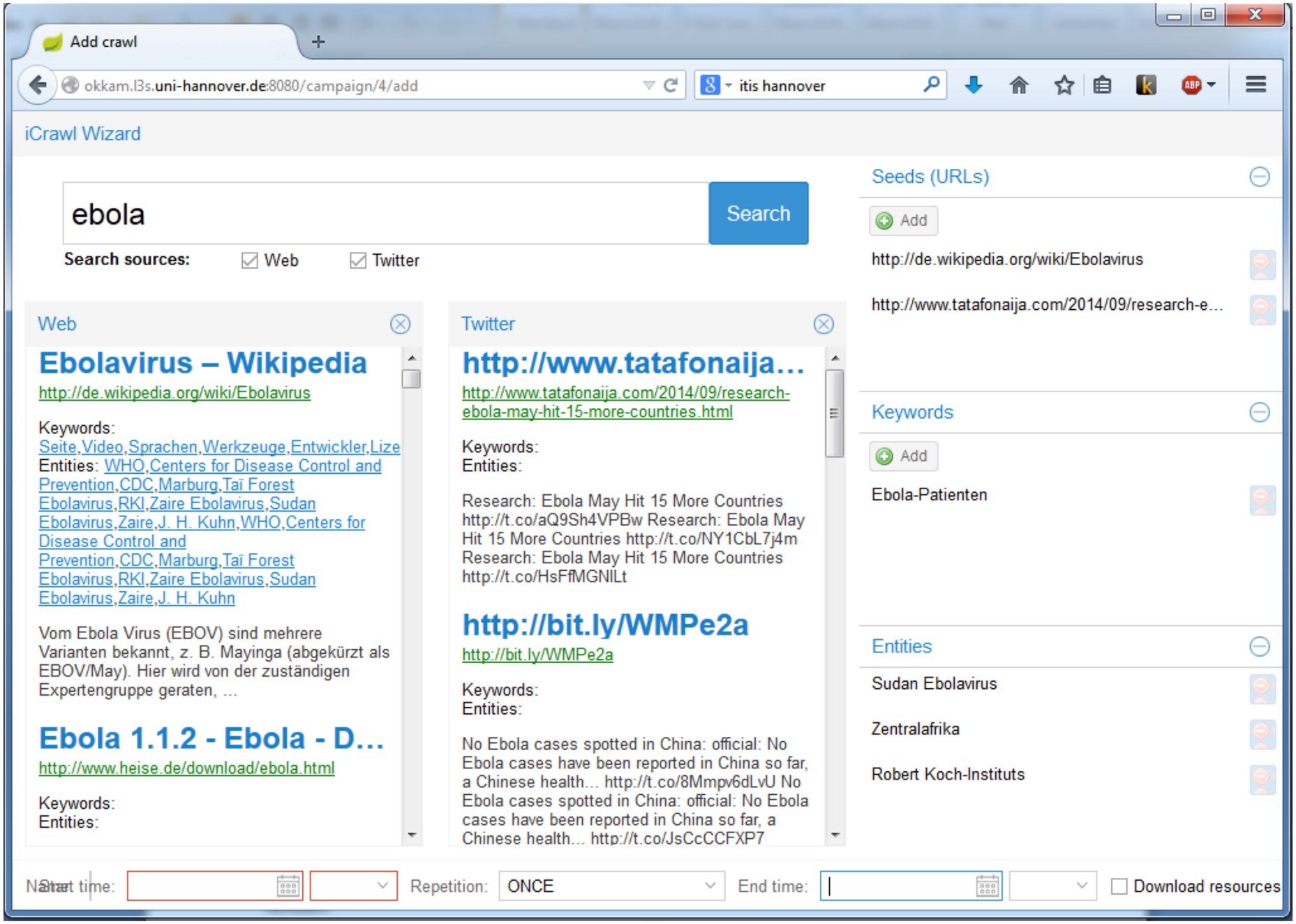
Reproducible

- Documentation of entire crawl process



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

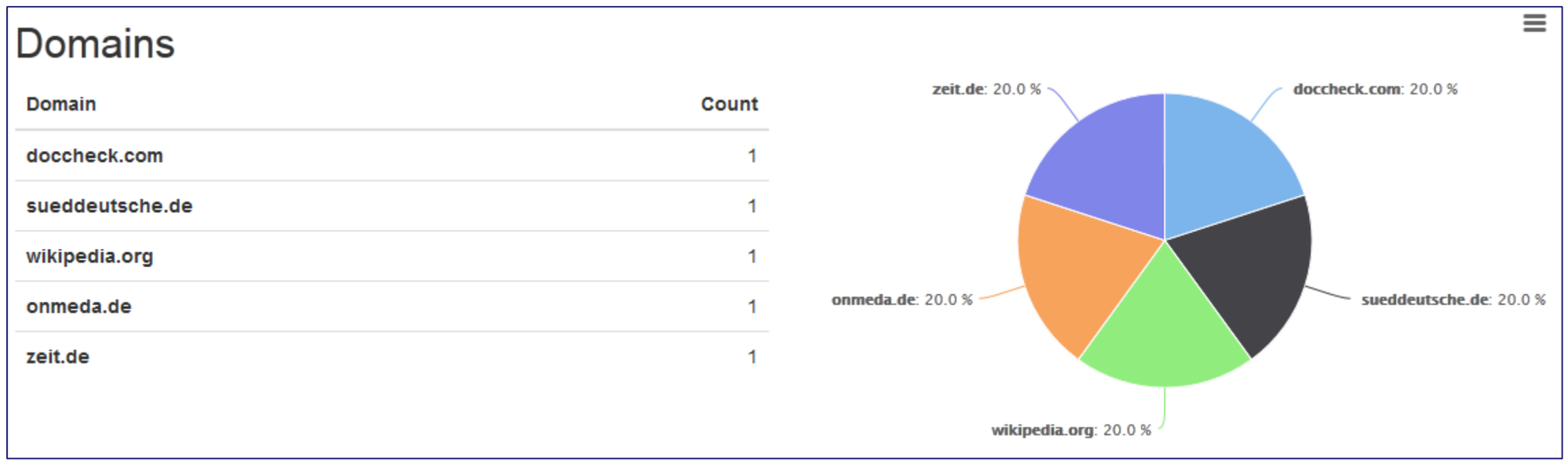
Crawl Specification Wizard



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



Full Information about Crawled Pages

iCrawl Campaigns Status Settings About

Ebola: Crawl 1 Edit History Statistics Data

Data for crawl “Ebola-News”

Selection FETCHED Filter

Title	Status	Fetched	URL	
http://alamandas.com/	fetched	15.10.2014 10:05:54	http://alamandas.com/	Wayback
http://ajax.aspnetcdn.com/ajax/jQuery/jquery-1.6.1.min.js	fetched	15.10.2014 10:05:52	http://ajax.aspnetcdn.com/ajax/jQuery/jquery-1.6.1.min.js	Wayback
http://imworld.aufeminin.com/dossiers/acc1_29869 /accNoTexte686x400a488591.jpg	fetched	15.10.2014 09:21:39	http://imworld.aufeminin.com/dossiers/acc1_29869 /accNoTexte686x400a488591.jpg	Wayback
http://imworld.aufeminin.com/manage/bloc/D20140915 /onmeda-085353_L.jpg	fetched	15.10.2014 09:21:49	http://imworld.aufeminin.com/manage/bloc/D20140915 /onmeda-085353_L.jpg	Wayback
http://admin.brightcove.com/js/BrightcoveExperiences.js	fetched	15.10.2014 09:21:39	http://admin.brightcove.com/js/BrightcoveExperiences.js	Wayback

iCrawl Campaigns Status

Ebola: Crawl 1 Edit History Statistics Data

Resource http://alamandas.com/

__g__dirty	java.nio.HeapByteBuffer[pos=0 lim=4 cap=4]
baseUrl	http://alamandas.com/
status	2
fetchTime	1413360354110
prevFetchTime	1410766562711
fetchInterval	2592000
retriesSinceFetch	0
modifiedTime	0
prevModifiedTime	0
protocolStatus	SUCCESS, args=[]
content	java.nio.HeapByteBuffer[pos=0 lim=61749 cap=61749]
contentType	application/xhtml+xml
prevSignature	–
signature	–
title	–
text	–
parseStatus	–
score	0.0
reprUrl	–
headers	

Key	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://wp.me/P466gB-78>; rel=shortlink

Wayback Machine View of Crawled Pages

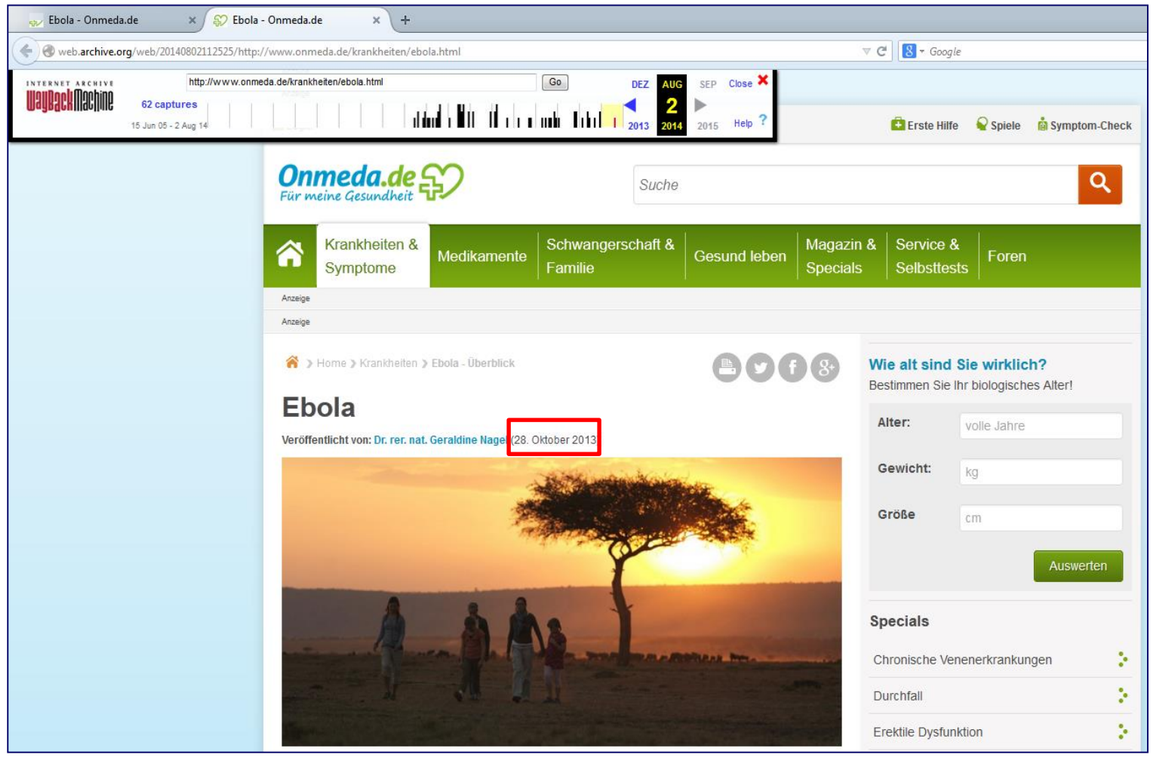
iCrawl capture: 15 September 2014

Updated: 13 August 2014

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

IA capture: 2 August 2014

Updated: 28 October 2013



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

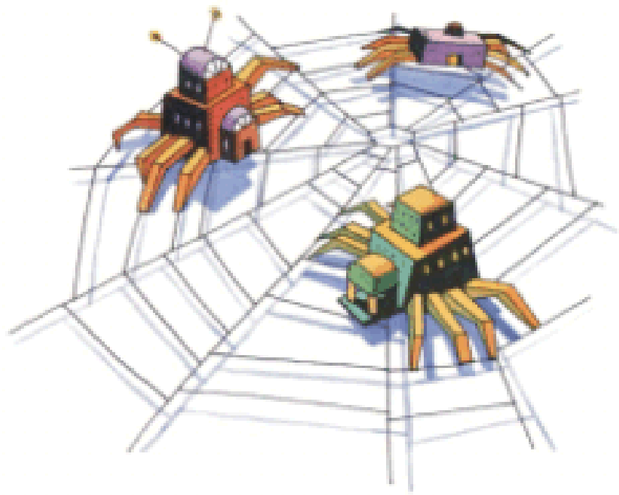
State of the art selection: by host, by domain

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

Thank You!

Dr. Elena Demidova
Forschungszentrum L3S
Appelstrasse 9a
30167 Hannover

E-Mail: demidova@L3S.de
Telefon: 0511 – 762 17732
Telefax: 0511 – 762 17779