Beyond the ‘Eternal Now’

preserving & accessing our past via Web archives

Hugo Huurdeman huurdeman@uva.nl

Digital Heritage - 8 March 2016



Introduction

- Our daily lives increasingly take place on the web
- However, web content is dynamic and ephemeral
- **Future understanding of our time is at risk!**
- Prevent our *future* heritage from disappearing



"if we lose our past, we will live in an **Orwellian world of the perpetual present,** where anybody who is in control of the information that is out there, will be able to say **what is true and what is not**"

Brewster Kahle, founder Internet Archive

DE VOLKSKRANT
DONDERDAG 4 OKTOBER 2012

Facebook-rellen in Haren



Posters die op de Facebook-pagina Project X-Haren verschenen, waarop werd opgeroepen om op 21 september naar het Groningse dorp te komen.





Faceb



et mu-
ek!',
ur van
n elkaar
ee te
n tekort
bier
uit
iem op
dering,
en on-
open-

ive ge-
den.
ijn, we-
at ze
zuele
n: met
naar Ha-
rij-
ord-
avond.

niet naar Haren mocht. 'Kom naar Haren grootste feest ooit', aldus Ruben

men: 'Loopt nog erger uit de hand dan in de film.'

12.500 Facebook-berichten achterhaald

Voor dit artikel kreeg *de Volkskrant* toegang tot een databank met 12.500 Facebook-berichten over Project X-Haren die is aangelegd door internetbedrijf Clockwork uit Groningen. De afgelopen dagen is het systeem aangevuld met 700 duizend tweets. Met de databank wil Clockwork voorkomen dat sociale-mediaberichten over de rellen in Haren van internet verdwijnen.

'Oorspronkelijk stonden deze gegevens allemaal online', zegt Rienk Prinsen, programmeur bij Clockwork. 'Op Facebook kun je een eind terugbladeren, maar die site is daar niet voor gemaakt. Er zit een grens aan wat je kunt terughalen. Ons systeem, Pulse, brengt de berichten weer allemaal in kaart.' Om de gegevens op te vragen, maakte Clockwork gebruik van speciale software-ingangen die Facebook en Twitter ter beschikking stellen aan programmeurs. Dat gebeurde onder meer op verzoek van de Groninger ict-ondernemer Ritzo ten Cate, die de Facebook-berichten over Project X-Haren onderzocht, maar niet verder kon terugbladeren dan 20 september – de dag voor de rellen. Donderdag wordt de databank gebruikt tijdens een 'hackersevenement' dat plaatsvindt op de Universiteit Utrecht. 'Het was bedoeld als alternatief voor de commissie-Cohen', zegt organisator Thomas Boeschoten. 'Van Job Cohen wordt wel gezegd dat hij geen verstand heeft van sociale media. We wilden de commissie een beetje helpen.' 'Echte' hackers zijn donderdag niet in Utrecht. Hoogleraar communicatiewetenschappen Jan van Dijk, lid van de commissie-Cohen, stuurt een medewerker langs. Verder doen mee: universitair onderzoekers uit Utrecht en Amsterdam, een rijksambtenaar, datajournalisten van nu.nl en *NRC Handelsblad*. Op een liveblog (tweetonderzoek.nl) worden de resultaten gepubliceerd.



DDS

PERSOONLIJK

POST

HOE ZUM HEER

A-Z

INDEX

PLATTEGROND

AANMELDEN ALS BEWONER

DIGI'S ADVERTENTIE KRANT

LAATSTE NIEUWS

ENGLISH INFORMATION

SPEL REGELS

PRIJS INFORMATIE

DDS KANTOOR

HELP DESK SERVICE CENTER

DDS CENTRAAL

D N P X

zijwegen

kiosk

cafe

discussie

POSTBANK


Main
Neighborhoods
Featured Content
Members
WELCOME TO OUR COMMUNITY
GeoCities
Saturday, 14-Feb-98 19:06:49 PST
ONE MILLION
MATCH YOUR INTEREST TO A NEIGHBOR
FREE HOME PAGES AND E-MAIL
ARTS
AUTOS
BUSINESS
COMPUTERS
CULTURE
EDUCATION
ENTERTAINMENT
ENVIRONMENT
FAMILY
FASHION
FOOD
GAMES
GAY & LESBIAN
GOVERNMENT
HEALTH
KIDS
MUSIC
PEOPLE
RECREATIO
SCIENCE FI
GEOCITIES
Your home on the web!

Internet Atrocity! GeoCities' Demise Erases Web History

By Dan Fletcher | Monday, Nov. 09, 2009

Like 5 Tweet 1 +1 18 Share Read Later

When Yahoo! switched off the servers for GeoCities, the Web posting service, on Oct. 27, some 7 million of the Internet's first websites went dark forever. The bulk of these were people's personal home pages, which were pulled offline with no backup and no permanent record of those users' frenetic early forays online.



The GeoCities home page in 1998

Now a ragtag effort by several groups of budding computer historians is feverishly — and angrily — trying to bring as much as they can back online. Founded in 1995 and bought by Yahoo! four years later, GeoCities had become a relic, and not a

RELATED

Hyves

Zoek vrienden, games en meer...

Gebruikersnaam

Wachtwoord

Inloggen

Onthoud mij

Wachtwoord vergeten?

Gebruikersnaam | Wachtwoord | Inloggen

Onthoud mij | Wachtwoord vergeten?

Hyves | Zoek vrienden, games en meer...

Hyves wordt een Games netwerk



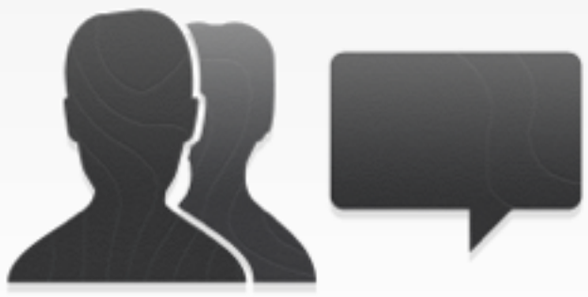
We kunnen ons voorstellen dat dit een hoop vragen bij je oproept.

Lees meer op Hyves

Meer info

Of lees het persbericht

Ontmoet je vrienden



Vind je vrienden terug op Hyves of maak nieuwe vrienden. Doe dit wel veilig, bekijk hier hoe we jouw veiligheid waarborgen en hoe je dit zelf kunt doen.

Gamen op Hyves



Houd jij van boerderij-, puzzel- of multiplayer-games? Op Hyves vind je ze allemaal! Speel tegen je vrienden en probeer elkaars highscores te verbeteren.

Deel foto's, video's en muziek



Respecteer foto's van je vrienden en deel zelf je leukste video's. Ben je artiest? Hyves is ook de p voor je eigen muziek.

Ik wil een profiel

Ik wil gamen

Ik wil foto's uploaden

WEBLOG.NL

Sanoma Media heeft Weblog.nl per 1 maart 2013 overgedragen aan WordPress.com

Exportbestand

Was jij de eigenaar van een weblog op Weblog.nl en wil je je blog voortzetten bij WordPress.com of een andere blogaanbieder? Dan heb je daarvoor een exportbestand nodig van je blog. Beschik jij nog niet over een exportbestand? Neem dan contact op met de Helpdesk.

Mis je afbeeldingen?

Het kan zijn dat je afbeeldingen mist op je nieuwe blog, terwijl je al eerder je weblog had voortgezet bij WordPress.com of een andere blog aanbieder. In dat geval verwijzen de URL's voor je mediabestanden naar je oude webloglocatie. Je kunt dit probleem oplossen door een nieuw exportbestand aan te vragen bij de Helpdesk en deze opnieuw te importeren bij WordPress.com of bij een andere blogaanbieder.

Problemen bij het importeren

Heb je problemen bij het importeren van jouw nieuwe exportbestand in WordPress.com? Neem dan contact op met support@wordpress.com.

1. What is web archiving?

1. What is web archiving?

- "the process of **collecting** portions of the World Wide Web, **preserving** the collections in an archival format, and then **serving** the archives for access and use" (IIPC, 2016)

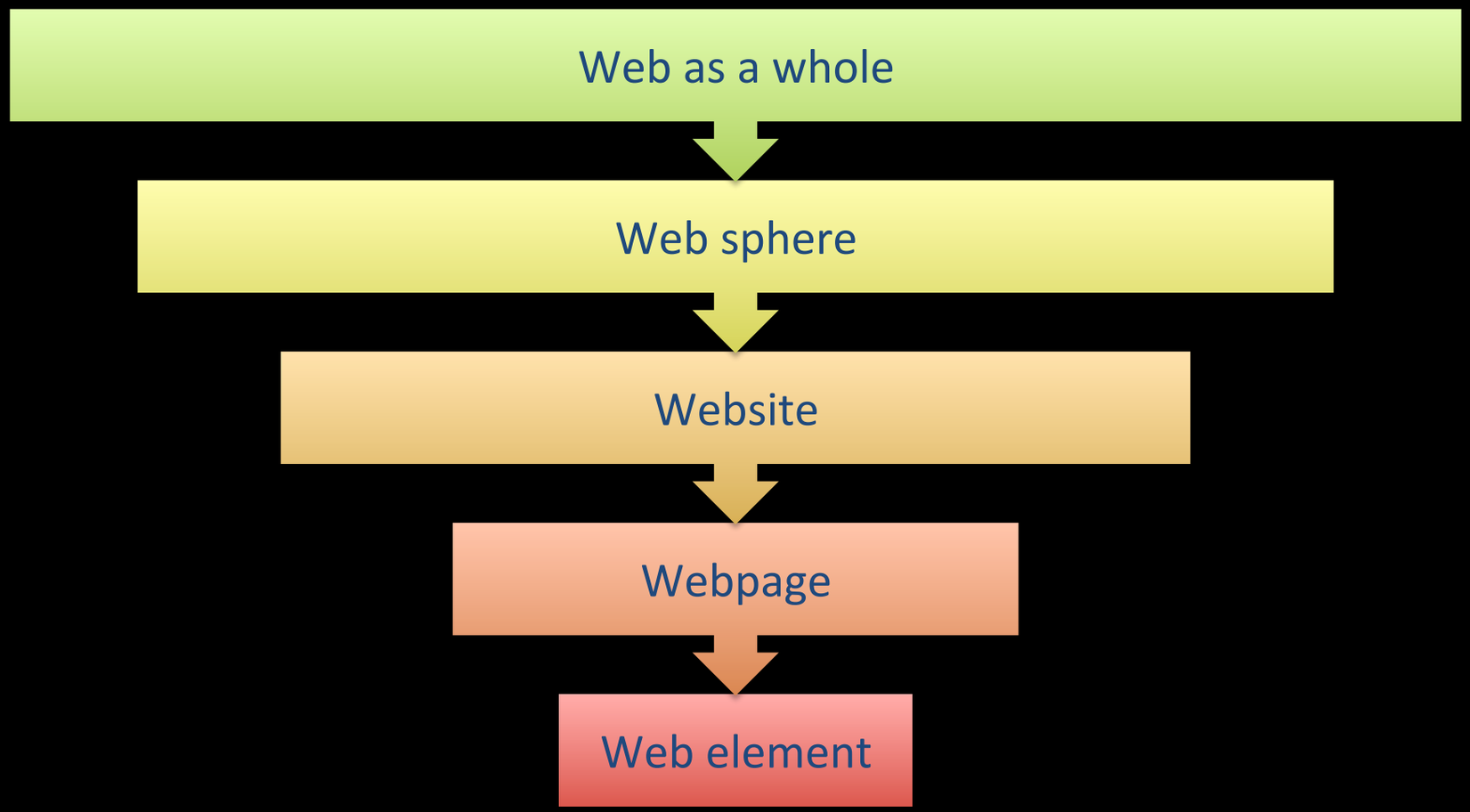
1. What is web archiving?

- "the process of **collecting** portions of the World Wide Web, **preserving** the collections in an archival format, and then **serving** the archives for access and use" (IIPC, 2016)
- "the **selection**, **collection**, **storage**, **retrieval**, and **maintenance** of the integrity of Web resources (Ball, 2010)

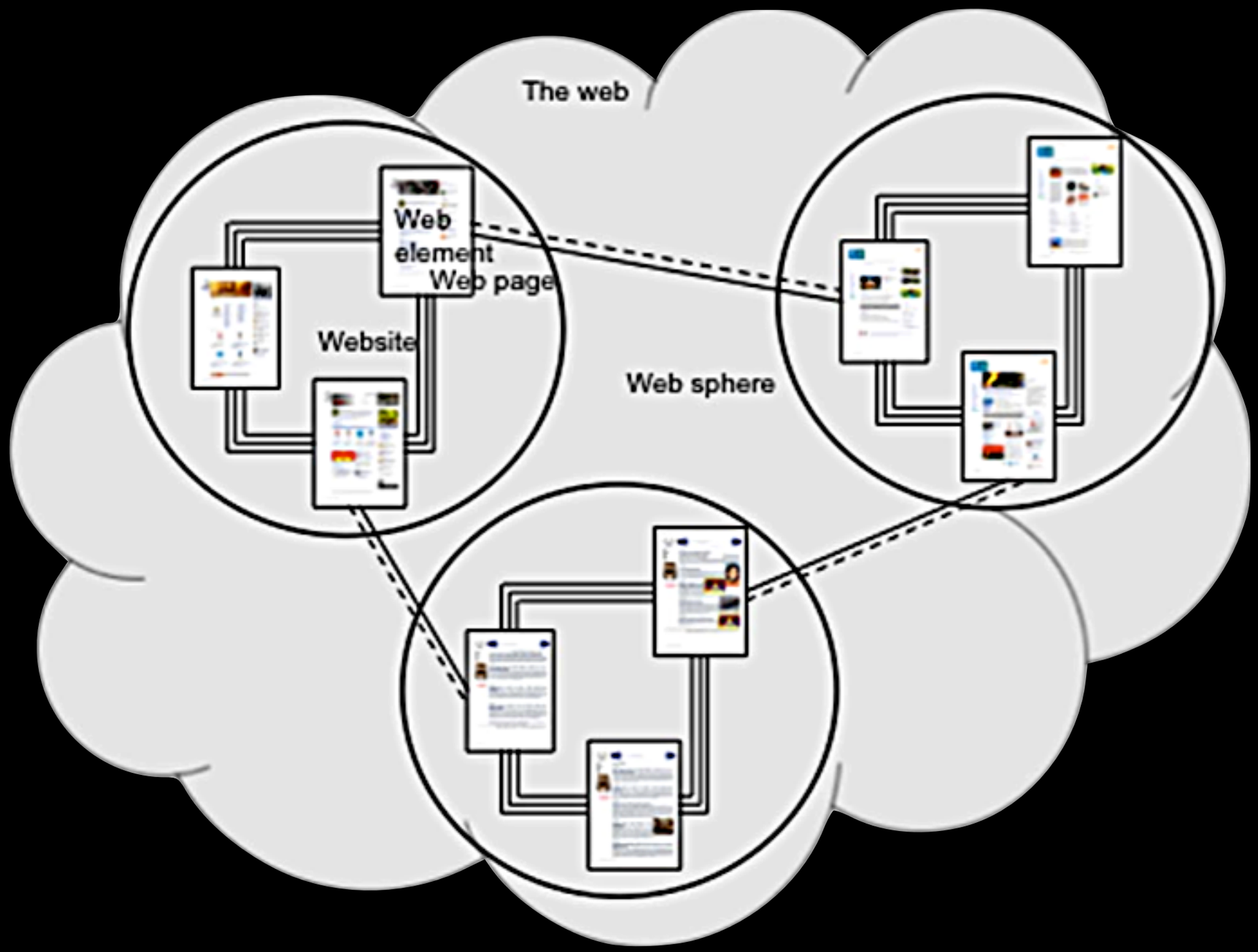
1. What is web archiving?

- "the process of **collecting** portions of the World Wide Web, **preserving** the collections in an archival format, and then **serving** the archives for access and use" (IIPC, 2016)
- "the **selection**, **collection**, **storage**, **retrieval**, and **maintenance** of the integrity of Web resources (Ball, 2010)
- "any form of **deliberate** and **purposive preserving** of **web material**" (Brügger, 2011, p. 25)

Web material?



Five analytical objects (Brügger 2009)



Analytical web objects (Brügger, 2012)

2. Who archives the Web?

2. Who archives the Web?

- National libraries & archives, research institutions, researchers, the *crowd,* …

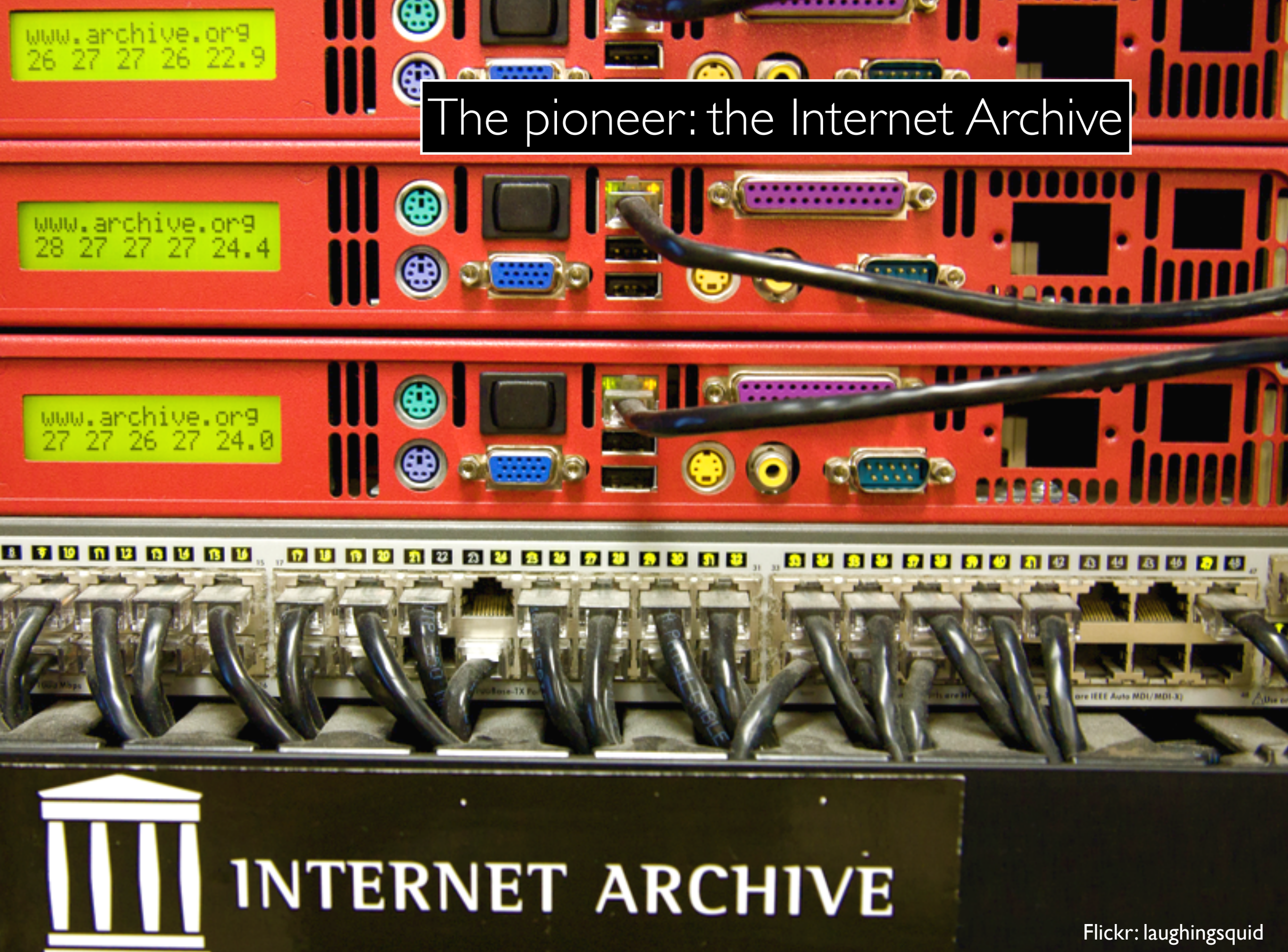
2. Who archives the Web?

- National libraries & archives, research institutions, researchers, the *crowd,* …
  - *collection building, preservation, research, legislation*

2. Who archives the Web?

- National libraries & archives, research institutions, researchers, the *crowd,* …
  - *collection building, preservation, research, legislation*
  - "macro" vs "micro" archiving (Brügger, 2009)

The pioneer: the Internet Archive

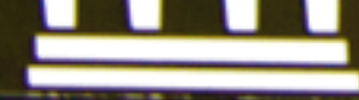


Flickr: laughingsquid

The pioneer: the Internet Archive

Since 1996:

- "mass archiving" the web



Flickr: laughingsquid

Who archives the Web: **National Libraries & Archives**



Flickr: Kungliga biblioteket

Who archives the Web: **National Libraries & Archives**



Approaches:

- **domain harvesting** (top domain of country), *and/or*
- **selective harvesting** (based on formal criteria), *and/or*
- **event harvesting** (related to current events)

Flickr: Kungliga biblioteket

HARVARD UNIVERSITY LIBRARY
Web Archive Collection Service

Harvard's Web Archive Collection Service (WA
central infrastructure for the capture, manage
sites for long-term archiving. More About WA







Universiteit Leiden

Home
DACHS Heidelberg

Citation Repository

Projects:
- Poetry
- SARS
- Tongzhi Culture

E-publications

Topical Issues

NEW!

萊頓大學漢學研究院

Talen en culturen van China / Sinologisch Instituut
Faculteit der Letteren

home contact help

Digital Archive for Chinese Studies
Leiden Division

Introduction: The Digital Archive for Chinese Studies (DACHS)

Our aim is a
Chinese Stu
reflected by

DACHS exis
at the Unive
joined the pi

For more ge
the introduc
Newsletter #
the XVth E

Collection

The collectic
snapshots c
political sut

University of Michigan Web Archives
Bentley Historical Library, University of Michigan

Home | About | Site List | Search | Help | Contact Us

Refine site list

lookup by site name

Go Clear

Site list by topic:

Administration
Architecture & Urban Planning
Athletics
College of Engineering

1000 Pitches Web Archives (University of Michigan)

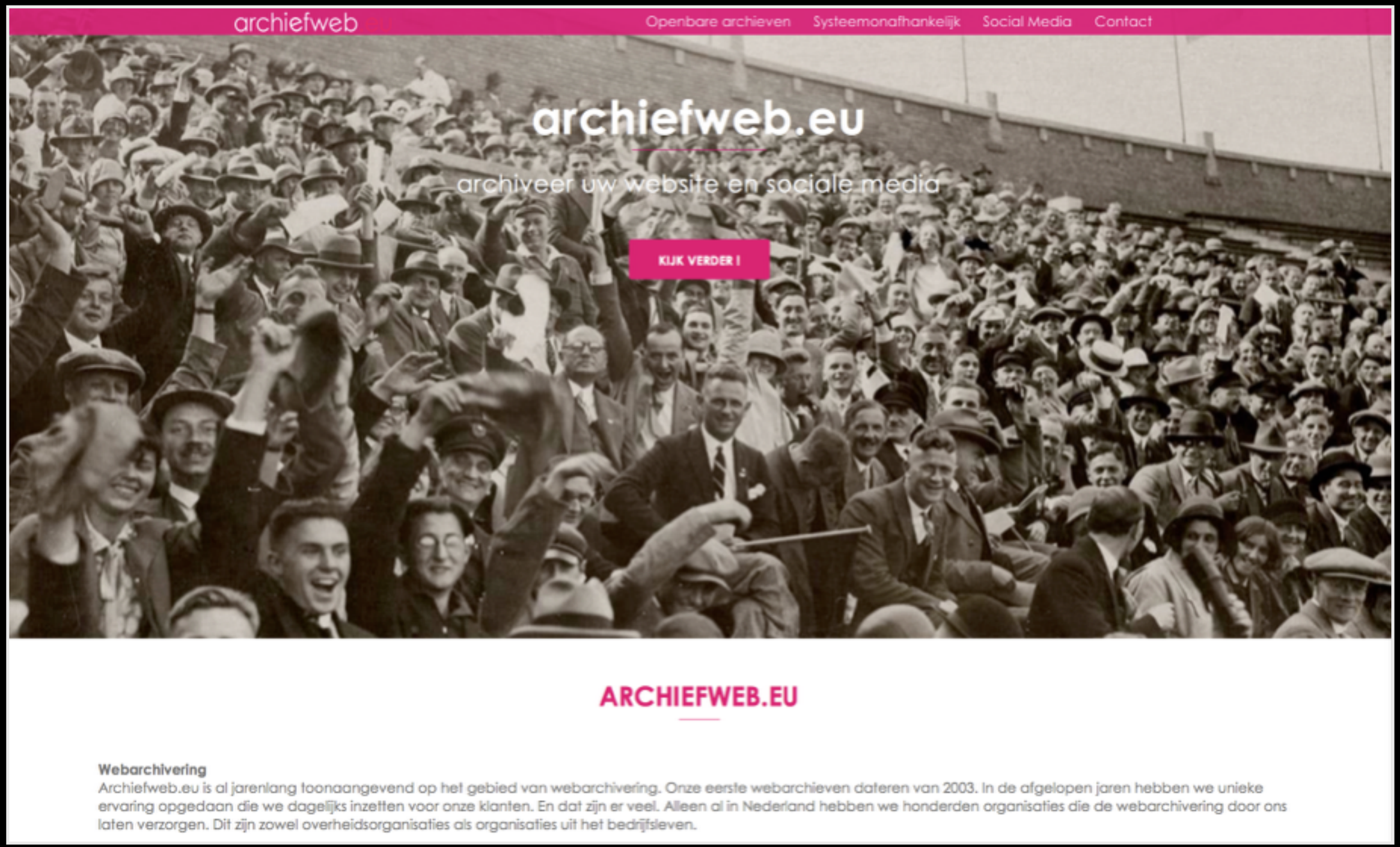
3D Lab Web Archives (University of Michigan)

A Alfred Taubman Medical Research Institute Web

Academic Support Services Web Archives (Universi

Accreditation Process: 2010 Web Archives (Univers

Who archives the Web: **Research institutions and researchers**



Who archives the Web: **Companies**

e.g. in the context of archive laws

Main Page

navigation

- Main page
- Community portal
- Current events
- Recent changes
- Random page
- Help

search

Go Search

toolbox

- What links here
- Related changes
- Special pages
- Printable version
- Permanent link

HISTORY IS OUR FUTURE



And we've been trashing our history

Archive Team is a loose collective of rogue archivists, programmers, writers and loudmouths dedicated to saving our digital heritage. Since 2009 this variant force of nature has caught wind of shutdowns, shutoffs, mergers, and plain old deletions - and done our best to save the history before it's lost forever. Along the way, we've gotten attention, resistance, press and discussion, but most importantly, we've gotten the message out: **IT DOESN'T HAVE TO BE THIS WAY**.

This website is intended to be an offloading point and information depot for a number of archiving projects, all related to saving websites or data that is in danger of being lost. Besides serving as a hub for team-based pulling down and mirroring of data, this site will provide advice on managing your own data and rescuing it from the brink of destruction.

Currently Active Projects (Get Involved Here!)

ARCHIVE TIME

What is What

- Who We Are and how you can join our cause!
- Deathwatch is where we keep track of sites that are sickly, dying or dead.
- Fire Drill is where we keep track of sites that seem fine but a lot depends on them.
- Projects is a comprehensive list of AT endeavors.
- Philosophy describes the ideas underpinning our work.
- Archives

Some Starting Points

- The Introduction is an overview of basic archiving methods.
- Why Back Up? Because they don't care about you.
- Back Up your Facebook Data Learn how to liberate your personal data from Facebook.

Who archives the Web: **'crowdsourced' archiving**

Log in / create account

page | discussion | view source | history

Main Page

GeoCities

GeoCities

URL	http://geocities.com
Project status	Offline
Archiving status	Saved!, but not yet in Wayback machine
Project source	Unknown
Project tracker	Unknown
IRC channel	#archiveteam

FUTURE



s caught wind of shutdowns,
ain old deletions - and done
ry before it's lost forever.
ten attention, resistance, press
importantly, we've gotten the
T HAVE TO BE THIS WAY.

to be an offloading point and
umber of archiving projects, all
s or data that is in danger of
ng as a hub for team-based
g of data, this site will provide
own data and rescuing it from

Currently Active Projects (Get Involved Here!)

ARCHIVE TIME

What is What

- Who We Are and how you can join our cause!
- Deathwatch is where we keep track of sites that are sickly, dying or dead.
- Fire Drill is where we keep track of sites that seem fine but a lot depends on them.
- Projects is a comprehensive list of AT endeavors.
- Philosophy describes the ideas underpinning our work.
- Archives

Some Starting Points

- The Introduction is an overview of basic archiving methods.
- Why Back Up? Because they don't care about you.
- Back Up your Facebook Data Learn how to liberate your personal data from Facebook.

Who archives the Web: '**crowdsourced**' archiving

Log in / create account

page | discussion | view source | history

Main Page

GeoCities

URL	http://geocities.com
Project status	Offline
Archiving status	Saved!, but not yet in Wayback machine
Project source	Unknown
Project tracker	Unknown
IRC channel	#archiveteam

Hyves



URL	http://hyves.nl
Project status	Offline 2013-12-02[1]
Archiving status	Saved!
Project source	https://github.com/ArchiveTeam/hyves-grab
Project tracker	http://tracker.archiveteam.org/hyves/
IRC channel	#angerthehyve

ARCHIVE TIME

u can join our cause!

keep track of sites that are sickly, dying or

ep track of sites that seem fine but a lot

sive list of AT endeavors.

ideas underpinning our work.

Currently Active Projects (Get Involved Here!)

- The Introduction is an overview of basic archiving methods.
- Why Back Up? Because they don't care about you.
- Back Up your Facebook Data Learn how to liberate your personal data from Facebook.

Who archives the Web: **'crowdsourced' archiving**

Log in / create account

page | discussion | view source | history

Main Page

GeoCities



URL	http://geocities.com
Project status	Offline
Archiving status	Saved!, but not yet in Wayback machine
Project source	Unknown
Project tracker	Unknown
IRC channel	#archiveteam

Hyves



URL	http://hyves.nl
Project status	Offline 2013-12-02[1]
Archiving status	Saved!
Project source	https://github.com/ArchiveTeam/hyves-grab
Project tracker	http://tracker.archiveteam.org/hyves/
IRC channel	#angerthehyve

Weblog.nl



URL	http://weblog.nl
Project status	Offline
Archiving status	Saved! - archive - lookup
Project source	https://github.com/ArchiveTeam/weblognl-grab
Project tracker	http://tracker.archiveteam.org/weblognl
IRC channel	#archiveteam

Currently Active Projects (Get Involved Here!)

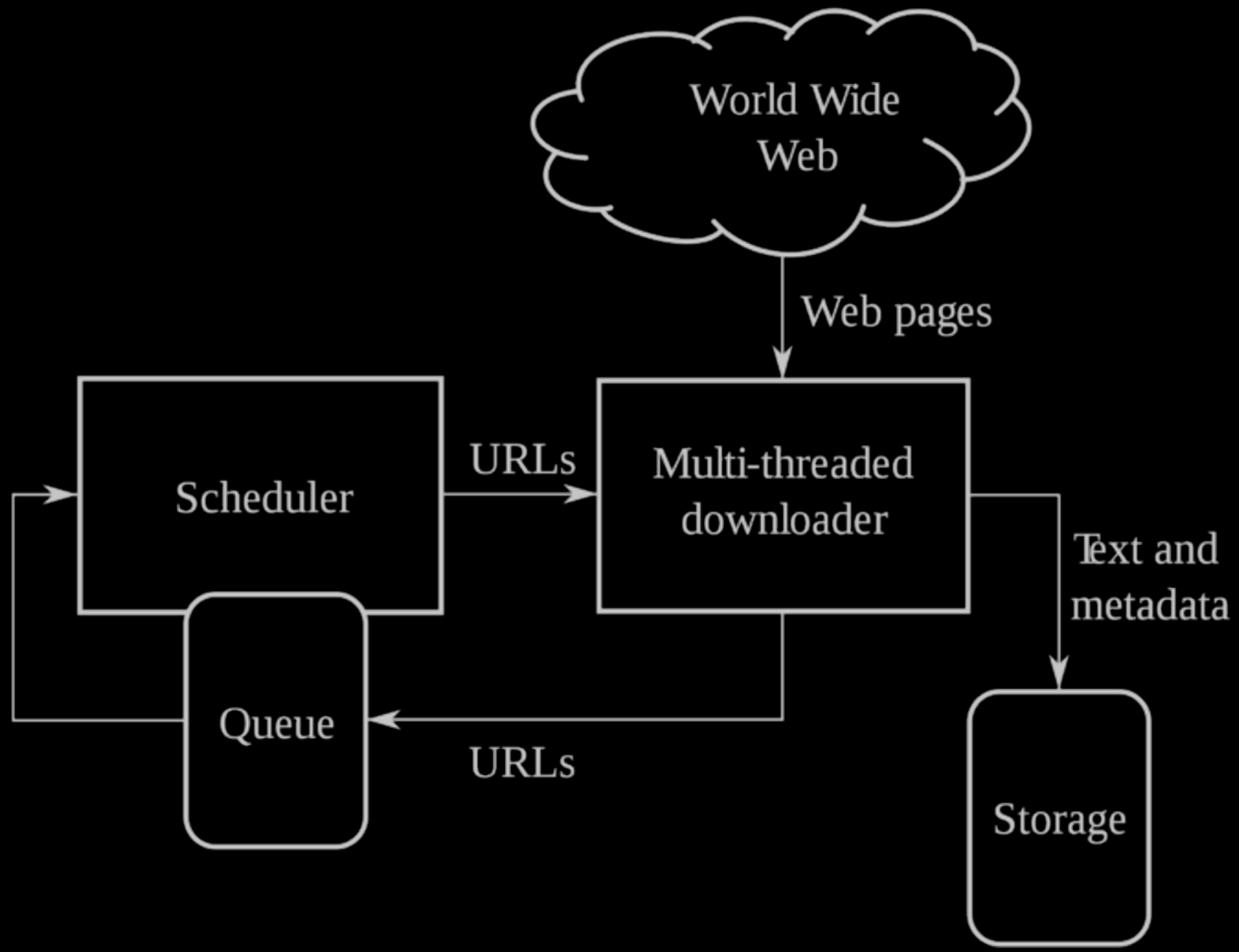
- Why Back Up? Because they don't care about you.
- Back Up your Facebook Data Learn how to liberate your personal data from Facebook.

Who archives the Web: '**crowdsourced**' archiving

3. So, problem solved..?

3. So, problem solved..?

…Not exactly :(


World Wide Web
Web pages
Scheduler
URLs
Multi-threaded downloader
Text and metadata
Queue
URLs
Storage

Limitations

- The Web is archived using *crawlers*
- However, crawlers **cannot capture everything**

Limitations

- The Web is archived using *crawlers*
- However, crawlers **cannot capture everything**
- Limitations in:
  - data formats (Flash, interactive forms, ..)
  - crawl frequency and depth
  - (legal) permissions

INTERNET ARCHIVE
WayBackMachine
http://www.omroep.nl/radio3/mainek.htm
Go
2 captures
7 Oct 00 - 29 Oct 00
SEP OCT
7
1999 2000



3FM LIVE ON THE NET

3FM

5 oktober Skik-dag

ACTIVE RADIO

HOME | PROGRAM-MA'S | I T V | LINKS | MUZIEK | E-MAIL | SHIT HAPPENS

INTERNET ARCHIVE WayBackMachine

http://www.omroep.nl/radio3/mainek.htm Go

2 captures
7 Oct 00 - 29 Oct 00

SEP OCT
1999 7 2000



3FM LIVE ON THE NET

3FM

5 oktober Skik-dag

> Missing images

ACTIVE RADIO

HOME | PROGRAM-MA'S | I T V | LINKS | MUZIEK | E-MAIL | SHIT HAPPENS

INTERNET ARCHIVE WayBackMachine

http://www.omroep.nl/radio3/mainek.htm Go

SEP OCT

1999 7 2000

2 captures
7 Oct 00 - 29 Oct 00

> few crawls



3FM LIVE ON THE NET

5 oktober Skik-dag

> Missing images

ACTIVE RADIO

HOME | PROGRAM-MA'S | I T V | LINKS | MUZIEK | E-MAIL | SHIT HAPPENS

INTERNET ARCHIVE WayBackMachine

http://www.omroep.nl/radio3/mainek.htm Go

SEP OCT

1999 7 2000

2 captures
7 Oct 00 - 29 Oct 00

> few crawls



3FM LIVE ON THE NET

5 oktober Skik-dag

> Missing images

> Links do not work

ACTIVE RADIO

HOME | PROGRAM-MA'S | I T V | LINKS | MUZIEK | E-MAIL | SHIT HAPPENS

INTERNET ARCHIVE WayBackMachine

http://www.omroep.nl/radio3/mainek.htm Go

2 captures
7 Oct 00 - 29 Oct 00

SEP OCT
7
1999 2000

> few crawls



3FM LIVE ON THE NET

5 oktober Skik-dag

> Missing images

> Links do not work

ACTIVE RADIO

HOME



INTERNET ARCHIVE WayBackMachine

http://www.omroep.nl/radio3/mainek.htm Go

2 captures
7 Oct 00 - 29 Oct 00

SEP OCT 7 1999 2000

> few crawls



3FM LIVE ON THE NET

3 FM

5 oktober Skik-dag

> Missing images

> Links do not work

ACTIVE RADIO

HOME

INTERNET ARCHIVE WayBackMachine

http://www.omroep.nl/tros/homepage/nonflash.html Go

48 captures
28 Nov 99 - 18 Jun 15

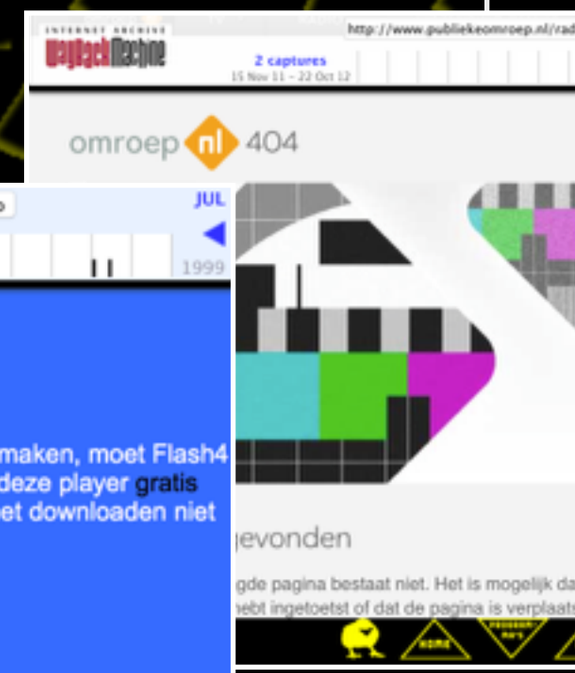
JUL 1999

FLASH

Om het **TROSWEB** goed te kunnen bekijken en van alle mogelijkheiden gebruik te kunnen maken, moet Flash4 geïnstalleerd zijn op uw computer. Dat is bij u nog niet het geval. Op deze pagina kunt u deze player gratis downloaden door op de "Get Flash"-knop te drukken. Omdat het een kleine player is, zal het downloaden niet lang duren.

GET FLASH

Na het downloaden wordt u precies verteld wat u moet doen. Succes !



Limitations

- The Web is archived using *crawlers*
- However, crawlers **cannot capture everything**
- Limitations in:
  - data formats (Flash, interactive forms, ..)
  - crawl frequency and depth
  - (legal) permissions
- ***As a result, web archives are incomplete at several levels***

4. Still, web archives have gathered **petabytes** of valuable information

4. Still, web archives have gathered **petabytes** of valuable information

So, they must be used on a **massive scale**, right?

4. Still, web archives have gathered **petabytes** of valuable information

So, they must be used on a **massive scale**, right?

Well, …

BRITISH LIBRARY

How often are web archives used?

- Archiving institutions' focus on data collection, not usage
- 19 of 29 IIPC members' archives (listed on website) have full or partial online access, often permission-based
- Large scale national web archives have restricted access – dark archives
  - eg Danish National Web Archive, over 280TB
    - online access for researchers with PhD or higher level
    - 20 users since 2005
- "Document-centric" access methods
- No agreed way of calculating / benchmarking access statistics
- Little evidence of scholarly use of web archives, making it difficult to understand requirements

(From: Hockx-Yu, Web Archiving and Scholarly Use of Web Archives, 2013)





- A big gap exists between the **large potential research community** addressed by web archives & the **small group actually using them** (Dougherty & Meyer, 2014)

Non-use: some reasons:

- Legal restrictions preventing access
- Data quantity and quality
- Lack of appropriate **access methods**



- There is a wide gap between **researchers needing appropriate datasets** and **archivists who have the expertise to build collections and tools** (Dougherty & Meyer, 2014)

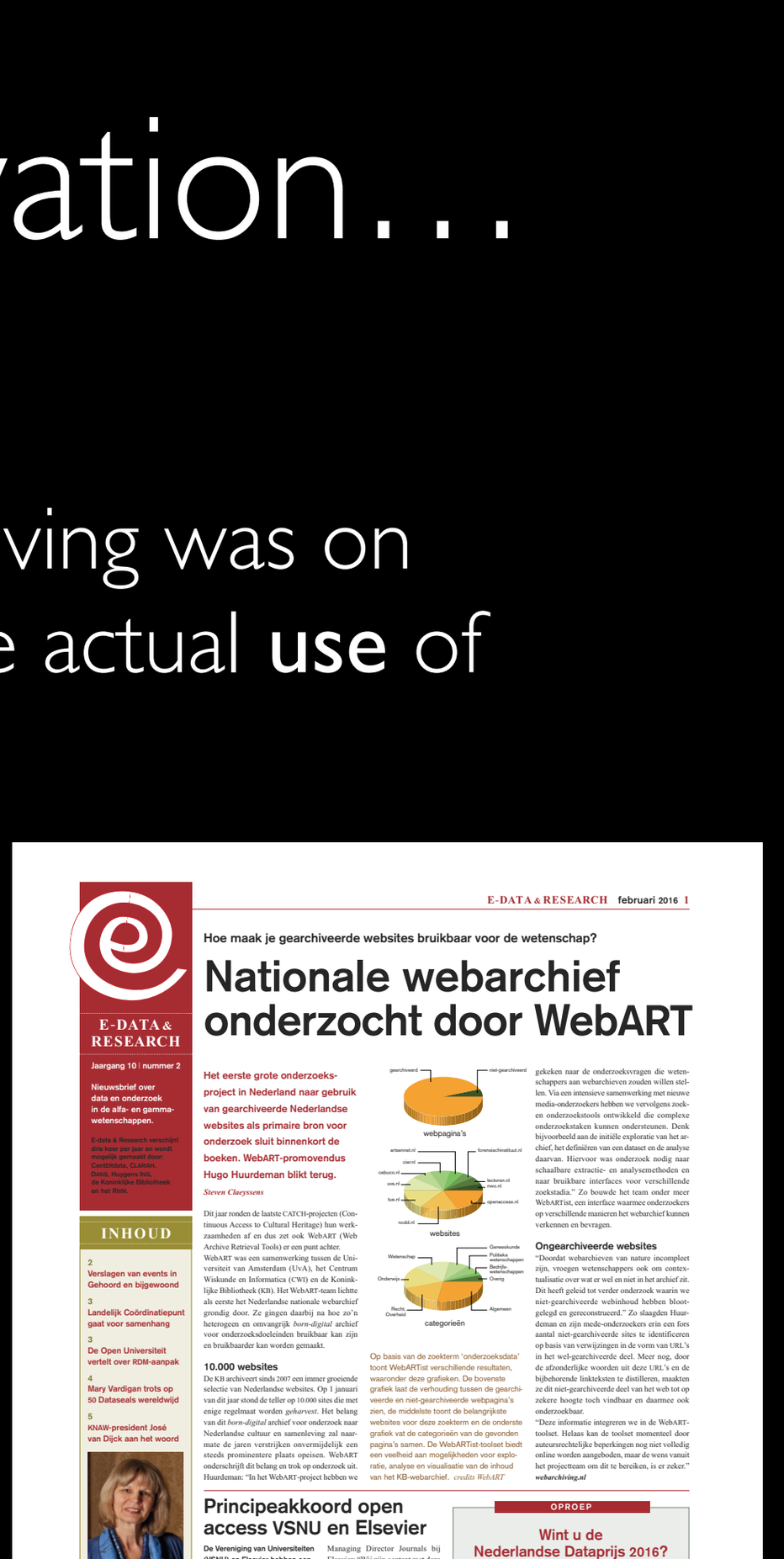
The *mission* of a web archive



Beyond preservation…

- The initial focus of web archiving was on **preservation,** but less on the actual **use** of the archived material

- The WebART project, a Dutch national research project, aims to improve upon this situation



E-DATA & RESEARCH

Jaargang 10 | nummer 2

Nieuwsbrief over data en onderzoek in de alfa- en gamma-wetenschappen.

E-data & Research verschijnt drie keer per jaar en wordt mogelijk gemaakt door: CentERdata, CLARIAH, DANS, Huygens ING, de Koninklijke Bibliotheek en het RIVM.

INHOUD

2 Verslagen van events in Gehoord en bijgewoond

3 Landelijk Coördinatiepunt gaat voor samenhang

3 De Open Universiteit vertelt over RDM-aanpak

4 Mary Vardigan trots op 50 Datasealss wereldwijd

5 KNAW-president José van Dijck aan het woord

E-DATA & RESEARCH februari 2016 1

Hoe maak je gearchiveerde websites bruikbaar voor de wetenschap?

Nationale webarchief onderzocht door WebART

Het eerste grote onderzoeksproject in Nederland naar gebruik van gearchiveerde Nederlandse websites als primaire bron voor onderzoek sluit binnenkort de boeken. WebART-promovendus Hugo Huurdeman blikt terug.

Steven Claeyssens

Dit jaar ronden de laatste CATCH-projecten (Continuous Access to Cultural Heritage) hun werkzaamheden af en dus zet ook WebART (Web Archive Retrieval Tools) er een punt achter. WebART was een samenwerking tussen de Universiteit van Amsterdam (UvA), het Centrum Wiskunde en Informatica (CWI) en de Koninklijke Bibliotheek (KB). Het WebART-team lichtte als eerste het Nederlandse nationale webarchief grondig door. Ze gingen daarbij na hoe zo'n heterogeen en omvangrijk *born-digital* archief voor onderzoeksdoeleinden bruikbaar kan zijn en bruikbaarder kan worden gemaakt.

10.000 websites

De KB archiveert sinds 2007 een immer groeiende selectie van Nederlandse websites. Op 1 januari van dit jaar stond de teller op 10.000 sites die met enige regelmaat worden *geharvest*. Het belang van dit *born-digital* archief voor onderzoek naar Nederlandse cultuur en samenleving zal naarmate de jaren verstrijken onvermijdelijk een steeds prominentere plaats opeisen. WebART onderschrijft dit belang en trok op onderzoek uit. Huurdeman: "In het WebART-project hebben we gekeken naar de onderzoeksvragen die wetenschappers aan webarchieven zouden willen stellen. Via een intensieve samenwerking met nieuwe media-onderzoekers hebben we vervolgens zoek- en onderzoekstools ontwikkeld die complexe onderzoekstaken kunnen ondersteunen. Denk bijvoorbeeld aan de initiële exploratie van het archief, het definiëren van een dataset en de analyse daarvan. Hiervoor was onderzoek nodig naar schaalbare extractie- en analysemethoden en naar bruikbare interfaces voor verschillende zoekstadia." Zo bouwde het team onder meer WebARTist, een interface waarmee onderzoekers op verschillende manieren het webarchief kunnen verkennen en bevragen.

Op basis van de zoekterm 'onderzoeksdata' toont WebARTist verschillende resultaten, waaronder deze grafieken. De bovenste grafiek laat de verhouding tussen de gearchiveerde en niet-gearchiveerde webpagina's zien, de middelste toont de belangrijkste websites voor deze zoekterm en de onderste grafiek vat de categorieën van de gevonden pagina's samen. De WebARTist-toolset biedt een veelheid aan mogelijkheden voor exploratie, analyse en visualisatie van de inhoud van het KB-webarchief. *credits WebART*

Ongearchiveerde websites

"Doordat webarchieven van nature incompleet zijn, vroegen wetenschappers ook om contextualistie over wat er wel en niet in het archief zit. Dit heeft geleid tot verder onderzoek waarin we niet-gearchiveerde webinhoud hebben blootgelegd en gereconstrueerd." Zo slaagden Huurdeman en zijn mede-onderzoekers erin een fors aantal niet-gearchiveerde sites te identificeren op basis van verwijzingen in de vorm van URL's in het wel-gearchiveerde deel. Meer nog, door de afzonderlijke woorden uit deze URL's en de bijbehorende linkteksten te distilleren, maakten ze dit niet-gearchiveerde deel van het web tot op zekere hoogte toch vindbaar en daarmee ook onderzoekbaar.

"Deze informatie integreren we in de WebART-toolset. Helaas kan de toolset momenteel door auteursrechtelijke beperkingen nog niet volledig online worden aangeboden, maar de wens vanuit het projectteam om dit te bereiken, is er zeker."

webarchiving.nl

Principeakkoord open access VSNU en Elsevier

De Vereniging van Universiteiten (VSNU) en Elsevier hebben een

Managing Director Journals bij

OPROEP

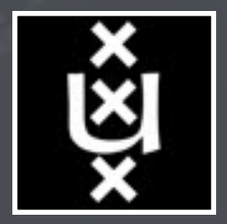
Wint u de Nederlandse Dataprijs 2016?

WebART project

Web Archive Retrieval Tools

Jaap Kamps, Richard Rogers, Arjen de Vries
Hildelies Balk, René Voorburg

Hugo Huurdeman, Thaer Sammar
Sanna Kumpulainen, Anat Ben-David







Flickr: LucViatour

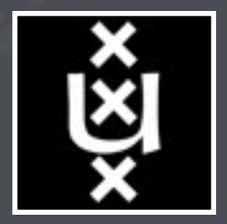
WebA[illegible]ect

We[illegible]Retrieval Tools

Supporting scholarly use of web archives

Jaap Kamps, Richard Rogers, Arjen de Vries
Hildelies Balk, René Voorburg

Hugo Huurdeman, Thaer Sammar
Sanna Kumpulainen, Anat Ben-David







Flickr: LucViatour

5. WebART & the KB



WebART

Web Archive Retrieval Tools

2012-2016

WebART
Web Archive Retrieval Tools

KB

Hildelies Balk

René Voorburg

Arjen de Vries

Jaap Kamps

Richard Rogers

CWI

Thaer Samar
PhD/programmer

Hugo Huurdeman
PhD researcher

Anat Ben-David ('12-'14)
Postdoc

WebART Goals

- Evaluating current curation and selection procedures of Web archives
- Getting insights into current use of Web archives
- Developing new methods and tools for research using Web archives

KB: Web archive since 2007

Selective approach

Statistics:

- 10,000+ websites
- 35,000+ harvests
- 16+ Terabyte

Flickr: koninklijkebibliotheek

KB-webarchief



In het KB-webarchief zijn honderden gearchiveerde websites vanaf 2007 te vinden.

Het KB-webarchief bevat een selectie Nederlandse websites die vanaf 2007 zijn verzameld door de KB. De sites worden één tot twee keer per jaar binnengehaald (geharvest) en duurzaam opgeslagen. Als u in het bezit bent van een KB-jaarpas kunt u het KB-webarchief in de leeszaal van de KB bekijken met behulp van de zogenaamde *Wayback Machine*.

Volop in ontwikkeling

Het KB-webarchief is nog volop in ontwikkeling en nog niet alle gearchiveerde sites zijn beschikbaar. De KB werkt in internationaal verband aan het verbeteren van de software voor het verwerven en tonen van de websites. In een latere fase zal het KB-webarchief ook online beschikbaar komen.

Ga naar het webarchief *(alleen toegankelijk in de leeszaal van de KB)*

KB-webarchief



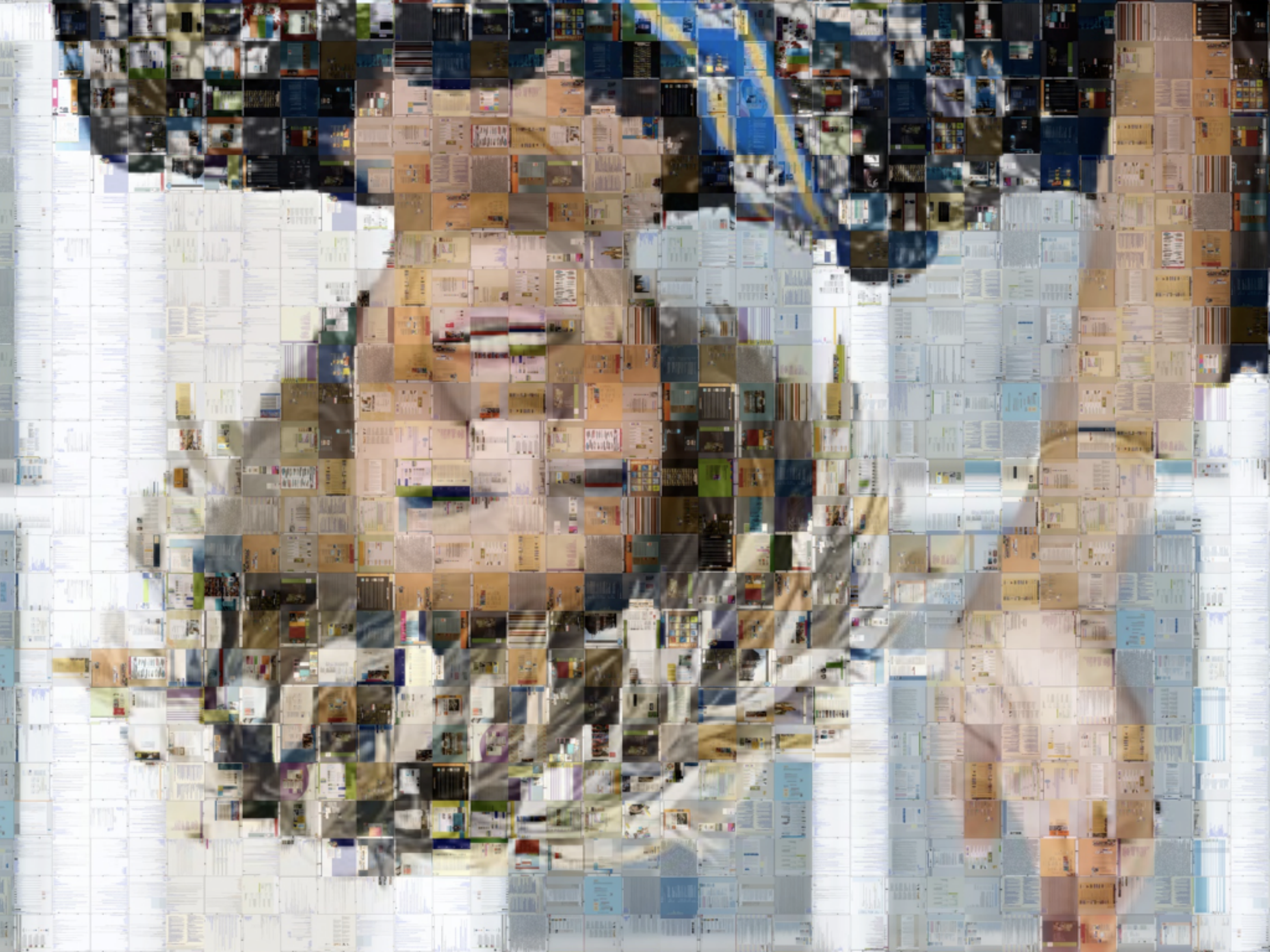
In het KB-webarchief zijn honderden gearchiveerde websites vanaf 2007 te vinden.

Het KB-webarchief bevat een selectie Nederlandse websites die vanaf 2007 zijn verzameld door de KB. De sites worden één tot twee keer per jaar binnengehaald (geharvest) en duurzaam opgeslagen. Als u in het bezit bent van een KB-jaarpas kunt u het KB-webarchief in de leeszaal van de KB bekijken met behulp van de zogenaamde *Wayback Machine*.

Volop in ontwikkeling

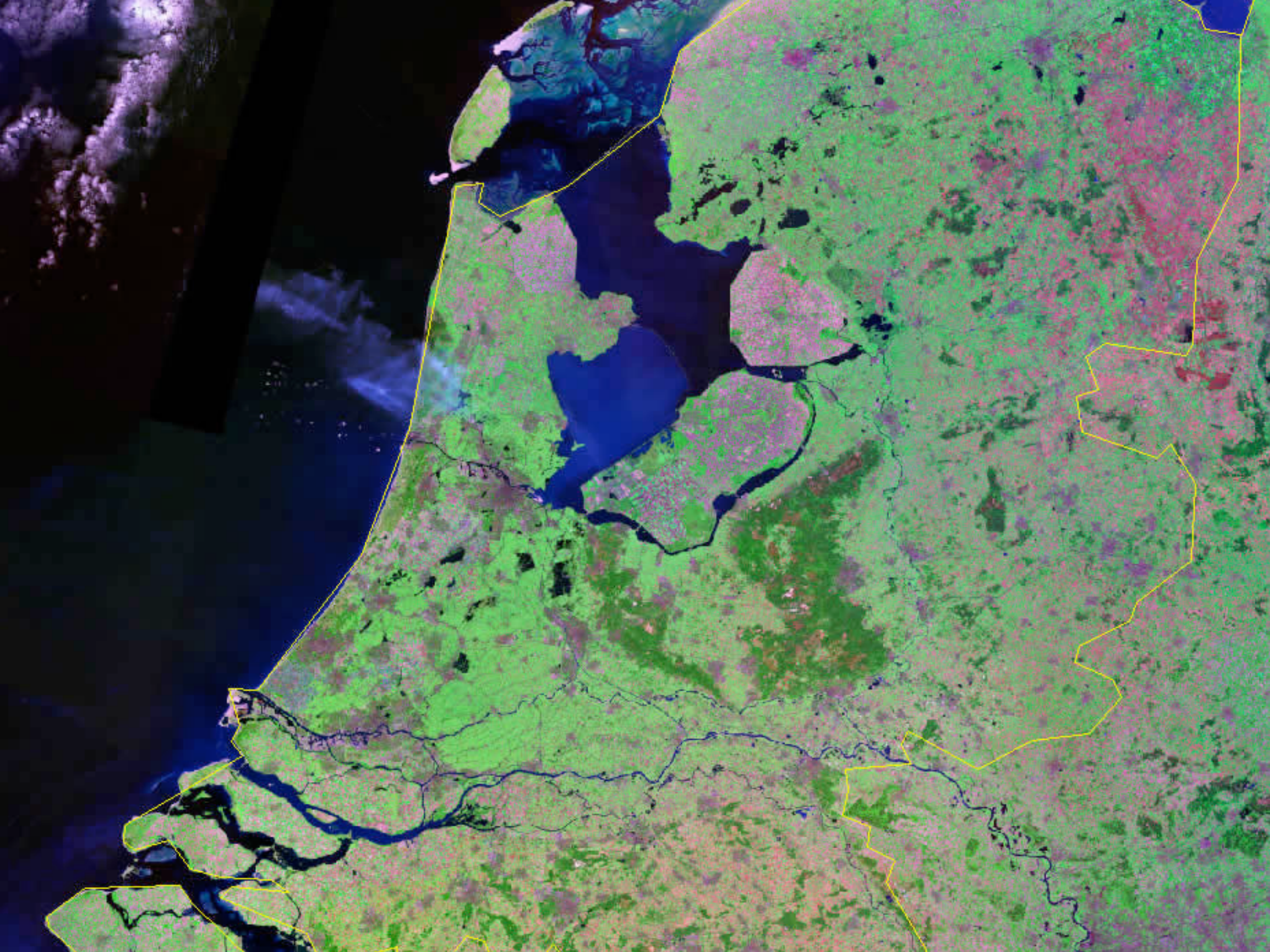
Het KB-webarchief is nog volop in ontwikkeling en nog niet alle gearchiveerde sites zijn beschikbaar. De KB werkt in internationaal verband aan het verbeteren van de software voor het verwerven en tonen van de websites. In een latere fase zal het KB-webarchief ook online beschikbaar komen.

Ga naar het webarchief *(alleen toegankelijk in de leeszaal van de KB)*





Original image: A N P



INTERNET ARCHIVE
WayBackMachine

KB Webarchief van Nederland
bèta

Kies naam site | of URL http://www.nu.nl | Alle | Zoeken | Uitgebreid Zoeken

Het KB-webarchief bevat een selectie Nederlandse websites, verzameld vanaf 2007.

Hoe werkt het?

Verken het webarchief door in het rechter zoekveld een URL in te geven en "Zoeken" te kiezen. Ook kunt u in het linker zoekveld de naam van een site invoeren. Klik op "Zoeken" voor een overzicht van alle gevonden versies van de gekozen site of URL. Klik de gewenste versie aan om deze te bekijken.
Zoeken is in deze fase alleen beschikbaar in de leeszaal van de KB.

Nog in ontwikkeling

Het KB-webarchief is nog volop in ontwikkeling. Door technische beperkingen kan het zijn dat (delen van) van gearchiveerde websites niet volledig getoond kunnen worden. Ook kan het voorkomen dat de archiefversie informatie toont vanuit het actuele internet. U kunt dit controleren via de adresbalk. Informatie uit het archief heeft altijd een URL die begint met domeinnaam "webaccess.kb.nl".
De KB werkt aan het verbeteren en het uitbreiden van het archief. Het vroegste materiaal (2007- half 2009) zal later in 2011 opgenomen worden. Daarnaast wordt de selectie voortdurend aangevuld.

Graag vernemen wij uw mening over inhoud en functionaliteit van het KB-webarchief. Maak hiervoor gebruik van het contactformulier.

| Home | Contactformulier |

Support: Wayback Machine

INTERNET ARCHIVE
WayBackMachine

KB Webarchief van Nederland
bèta

Kies naam site | of URL http:// | Alle | Zoeken | Uitgebreid Zoeken

Gezocht naar http://www.nu.nl

Kies tijdsvenster: none | 982 Resultaten

Resultaten voor periode 01-01-1996 tot 31-12-2014

Jan 1996 - Dec 1997	Jan 1998 - Dec 1999	Jan 2000 - Dec 2001	Jan 2002 - Dec 2003	Jan 2004 - Dec 2005	Jan 2006 - Dec 2007	Jan 2008 - Dec 2009	Jan 2010 - Dec 2011	Jan 2012 - Dec 2013	Jan 2014 - Dec 2015
0 pagina's	0 pagina's	0 pagina's	0 pagina's	0 pagina's	4 pagina's	11 pagina's	157 pagina's	778 pagina's	32 pagina's
					12-01-2007 *	25-02-2009 *	09-01-2010 *	02-01-2012 *	02-01-2014 *
					24-01-2007 *	09-09-2009 *	11-01-2010 *	02-01-2012 *	03-01-2014 *
					04-02-2007 *	10-09-2009 *	12-01-2010 *	03-01-2012 *	03-01-2014 *
					22-03-2007 *	01-10-2009 *	25-02-2010 *	03-01-2012 *	03-01-2014
						01-10-2009 *	07-04-2010 *	03-01-2012 *	04-01-2014 *
						06-10-2009 *	09-04-2010 *	04-01-2012 *	05-01-2014 *
						06-10-2009 *	11-04-2010 *	04-01-2012 *	06-01-2014 *
						26-10-2009 *	22-05-2010 *	06-01-2012 *	07-01-2014 *
						01-12-2009 *	25-05-2010 *	09-01-2012 *	08-01-2014 *

Support: Wayback Machine

INTERNET ARCHIVE
WayBackMachine BETA

http://www.nu.nl/ Toon DEC JAN 12 2007 MAR »

969 versies
12 Jan 07 - 31 Jan 14

2006 2009 »

Sluiten

Vr. 12 januari 2007. Het laatste nieuws lees je het eerst op NU.nl | Zoek nieuws zoek

Algemeen
Economie
Internet
Sport
Achterklap
Beurs
Wetenschap
Plugged
Auto

Weer
Verkeer

Muziek
TVFilm
Dvd
Film
Boek
Games

NU-blog
Yoshe
Column
Cartoon
Lezersfoto

Video



RandstadRail gaat testen in februari

Onduidelijk wanneer treinen weer personen gaan vervoeren

- **Nederlanders in vuurgevecht met Taliban**
- **Eerste zitting verbod Hells Angels**
- **Man rijdt kilometers mee op motorkap**
- **Nederlands meisje in Nigeria vermoord**
- **'Meerderheid jeugd negatief over moslims'**
- **Enorme toename Europeanen met drank op achter het stuur**
- **Voorarrest verdachte moord Verstappen verlengd**

Strengere regels voor krediet

AFM pakt 'onverantwoorde' gelduitleen aan

Weer

Weer van 17:00:
11°C

Verkeer

Er zijn **24** files met een totale lengte van 95 km.

Achterklap



Column

"Topsport is uitgevonden om te winnen - hoe dan ook"

Support: Wayback Machine

Support

- **Wayback Machine:** "Single URL" approach:
  - knowledge of URL required (no search functionality)
  - one URL at a time
  - hard to make comparisons

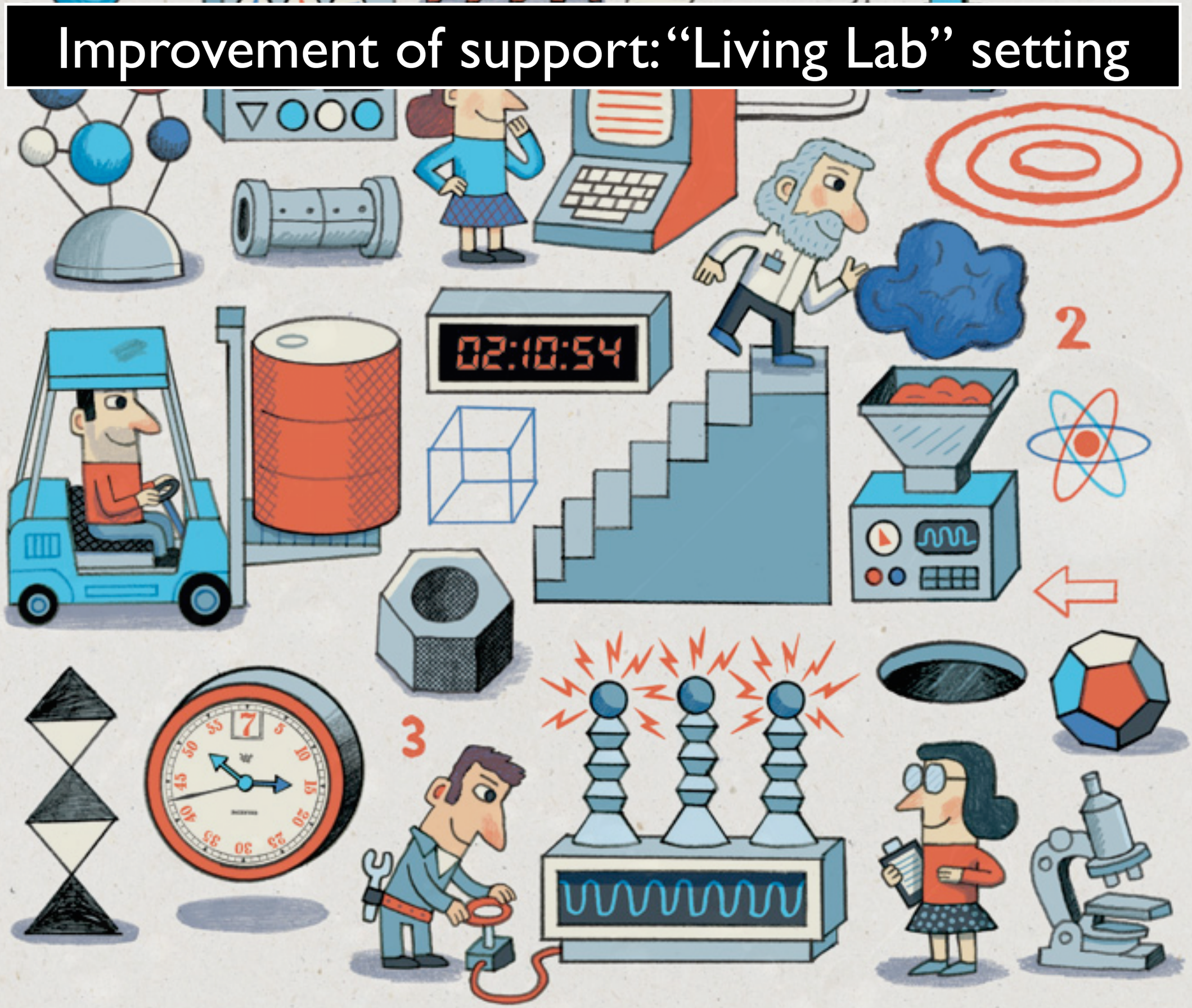
[Ben-David e.a. '14, Rogers, '13]

Improvement of support: "Living Lab" setting



allansanders.co.uk

Improvement of support: "Living Lab" setting



allansanders.co.uk

6. WebART Workshops & Tools

WebART workshops



- **09/12 DMI Summer School**
- 11/12 Web Archiving: Theorized Practices
- **01/13 DMI Winter School**
- 05/13 Exploring Israeli Politics online
- **09/13 DMI "Web Archiving day" - evaluation WebART tools**
- 11/13 New Media Research Masters - Web archive research scenarios

Transitions…

6.1 'Traditional' Web archive access

- Initial research explorations existing tools



6.2 Towards search-based access

- Living Lab, workshops and focus group



6.3 Beyond search-based access

6.1 "Traditional Web archive access"

INTERNET ARCHIVE WayBackMachine BETA

http://www.nu.nl/ Toon DEC JAN 12 2007 MAR »

969 versies
12 Jan 07 - 31 Jan 14

2006 2009 »

Sluiten

nu.nl

Vr. 12 januari 2007. Het laatste nieuws lees je het eerst op NU.nl | Zoek nieuws zoek

Algemeen
Economie
Internet
Sport
Achterklap
Beurs
Wetenschap
Plugged
Auto

Weer
Verkeer

Muziek
TVFilm
Dvd
Film
Boek
Games

NU-blog
Yoshe
Column
Cartoon
Lezersfoto

Video



RandstadRail gaat testen in februari

Onduidelijk wanneer treinen weer personen gaan vervoeren

- **Nederlanders in vuurgevecht met Taliban**
- **Eerste zitting verbod Hells Angels**
- **Man rijdt kilometers mee op motorkap**
- **Nederlands meisje in Nigeria vermoord**
- **'Meerderheid jeugd negatief over moslims'**
- **Enorme toename Europeanen met drank op achter het stuur**
- **Voorarrest verdachte moord Verstappen verlengd**

Strengere regels voor krediet

AFM pakt 'onverantwoorde' gelduitleen aan

Weer

Weer van 17:00:
11°C

Verkeer

Er zijn **24** files met een totale lengte van 95 km.

Achterklap



Column

"Topsport is uitgevonden om te winnen - hoe dan ook"

Support: Wayback Machine, DMI tools

6.1 "Traditional Web archive access"





Support: Wayback Machine, DMI tools

6.1 DMI Summer School (2012)

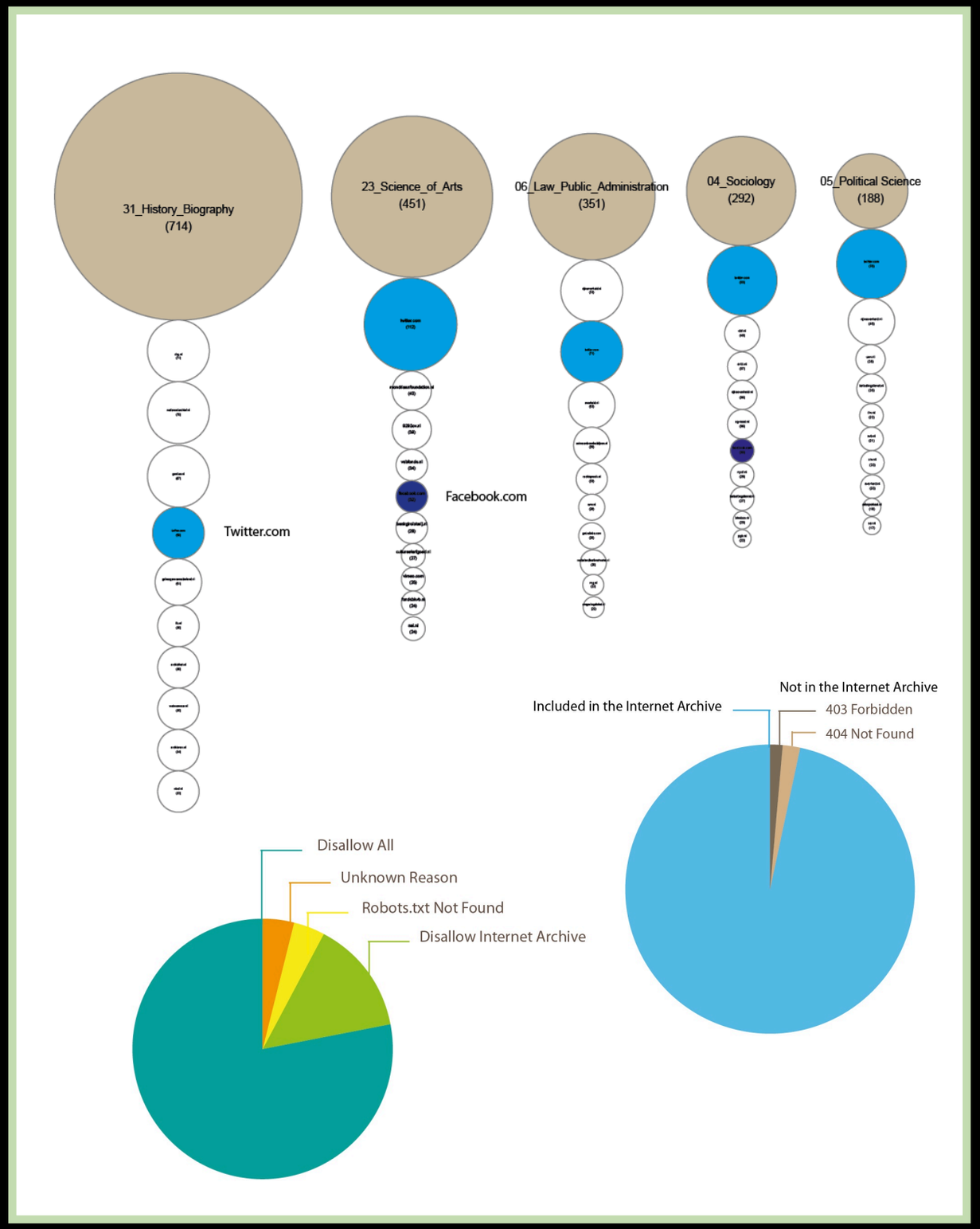






Flickr: Silvertje

Data: Selection lists KB





KB_collection_unesco_ca31_history_biography

Co-link Map Details:

Author:	Anat Ben-David
Email:	anatbd@gmail.com
Crawl start:	4 Jul 2012 - 20:11
Crawl end:	4 Jul 2012 - 23:55
Privilege starting points:	off
Co-link Analysis Mode:	page
Iterations:	2
Crawl Depth:	2
Node count:	50

Map generated from Issuecrawler.net by the Govcom.org Foundation, Amsterdam.

Legend:

(.nl) (.nu) (.info) (.net) (.org) (.com)

Select links layers:

links on

Statistics:

DMI Summer School (2012)

6.2 Towards search-based access

- **DMI Winter School** (January 2013)
- Israel Workshop (May 2013)
- **DMI 'Web Archiving Day'** (September 2013)
- New Media Research Masters proposals (November 2013)

6.2 Towards search-based access

- **DMI Winter School** (January 2013)
- Israel Workshop (May 2013)
- **DMI 'Web Archiving Day'** (September 2013)
- New Media Research Masters proposals (November 2013)

6.2.1 DMI Winter School (2013)



WebART & Anne Helmond, David Moats, Natalia Sánchez, Catherine Somzé

6.2.1 DMI Winter School (2013)

- Digital Methods Winter School (Jan. '13)
  - *Co-design* workshop ("Living Lab")
    - New Media researchers & system designers

WebART & Anne Helmond, David Moats, Natalia Sánchez, Catherine Somzé

6.2.1 DMI Winter School (2013)

- Digital Methods Winter School (Jan. '13)
  - *Co-design* workshop ("Living Lab")
    - New Media researchers & system designers



nu.nl subset webarchief
(432 crawls, 14 GB)

WebARTist
Web Archive Temporal Information Search Tools

Full-text Search Engine

WebARTist
Web Archive Temporal Information Search Tools

» Search tools

koninklijke bibliotheek Search

386 results for koninklijke bibliotheek, 386 filtered/shown results. Image search

1 Webarchief **Koninklijke Bibliotheek** telt twee miljoen krantenpagina's | nu.nl/media | Het laatste nieuws het eerst op nu.nl *archived page*



Via de De kranten bieden onder andere inzicht in de watersnoodramp in 1953 met foto's, kaartjes van het getroffen gebied en het dagelijks oplopende dodental. Naast kranten uit Nederland, zoals De Telegraaf en het Nieuwsblad van het Noorden, zijn ook negentiende-eeuwse publicaties uit Suriname opgenomen. In het krantenarchief is ook ruimte voor vermaak. Zo nam de bibliotheek de parodie op het huis-aan-huisblad van Kees van Kooten en Wim de Bie, de Juinensche Courant, in de collectie op. In de loop van volgend jaar moet het webarchief bestaan uit negen miljoen krantenpagina's.

url: http://www.nu.nl/media/2622346/webarchief-koninklijke-bibliotheek-telt-twee-miljoen-krantenpaginas.html collection: KB **site:** nu.nl size: 64671 crawl: 22/09/2011 **page:** 22/09/2011 **outlinks:** 19

2 Grondwet legt het af tegen Donald Duck | nu.nl/boek | Het laatste nieuws het eerst op nu.nl *archived page*



Het vrolijke weekblad versloeg daarmee onder meer de grondwet uit 1848. Die eindigde op de derde plek. Op de tweede plaats staat het Plakkaat van Verlatinge uit 1581. In die onafhankelijkheidsverklaring zette een aantal Nederlandse provincies koning Filips II van Spanje af als heerser van het land. De verkiezing van het drukwerk met de grootste betekenis voor Nederland werd georganiseerd door de Koninklijke Bibliotheek en het National Archief in Den Haag, die de uitslag van de online verkiezing donderdag bekendmaakten. In totaal brachten ongeveer duizend mensen hun stem uit. De top drie en andere genomineerde drukwerken zijn tot 28 augustus te zien in de expositieruimte van het Nationaal Archief en de Koninklijke Bibliotheek, De Verdieping van Nederland. Het nieuwe boekenseizoen is dit weekend feestelijk van start gegaan op boekenbeurs Manuscripta in Amsterdam. NU.nl was erbij. 'Je hebt schrijvers en je hebt auteurs', zo betoogde De genomineerden voor de

url: http://www.nu.nl/boek/2592705/grondwet-legt-af-donald-duck.html collection: KB **site:** nu.nl size: 65656 crawl: 06/09/2011 **page:** 18/08/2011 **outlinks:** 20

3 'Dak eraf' tijdens Haagse Museumnacht | nu.nl/binnenland | Het laatste nieuws het eerst op nu.nl *archived page*



De drummers Cesar Zuiderwijk van de rockband Golden Earring en Koen Herfst van dj Armin Van Buuren gaven met een zogeheten drumbattle het startschot in de Koninklijke Schouwburg. De woordvoerder van het evenement spreekt van een groot succes. "We hadden geen betere opening kunnen hebben. Het dak ging eraf." De Haagse Museumnacht werd dit jaar voor de tweede keer gehouden. Na het succes van vorig jaar had de organisatie meer bezoekers verwacht. Door het mooie zomerweer verkozen veel mensen het terras boven het museum, vermoedt de zegsman. In totaal deden 38 culturele instellingen mee waaronder verscheidene musea, galerie Wie daarna nog even wil feesten, kan nog verder dansen op afterparty's in onder andere de Centrale Bibliotheek en het poppodium Paard van Troje. De volgende editie van de Museumnacht in Den Haag is op 1 september 2012. Het nieuwe boekenseizoen is dit weekend feestelijk van start gegaan op boekenbeurs Manuscripta in Amsterdam. NU.nl was erbij.

url: http://www.nu.nl/binnenland/2606409/dak-eraf-tijdens-haagse-museumnacht.html collection: KB **site:** nu.nl size: 73089 crawl: 06/09/2011 **page:** 04/09/2011 **outlinks:** 36

WebARTist

» Search tools

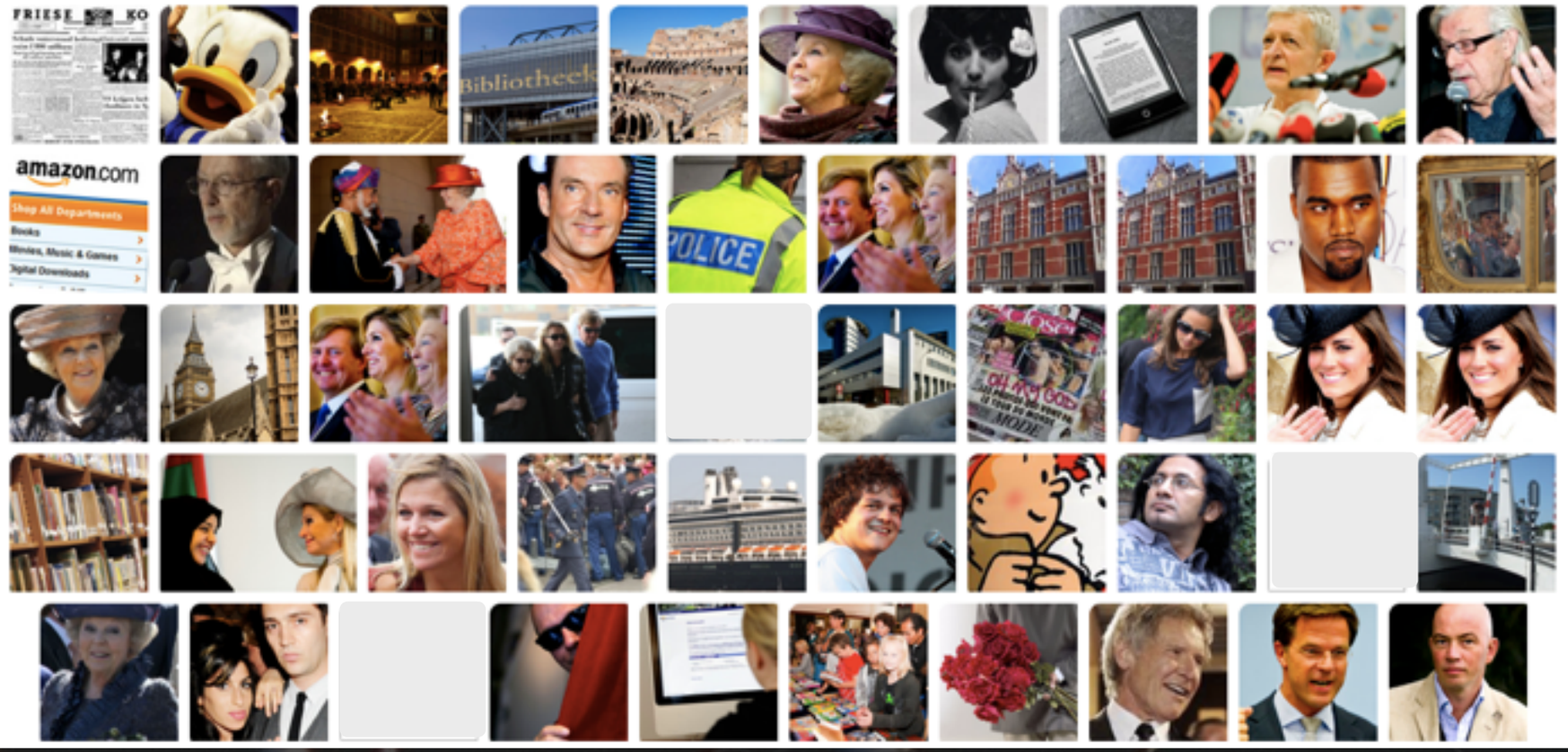
WebARTist

Web Archive Temporal Information Search Tools

» Search tools

koninklijke bibliotheek | DPH | Search

386 results for koninklijke bibliotheek, 386 filtered/shown results. Text search



WebARTist

» Search tools

WebARTist

» Search tools

WebARTist

Web Archive Temporal Information Search Tools

» Search tools

koninklijke familie date:2011 collection:kb qe:on Search

5499 results for koninklijke familie date:2011 collection:kb qe:on, 1001 filtered/shown results. Image search

lech ski koninginnedag foto prins mabel lawin veenenda

beatrix wachtkamer oranj oostenrijks rhenen friso leden maxima

koningin

Export query expansion

1 Projectiel naar Gouden Koets gegooid I nu.nl/binnenland I Het laatste nieuws het eerst op nu.nl *archived page*



De politie heeft de De schade aan de Gouden Koets is niet groot. De koningin kon haar rijtoer zonder vertraging voortzetten. Dat heeft de politie gemeld. De man is direct naar een politiebureau gebracht. De woordvoerder sprak van een vervelend incident, maar zei dat de sfeer "plezierig bleef'. Even voor het incident in de Heulstraat was er ook al een persoon aangehouden die te veel gedronken had, meldde de politie. Het is niet de eerste keer dat de koninklijke familie het doelwit is. Vorig jaar op Koninginnedag in Apeldoorn reed een man door de afzetting richting de bus waarmee leden van het koningshuis een rijtoer maakten door de stad. Bij die aanslag kwamen acht mensen, onder wie dader Karst T., om het leven. In 2002 gooide op de huwelijksdag in Amsterdam van prins Willem-Alexander en prinses M Het nieuwe boekenseizoen is dit weekend feestelijk van start gegaan op boekenbeurs Manuscripta in Amsterdam. NU.nl was erbij.

url: http://www.nu.nl/binnenland/2338324/projectiel-gouden-koets-gegooid.html collection: KB site: nu.nl size: 70552 crawl: 06/09/2011 page: 04/06/2011 outlinks: 39

2 Gooier waxinelichthouder moet naar Pieter Baan Centrum I nu.nl/binnenland I Het laatste nieuws het eerst op nu.nl *archived page*

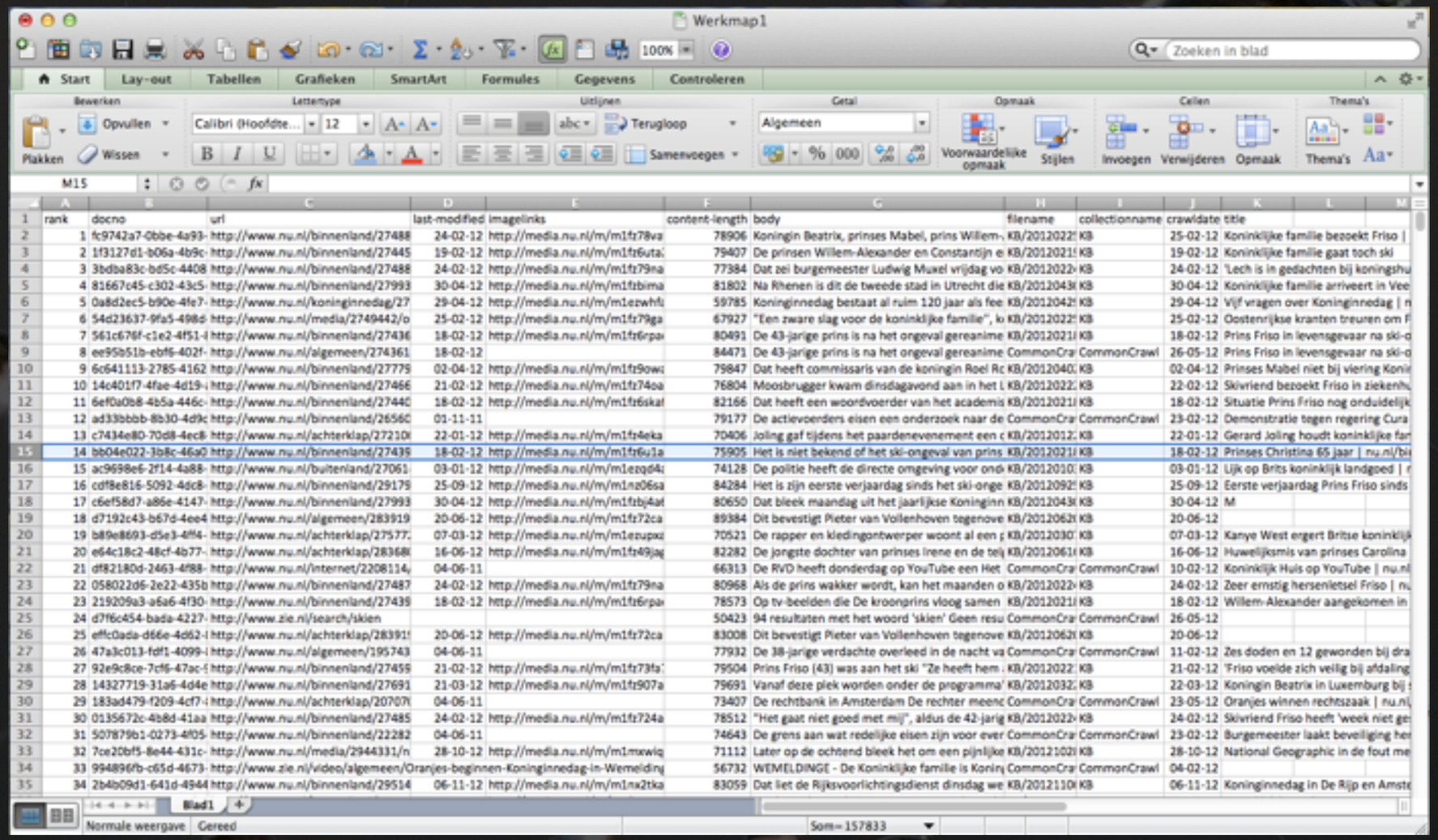


De rechtbank in Den Haag heeft vrijdag beslist dat Erwin L. geobserveerd moet worden in het Pieter Baan Centrum (PBC). Pas daarna kan de rechtbank een afgewogen oordeel geven over de eventuele schuld van de 30-jarige man en over zijn toerekeningsvatbaarheid. De 30-jarige L. gooide de kaarsenhouder op 21 september naar het rijtuig waarmee koningin Beatrix, prins Willem-Alexander en prinses M L. is wel onderzocht door een psycholoog en een psychiater, maar weigerde tot nu toe mee te werken aan meer onderzoek naar zijn geestestoestand. De rechtbank wil hem nu toch laten observeren in het PBC, in de hoop daarover meer te weten te komen. Een deskundige van het Nederlands Instituut voor Forensische Psychiatrie en Psychologie (NIFP) heeft gezegd dat L. lijdt aan een waanstoornis. Volgens de psychiater zou hij volledig ontoerekeningsvatbaar moeten worden verklaard. Het Openbaar Ministerie beschuldigt de man ervan dat hij de koningin, haar troonopvolger en diens echtgenote heeft beledigd, de Gouden Koets heeft vernield en de intentie heeft gehad de koninklijke familie zwaar te mishandelen. De officier van justitie sloot zich dinsdag aan bij de conclusie van de psychiater en eiste daarom ontslag van rechtsvervolging en opname in een psychiatrisch ziekenhuis. Het is onbekend wanneer de rechtbank een definitieve uitspraak kan doen. Het nieuwe boekenseizoen is dit weekend feestelijk van start gegaan op boekenbeurs Manuscripta in Amsterdam. NU.nl was erbij.

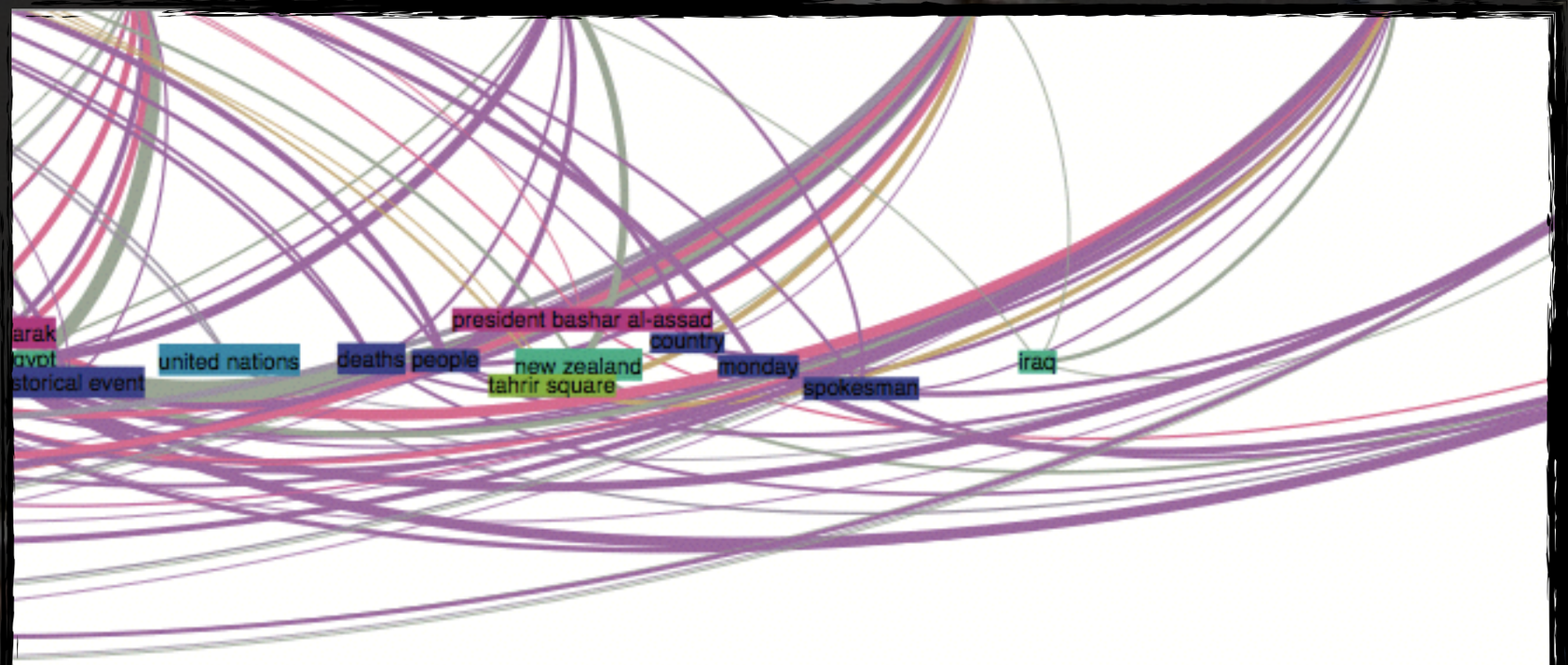
url: http://www.nu.nl/binnenland/2471481/gooier-waxinelichthouder-moet-pieter-baan-centrum.html collection: KB site: nu.nl size: 72110 crawl: 06/09/2011 page: 02/06/2011 outlinks: 36

WebARTist

Web Archive Temporal Information Search Tools



rank	docno	url	last-modified	imagelinks	content-length	body	filename	collectionname	crawldate	title
1	fc9742a7-0bbe-4a93-	http://www.nu.nl/binnenland/27488	24-02-12	http://media.nu.nl/m/m1fz78va	78906	Koningin Beatrix, prinses Mabel, prins Willem-	KB/20120222	KB	25-02-12	Koninklijke familie bezoekt Friso |
2	1f3127d1-b06a-4b9c-	http://www.nu.nl/binnenland/27445	19-02-12	http://media.nu.nl/m/m1fz6uta	79407	De prinsen Willem-Alexander en Constantijn e	KB/20120215	KB	19-02-12	Koninklijke familie gaat toch ski
3	3bdba83c-bd5c-4408	http://www.nu.nl/binnenland/27488	24-02-12	http://media.nu.nl/m/m1fz79na	77384	Dat zei burgemeester Ludwig Muxel vrijdag vo	KB/20120222	KB	24-02-12	'Lech is in gedachten bij koningshu
4	81667c45-c302-43c5-	http://www.nu.nl/binnenland/27993	30-04-12	http://media.nu.nl/m/m1fzbima	81802	Na Rhenen is dit de tweede stad in Utrecht die	KB/20120430	KB	30-04-12	Koninklijke familie arriveert in Vee
5	0a8d2ec5-b90e-4fe7-	http://www.nu.nl/koninginnedag/27	29-04-12	http://media.nu.nl/m/m1ezwhfu	59785	Koninginnedag bestaat al ruim 120 jaar als fee	KB/20120429	KB	29-04-12	Vijf vragen over Koninginnedag | n
6	54d23637-9fa5-498d-	http://www.nu.nl/media/2749442/o	25-02-12	http://media.nu.nl/m/m1fz79ga	67927	"Een zware slag voor de koninklijke familie", k	KB/20120225	KB	25-02-12	Oostenrijkse kranten treuren om F
7	561c676f-c1e2-4f51-(	http://www.nu.nl/binnenland/2743(	18-02-12	http://media.nu.nl/m/m1fz6rpa	80491	De 43-jarige prins is na het ongeval gereanime	KB/20120218	KB	18-02-12	Prins Friso in levensgevaar na ski-o
8	ee95b51b-ebf6-402f-	http://www.nu.nl/algemeen/274361	18-02-12		84471	De 43-jarige prins is na het ongeval gereanime	CommonCra	CommonCrawl	26-05-12	Prins Friso in levensgevaar na ski-o
9	6c641113-2785-4162	http://www.nu.nl/binnenland/2777(	02-04-12	http://media.nu.nl/m/m1fz9owz	79847	Dat heeft commissaris van de koningin Roel Rc	KB/20120402	KB	02-04-12	Prinses Mabel niet bij viering Konir
10	14c401f7-4fae-4d19-(	http://www.nu.nl/binnenland/27466	21-02-12	http://media.nu.nl/m/m1fz74oa	76804	Moosbrugger kwam dinsdagavond aan in het L	KB/20120222	KB	22-02-12	Skivriend bezoekt Friso in ziekenhu
11	6ef0a0b8-4b5a-446c-	http://www.nu.nl/binnenland/2744(	18-02-12	http://media.nu.nl/m/m1fz6skat	82166	Dat heeft een woordvoerder van het academis	KB/20120218	KB	18-02-12	Situatie Prins Friso nog onduidelijk
12	ad33bbbb-8b30-4d9c	http://www.nu.nl/binnenland/2656(	01-11-11		79177	De actievoerders eisen een onderzoek naar de	CommonCra	CommonCrawl	23-02-12	Demonstratie tegen regering Cura
13	c7434e80-70d8-4ec8-	http://www.nu.nl/achterklap/2721(	22-01-12	http://media.nu.nl/m/m1fz4eka	70406	Joling gaf tijdens het paardenevenement een c	KB/20120122	KB	22-01-12	Gerard Joling houdt koninklijke fan
14	bb04e022-3b8c-46a0	http://www.nu.nl/binnenland/27439	18-02-12	http://media.nu.nl/m/m1fz6u1a	75905	Het is niet bekend of het ski-ongeval van prins	KB/20120218	KB	18-02-12	Prinses Christina 65 jaar | nu.nl/bi
15	ac9698e6-2f14-4a88-	http://www.nu.nl/buitenland/27061	03-01-12	http://media.nu.nl/m/m1ezqd4z	74128	De politie heeft de directe omgeving voor ond	KB/20120103	KB	03-01-12	Lijk op Brits koninklijk landgoed | r
16	cdf8e816-5092-4dc8-	http://www.nu.nl/binnenland/29179	25-09-12	http://media.nu.nl/m/m1nz06sa	84284	Het is zijn eerste verjaardag sinds het ski-onge	KB/20120925	KB	25-09-12	Eerste verjaardag Prins Friso sinds
17	c6ef58d7-a86e-4147-	http://www.nu.nl/binnenland/27993	30-04-12	http://media.nu.nl/m/m1fzbj4at	80650	Dat bleek maandag uit het jaarlijkse Koninginn	KB/20120430	KB	30-04-12	M
18	d7192c43-b67d-4ee4	http://www.nu.nl/algemeen/283919	20-06-12	http://media.nu.nl/m/m1fz72ca	89384	Dit bevestigt Pieter van Vollenhoven tegenove	KB/20120620	KB	20-06-12	
19	b89e8693-d5e3-4ff4-	http://www.nu.nl/achterklap/27577(	07-03-12	http://media.nu.nl/m/m1ezupxz	70521	De rapper en kledingontwerper woont al een p	KB/20120307	KB	07-03-12	Kanye West ergert Britse koninklijk
20	e64c18c2-48cf-4b77-	http://www.nu.nl/achterklap/28368(	16-06-12	http://media.nu.nl/m/m1fz49jag	82282	De jongste dochter van prinses Irene en de tei	KB/20120616	KB	16-06-12	Huwelijksmis van prinses Carolina
21	df82180d-2463-4f88-	http://www.nu.nl/internet/2208114/	04-06-11		66313	De RVD heeft donderdag op YouTube een Het	CommonCra	CommonCrawl	10-02-12	Koninklijk Huis op YouTube | nu.nl
22	058022d6-2e22-435b	http://www.nu.nl/binnenland/27487	24-02-12	http://media.nu.nl/m/m1fz79na	80968	Als de prins wakker wordt, kan het maanden o	KB/20120222	KB	24-02-12	Zeer ernstig hersenletsel Friso | nu.
23	219209a3-a6a6-4f30-	http://www.nu.nl/binnenland/27439	18-02-12	http://media.nu.nl/m/m1fz6rpa	78573	Op tv-beelden die De kroonprins vloog samen	KB/20120218	KB	18-02-12	Willem-Alexander aangekomen in
24	d7f6c454-bada-4227-	http://www.zie.nl/search/skien			50423	94 resultaten met het woord 'skien' Geen resu	CommonCra	CommonCrawl	26-05-12	
25	effc0ada-d66e-4d62-(	http://www.nu.nl/achterklap/28391(	20-06-12	http://media.nu.nl/m/m1fz72ca	83008	Dit bevestigt Pieter van Vollenhoven tegenove	KB/20120620	KB	20-06-12	
26	47a3c013-fdf1-4099-(	http://www.nu.nl/algemeen/195743	04-06-11		77932	De 38-jarige verdachte overleed in de nacht va	CommonCra	CommonCrawl	11-02-12	Zes doden en 12 gewonden bij dra
27	92e9c8ce-7cf6-47ac-(	http://www.nu.nl/binnenland/27459	21-02-12	http://media.nu.nl/m/m1fz73fa	79504	Prins Friso (43) was aan het ski "Ze heeft hem	KB/20120222	KB	21-02-12	'Friso voelde zich veilig bij afdaling
28	14327719-31a6-4d4e	http://www.nu.nl/binnenland/27691	21-03-12	http://media.nu.nl/m/m1fz907a	79691	Vanaf deze plek worden onder de programma'	KB/20120323	KB	22-03-12	Koningin Beatrix in Luxemburg bij
29	183ad479-f209-4cf7-(	http://www.nu.nl/achterklap/20707(	04-06-11		73407	De rechtbank in Amsterdam De rechter meenc	CommonCra	CommonCrawl	23-05-12	Oranjes winnen rechtszaak | nu.nl
30	0135672c-4b8d-41aa	http://www.nu.nl/binnenland/27485	24-02-12	http://media.nu.nl/m/m1fz724a	78512	"Het gaat niet goed met mij", aldus de 42-jarig	KB/20120222	KB	24-02-12	Skivriend Friso heeft 'week niet ge
31	507879b1-0273-4f05-	http://www.nu.nl/binnenland/22282	04-06-11		74643	De grens aan wat redelijke eisen zijn voor ever	CommonCra	CommonCrawl	23-02-12	Burgemeester laakt beveiliging he
32	7ce20bf5-8e44-431c-	http://www.nu.nl/media/2944331/n	28-10-12	http://media.nu.nl/m/m1mxwiq	71112	Later op de ochtend bleek het om een pijnlijke	KB/20121028	KB	28-10-12	National Geographic in de fout me
33	994896fb-c65d-4673-	http://www.zie.nl/video/algemeen/Oranjes-beginnen-Koninginnedag-in-Wemeldinge			56732	WEMELDINGE - De Koninklijke familie is Konin	CommonCra	CommonCrawl	04-02-12	
34	2b4b09d1-641d-4944	http://www.nu.nl/binnenland/29514	06-11-12	http://media.nu.nl/m/m1nx2tka	83059	Dat liet de Rijksvoorlichtingsdienst dinsdag we	KB/20121106	KB	06-11-12	Koninginnedag in De Rijp en Amste



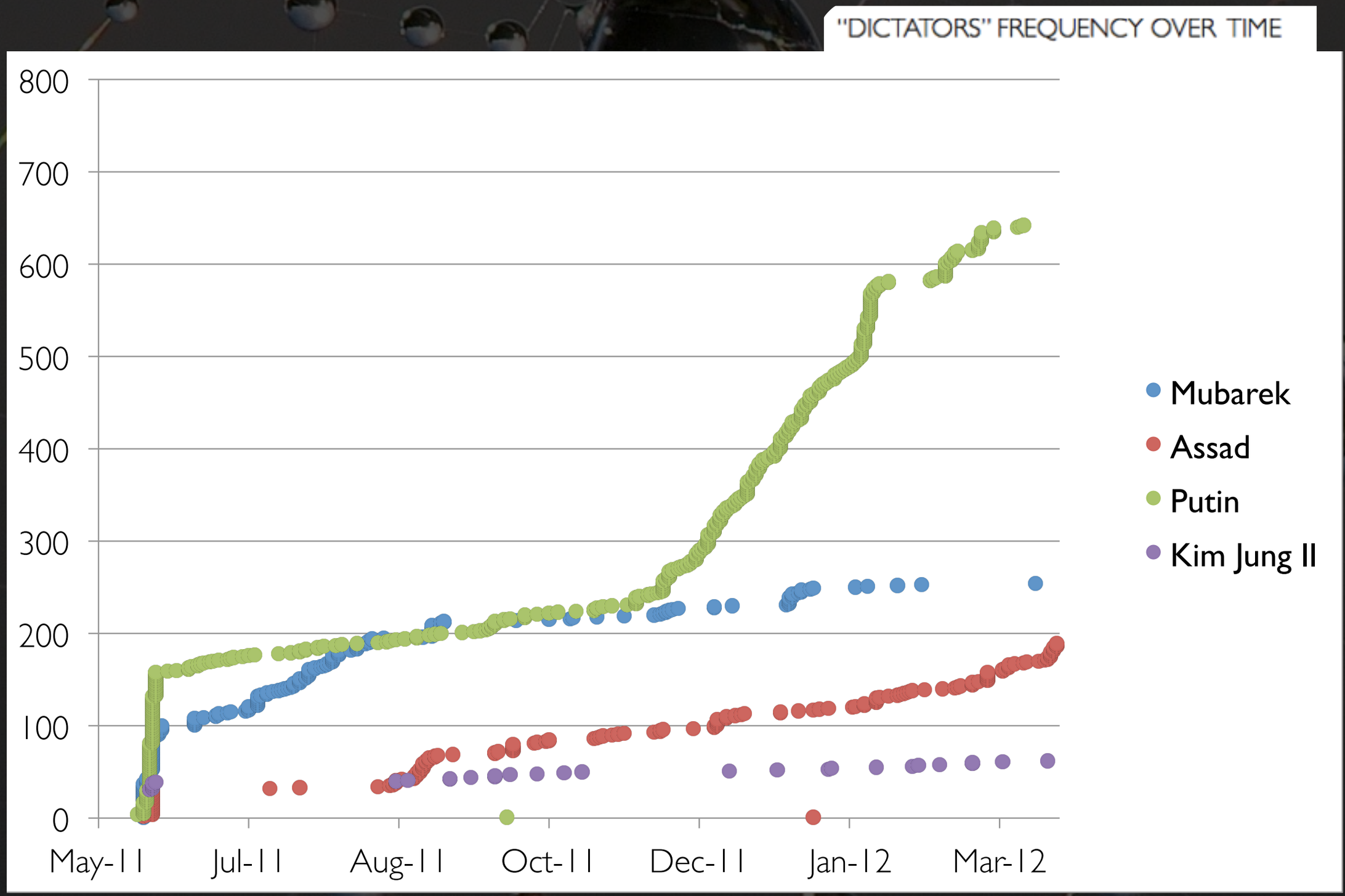
A YEAR IN THE NEWS:

Sprint-Exploring the Nu.nl Archive 2011-2012

Anat Ben David
Anne Helmond
Hugo Huurdeman
David Moats
Natalia Sánchez
Thaer Sammar
Catherine Somzé

Data Source: KB, CommonCrawl

DMI Winter School 2013

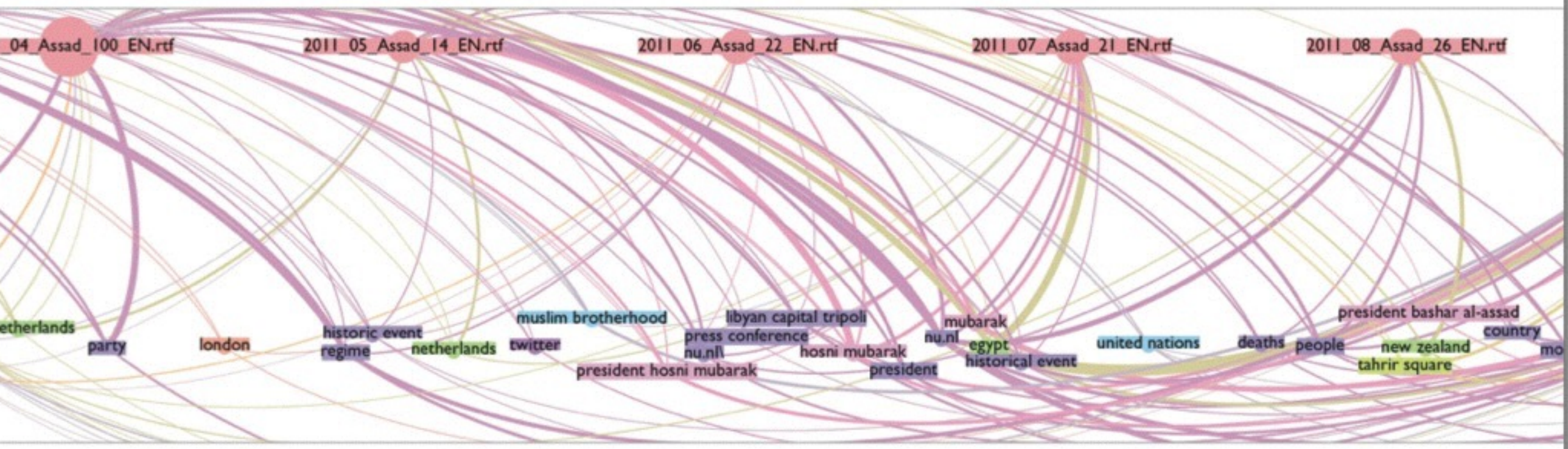


Content analyse: Word frequency analysis

CO-WORD ANALYSIS 2011



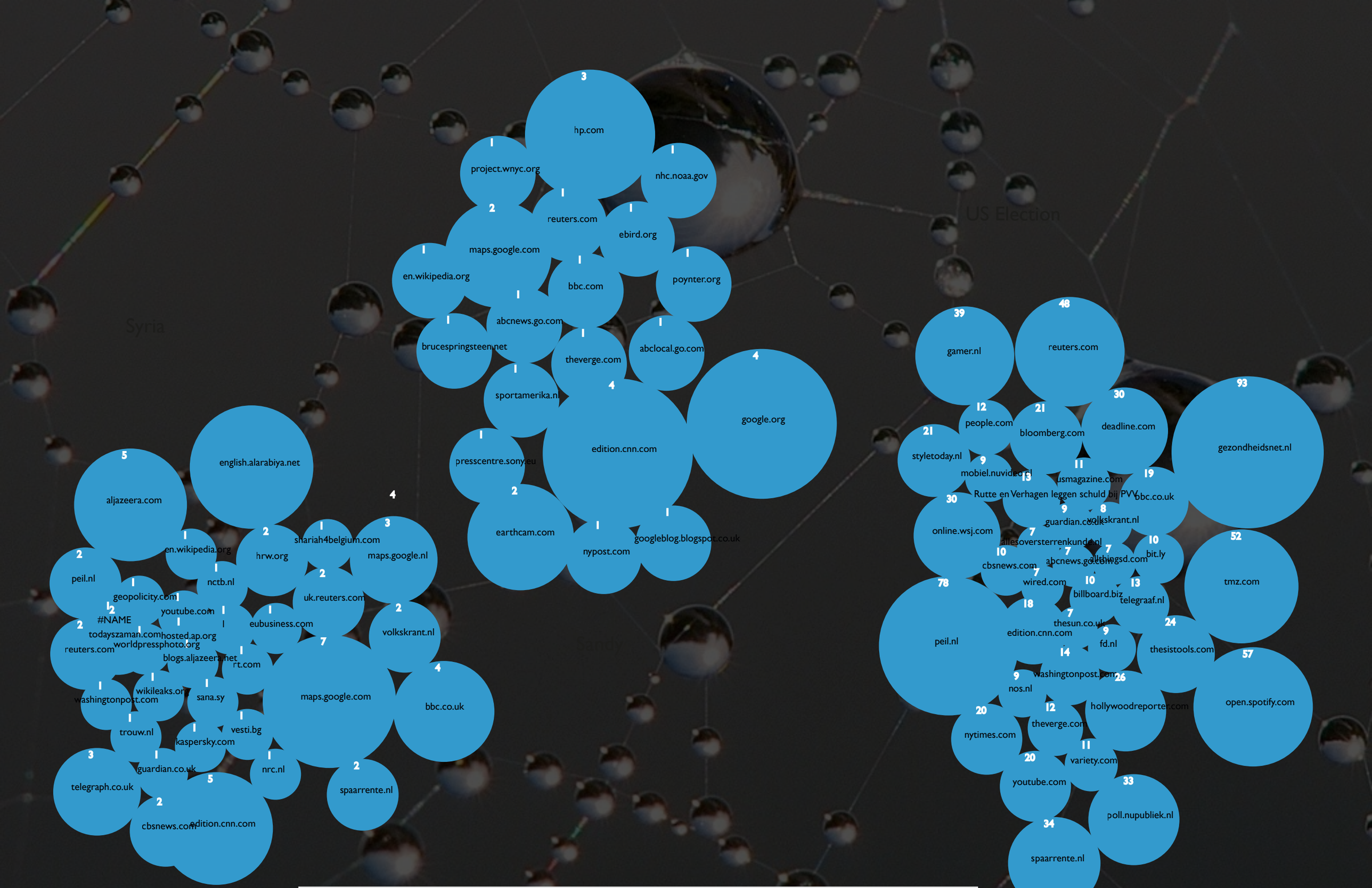
Co-word anlysis of Assad articles: J 2011



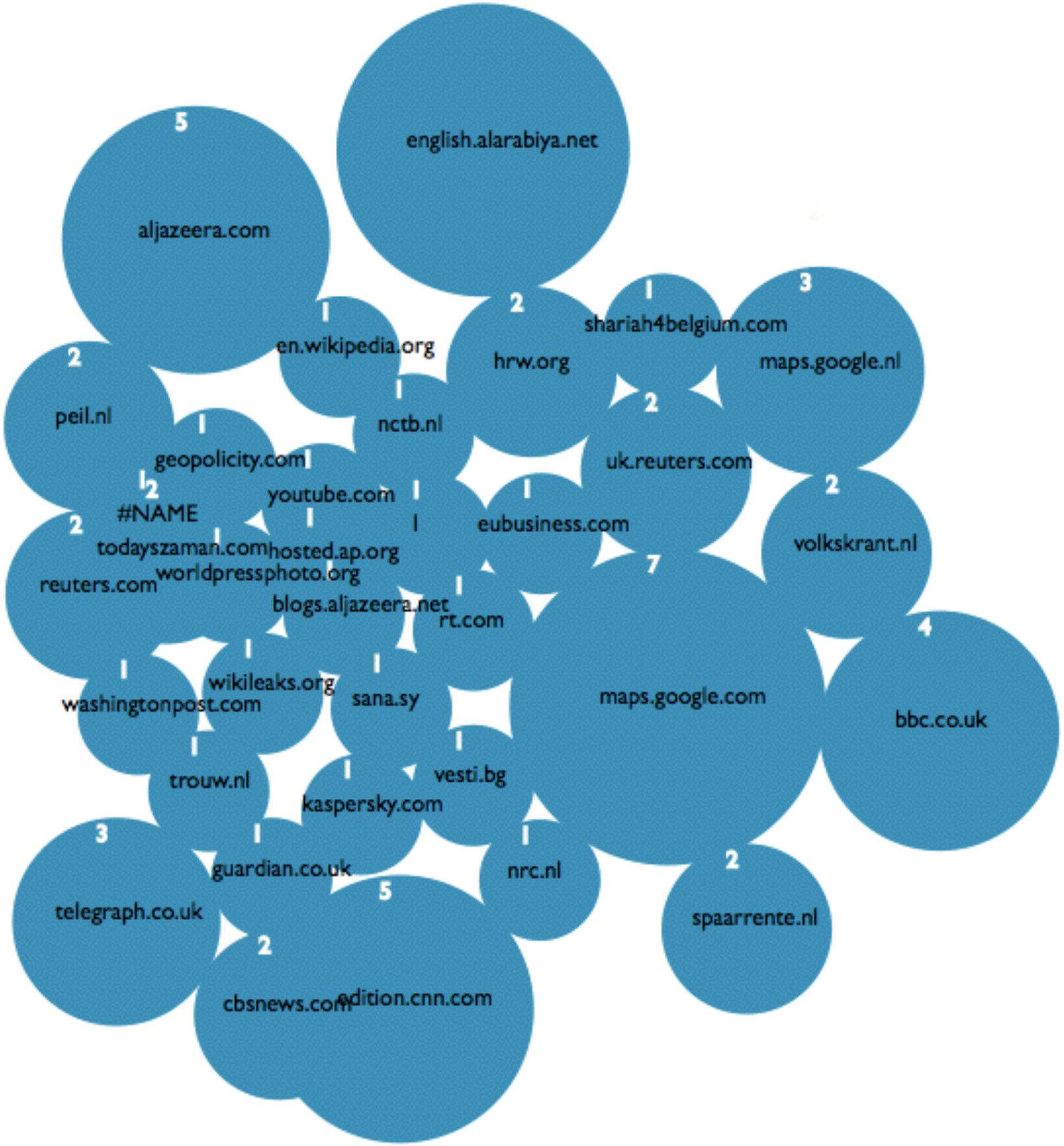
Co-Word Analyse

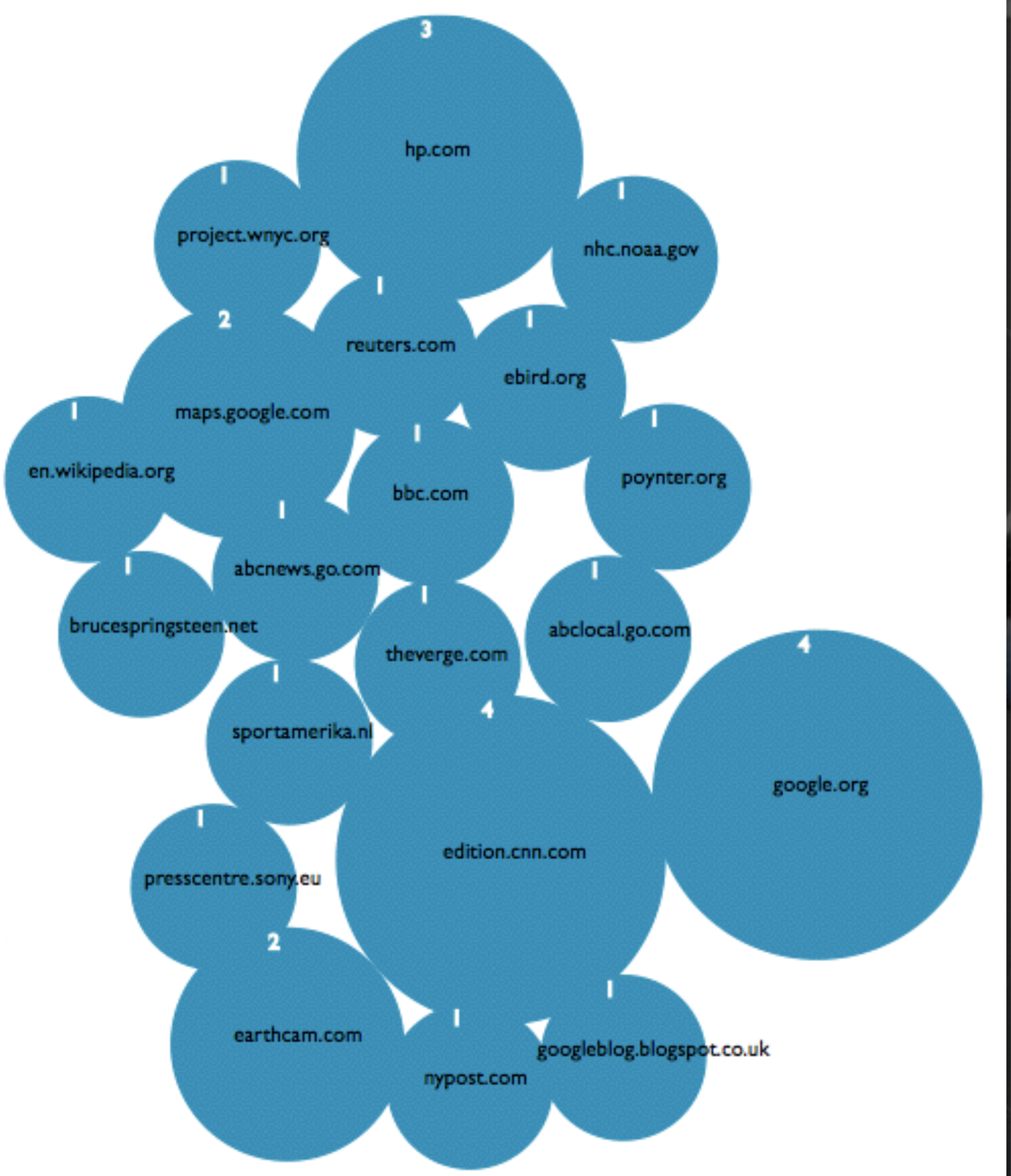
Geomapping: locatie bron nieuws

Linkstructuur: Outlink Analysis



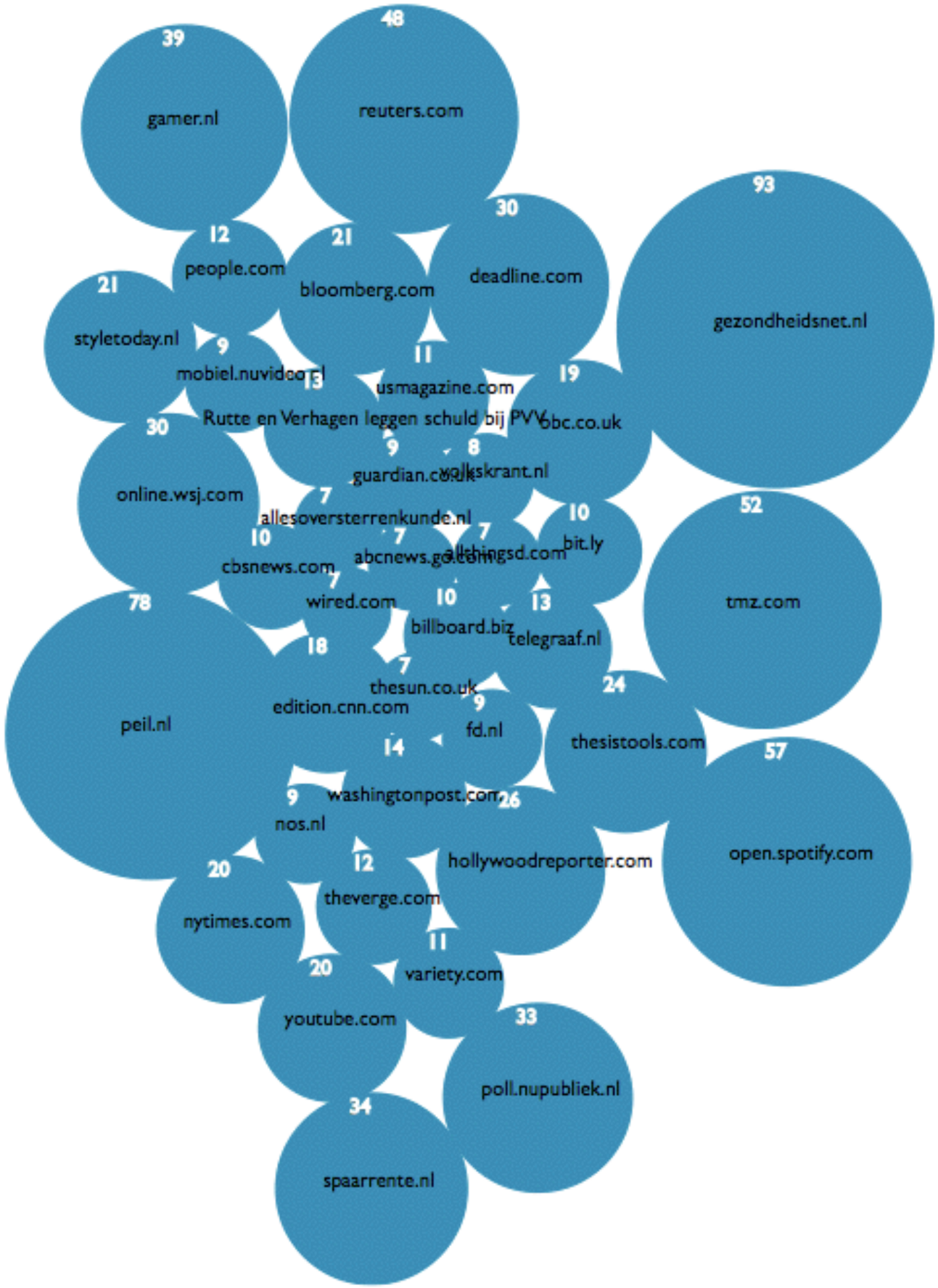
SYRIA


5
aljazeera.com
english.alarabiya.net
en.wikipedia.org
2
hrw.org
1
shariah4belgium.com
3
maps.google.nl
2
peil.nl
nctb.nl
2
uk.reuters.com
1
geopolicity.com
youtube.com
2
#NAME
eubusiness.com
2
volkskrant.nl
2
todayszaman.com
hosted.ap.org
reuters.com
worldpressphoto.org
7
blogs.aljazeera.net
rt.com
4
wikileaks.org
sana.sy
maps.google.com
bbc.co.uk
washingtonpost.com
trouw.nl
vesti.bg
kaspersky.com
3
guardian.co.uk
nrc.nl
2
5
telegraph.co.uk
spaarrente.nl
2
cbsnews.com
edition.cnn.com



HURRICANE SANDY

US ELECTION


39
gamer.nl
48
reuters.com
12
people.com
21
bloomberg.com
30
deadline.com
93
gezondheidsnet.nl
21
styletoday.nl
9
mobiel.nuvid.com
11
usmagazine.com
19
bbc.co.uk
Rutte en Verhagen leggen schuld bij PVV
30
online.wsj.com
9
8
volkskrant.nl
7
allesoversterrenkunde.nl
10
bit.ly
10
cbsnews.com
7
abcnews.go.com
7
allthingsd.com
52
tmz.com
78
7
wired.com
10
billboard.biz
13
telegraaf.nl
18
7
thesun.co.uk
9
edition.cnn.com
24
peil.nl
fd.nl
14
thesistools.com
57
9
washingtonpost.com
26
nos.nl
20
12
hollywoodreporter.com
open.spotify.com
theverge.com
nytimes.com
11
20
variety.com
youtube.com
33
poll.nupubliek.nl
34
spaarrente.nl

Media: Temporal Image Analyse

lees meer: [Huurdeman et al, 2013]

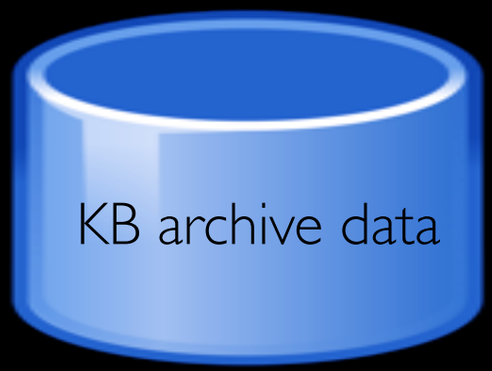
6.2.1 DMI Winter School (2013)

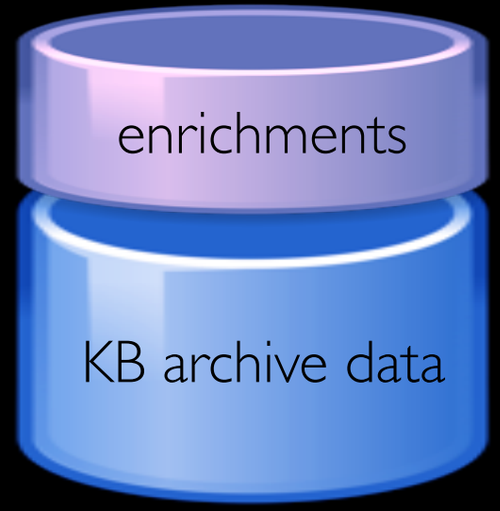
- New possibilities for exploration and use of archived material
  - from *single site browsing* naar *(cross-)collection search*
- *bottom-up* approach: adaptation functionality of system to needs users

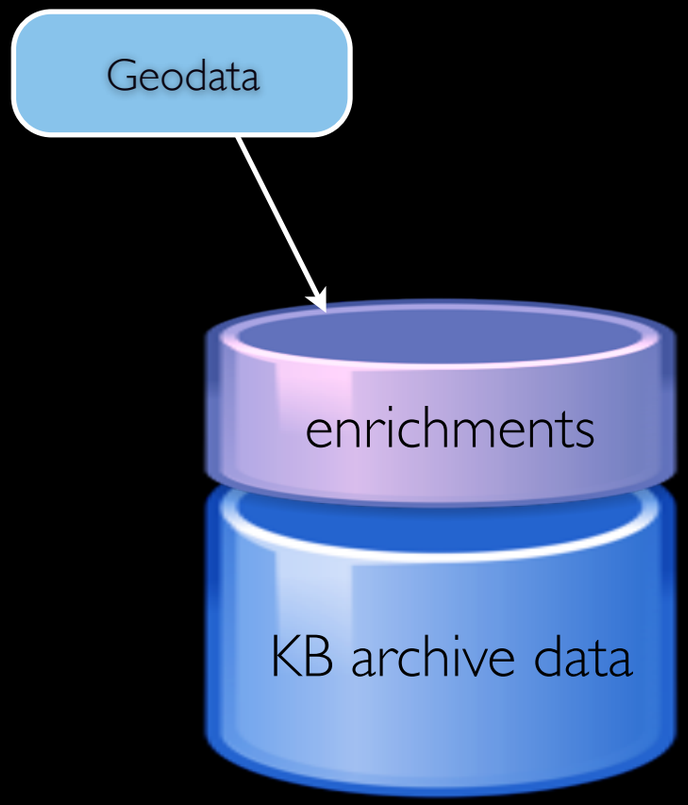
lees meer: [Huurdeman et al, 2013]

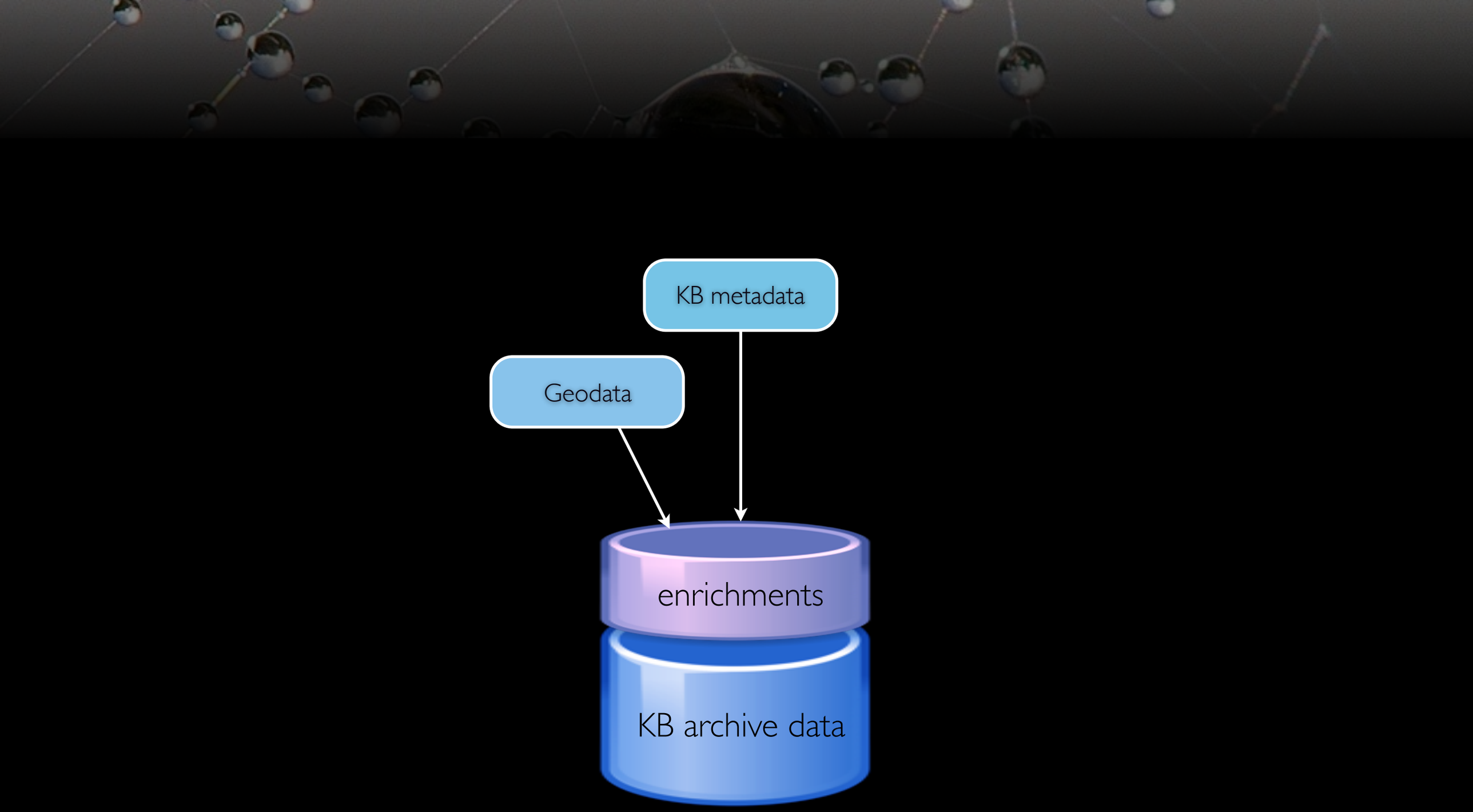
6.2.1 DMI Winter School (2013)

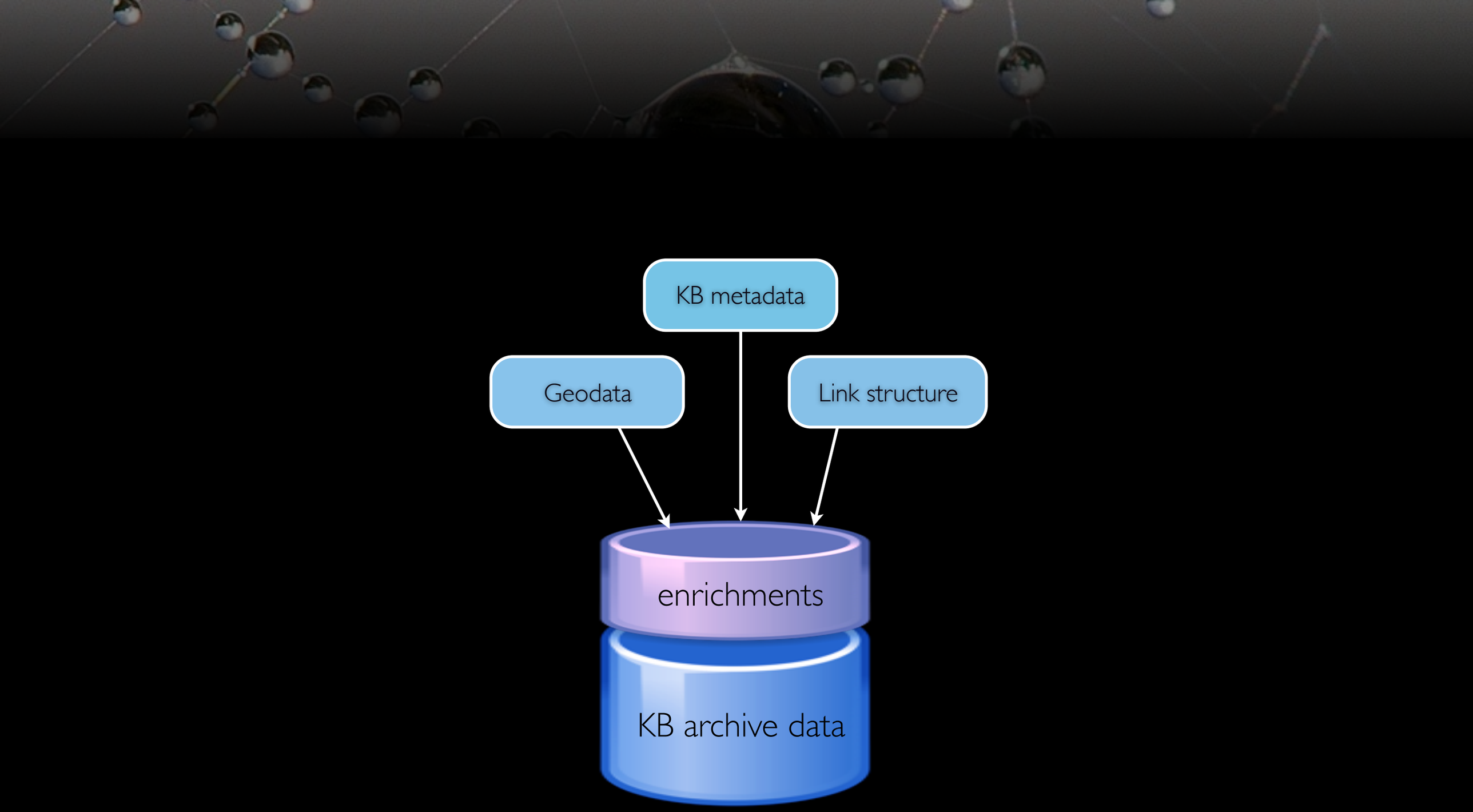
- New possibilities for exploration and use of archived material
  - from *single site browsing* naar *(cross-)collection search*
  - *bottom-up* approach: adaptation functionality of system to needs users
- **But:** strong demand for export functionality for analysis — in researchers *own tools*
  - How can we support more types of analysis?


KB archive data


enrichments
KB archive data

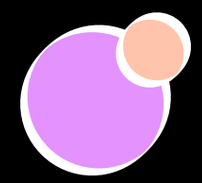

Geodata
enrichments
KB archive data


KB metadata
Geodata
enrichments
KB archive data


KB metadata
Geodata
Link structure
enrichments
KB archive data

nu.nl

host+1



nu.nl

Full ‘index’ KB Web archive

host+1

nu.nl

Full ‘index’ KB Web archive

43.533.104 documents

host+1
253.649 documents



nu.nl
57.913 documents

WebARTist

Web Archive Temporal Information Search Tools

mubarak Search Results order: relevance

Results | Timeline | Aggregations | Visualizations

[1] Achtergrond: **Mubarak** regeerde als moderne farao | nu.nl/egypte | Het laatste nieuws het eerst op nu.nl
page: original archived page: National Library, Internet Archive



hulp van de VS. Maar **Mubaraks** critici menen dat hij het land niet stabiel hield, maar op zijn zachtst gezegd stagneerde met het corrupte partijapparaat van zijn Nationaal Democratische Partij (NDP). Mubarak kwam ter wereld in 1928 in een dorpje in de Nijldelta. Hij bezocht de militaire academie en maakte carrière bij de luchtmacht, vooral tijdens de oorlog van 1973 tegen Israël. President Anwar Sadat benoemde Mubarak in 1975 tot vicepresident. Mubarak veranderde de prowesterse koers niet toen hij in oktober 1981 opvolger werd van de doodgeschoten Sadat. De noodtoestand die Sadat had uitgeroepen, trok Mubarak ... In de jaren '90 sloeg **Mubaraks** regime met succes islamitisch extremisme en terrorisme neer. Hij werd geholpen door een enorm aantal verklikkers en doordat de 83 milj



Nu.nl : troonswisseling url: nu.nl/egypte/2824454/achtergrond-mubarak-regeerde-als-moderne-farao.html category: Recht, Overheid
size: 5068 crawl: 01/06/2012 page: 01/06/2012 outlinks: 5

[2] Moslimbroeder leidt in tellingen Egypte | nu.nl/egypte | Het laatste nieuws het eerst op nu.nl
page: original archived page: National Library, Internet Archive



De beweging eiste eerder al de overwinning op en verklaarde Mursi tot "de eerste president van de republiek die door het volk is gekozen". Het kamp van de tegenstander, Ahmed Shafiq, claimde echter eveneens de winst. Een lid van het kiescomité bevestigde maandag dat Mursi aan kop ligt, maar zei dat de stemmen nog lang niet allemaal geteld zijn. "De resultaten die het campagneteam van Mursi publiceert stroken in grote lijnen met de werkelijkheid", zei hij. In 98 procent van de stembureaus zouden de stemmen zijn geteld. Waarnemers van de Broederschap ...



Nu.nl : troonswisseling url: nu.nl/egypte/2837794/moslimbroeder-leidt-in-tellingen-egypte.html category: Recht, Overheid size: 4888
crawl: 18/06/2012 page: 18/06/2012 outlinks: 7

[3] Doodstraf geëist tegen **Mubarak** | nu.nl/onrust midden-oosten | Het laatste nieuws het eerst op nu.nl
page: original archived page: National Library, Internet Archive

WebARTist
Web Archive Temporal Information Search Tools

mubarak Search Results order: date

Results | Timeline | Aggregations | Visualizations

/2011 09/09/2011 15/11/2011 22/11/2011 **23/11/2011** 24/11/2011 02/01/2012 06/01/2012

NTC maakt namen regering Libië bekend | nu.nl/onrust midden-oosten | H...



"Heel Libië is vertegenwoordigd", aldus interim-premier Abdurrahim El-Keib tijdens een persconferentie. Libië beloonde de strijders die afgelopen zaterdag Saif al-Islam Kaddafi oppakten. Hun aanvoerder Osama Al-Juwali wordt minister van Defensie. Minister van Buitenlandse Zaken wordt Ashour Bin Hayal. Hij is een vrij onbekende diplomaat uit het oosten van het land. Vooraf leek de diplomaat Ibrahim Dabbashi favoriet voor de positie. Hij mobiliseerde andere diplomaten om het tegen Kaddafi op te nemen. De benoeming van Hayal heeft vermoedelijk te maken met regionale twisten in Libië. Andere sleutelposten zijn voor Hassan Ziglam (Fina...

Nu.nl : troonswisseling crawl: 23/11/2011 page: 23/11/2011 category: Recht, Overheid

WebARTist
Web Archive Temporal Information Search Tools

mubarak Search

Results | Timeline | Aggregations | Visualizations

top sites:	*top sites by name:*	*top unesco categories:*	*top dates:*	*top crawldates:*
nu.nl (46)	Nu.nl : troonswisseling (46)	Recht, Overheid (48)	11/2010 (2)	11/2010 (2)
cidi.nl (43)	Centrum Informatie en Documentatie Isra?l (CIDI) (43)	Geschiedenis (47)	02/2011 (3)	01/2011 (3)
retecool.com (27)	Amazing Retecool (27)	Algemeen (29)	03/2011 (1)	02/2011 (29)
krakatau.nl (25)	Krakatau (25)	Dichtkunst (25)	05/2011 (20)	03/2011 (1)
arnongrunberg.com (4)	ArnonGrunberg.com (4)	Literatuurgeschiedenis (4)	07/2011 (1)	04/2011 (7)
fnvsport.nl (4)	FNV Sport (4)	Sociologie, Statistiek (4)	09/2011 (19)	05/2011 (23)
innl.nl (3)	Nationaal Historisch Museum (3)	Politieke wetenschappen (4)	10/2011 (5)	07/2011 (1)
nederlandtegenterrorisme.nl (2)	Nederland tegen terrorisme (2)	Stedenbouw, Kunst (3)	11/2011 (26)	08/2011 (1)
nos.nl (2)	NOS.nl (2)	Godsdienst (3)	01/2012 (6)	09/2011 (20)
wikipedia.org (2)	Wikipedy [Wikipedia in de Friese taal] (2)	Wijsbegeerte, Psychologie (2)	03/2012 (1)	10/2011 (11)
blogspot.com (1)	Blog van de Historische Vereniging Noordoost-Friesland (1)		05/2012 (5)	11/2011 (27)
cbc.ca (1)	Christians for Truth Christenen voor de Waarheid (1)		06/2012 (8)	12/2011 (1)
cftnederland.nl (1)	Irgoen Olei Holland (IOH) (1)		07/2012 (3)	01/2012 (6)
co.uk (1)	Vereniging De Vrije Gedachte (1)		08/2012 (2)	03/2012 (1)
devrijegedachte.nl (1)	Katholiek Nieuwsblad (1)		09/2012 (1)	04/2012 (6)
examiner.com (1)	Kerk in Actie (1)		11/2012 (3)	05/2012 (6)
huffingtonpost.com (1)	Stichting Kerk in Nood (1)		12/2012 (1)	06/2012 (8)
katholieknieuwsblad.nl (1)	NCDO (1)			07/2012 (3)
kerkinactie.nl (1)	Katholieke Vereniging voor Oecumene (1)			08/2012 (2)
kerkinnood.nl (1)	Women, Inc. (1)			09/2012 (1)
msn.com (1)	World Press Photo (1)			10/2012 (9)
ncdo.nl (1)				11/2012 (6)
nd.nl (1)				12/2012 (1)
oecumene.nl (1)				
vn.nl (1)				
womeninc.nl (1)				
worldpressphoto.org (1)				

top locations:	*top outlinks per site:*	*top outlinks:*	*top occurring words:*	*included in seed list:*
Den Haag (28)	arnongrunberg.com→feedburner.com (14)	google.com (378)	gedichten (791)	yes (169)
De Weere (16)	arnongrunberg.com→nytimes.com (4)	blogspot.com (353)	the (612)	no (6)
Rotterdam (12)	arnongrunberg.com→onestat.com (4)	woestenledig.com (275)	zie (437)	
Beers (4)	arnongrunberg.com→onestatfree.com (4)	zie.nl (161)	krakatau (419)	
De Woude (4)	arnongrunberg.com→twitter.com (2)	twitter.com (126)	redactie (404)	

WebARTist

Web Archive Temporal Information Search Tools

mubarak Search

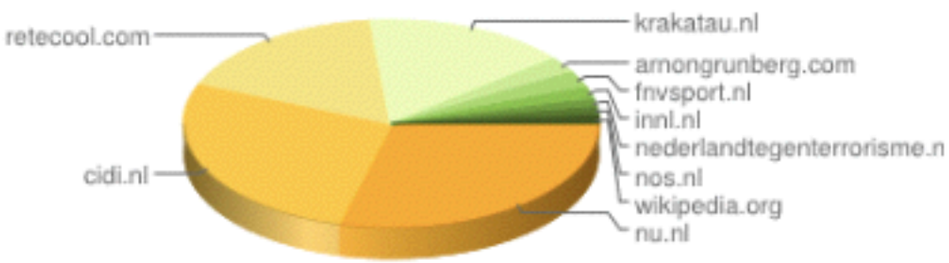
Results | Timeline | Aggregations | Visualizations

occurring words:

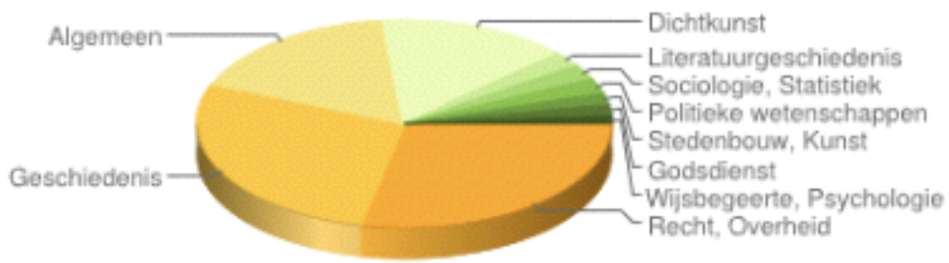
alle alleen arabische are augustus broederschap categorie categories comments dag dagen duprie egypte egyptische even februari for gaan gaat gedicht gedichten genoeg goede gratis hamas iran israel israelische jaar janus kaddafi krakatau laatste land laten lees leven libi libische maart man mensen militaire moslim mubarak nieuwe nieuws nooit not onder opstand politieke president reacties rebellen redactie regime september sinds staat steeds tags that tijd toe tripoli verder volgens vraag vrienden vrouw waar weer wereld weten willen zand zegt zie zien

statistics:

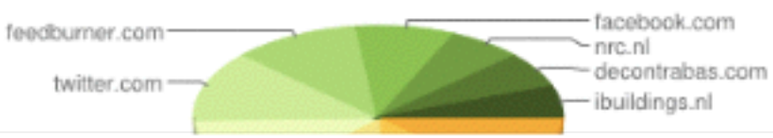
site



unescocode



outlinks



seedlist



6.2.2 DMI "Web Archiving Day" (2013)

mubarak

Search

Results | Timeline | Aggregations | Visualizations

175 results for mubarak, 175 filtered/shown results. (homepage+1 collection) Edit settings Help

alle alleen arabische are augustus broederschap categorie categories comments dag dagen duprie egypte egyptische even februari for gaan gaat gedich
gedichten genoeg goede gratis hamas iran israel israelische jaar janus kaddafi krakatau laatste land laten lees leven libi libische maart man mensen
militaire moslim mubarak nieuwe nieuws nooit not onder opstand politieke president reacties rebellen redactie regime september sinds staat steeds tags th
toe tripoli verder volgens vraag vrienden vrouw waar weer wereld weten willen zand zegt zie zien

statistics:

site

unescocode



outlinks



seedlist



6.2.2 DMI "Web Archiving Day" (2013)







6.2.2 DMI "Web Archiving Day" (2013)







Remarks researchers:

- "looking at data rather than single sites"
- "supports the shift to studying Web archives through queries"
- "aggregate views and bar graphs are extremely useful"

6.2.2 DMI "Web Archiving Day" (2013)

- Interests New Media researchers:
  - **"derive periodizations of the Web"** (Web history)
    - e.g. 'the rise of social media'
  - **"source hierarchy"** (dominant sources in archive)
    - e.g. 'financial crisis'
  - **"keyword uptake"** (terms over time)
    - e.g. 'geenstijl language in archive'
  - **"accidental"/"incidental" archiving**
    - e.g. 'the guilty pleasures of the Web of innocence'

6.2.2 DMI "Web Archiving Day" (2013)

- Suggestions researchers for extensions WebARTist:
  - **selections:** e.g. sampling en subsets (site, lists)
  - **analysis:** e.g. comparisons (resultset differences)
  - **collections:** e.g. create own collections and annotations
  - **transparency:** e.g. selection procedures, algorithms and (in)completeness

6.3. From *search* to *research* tools

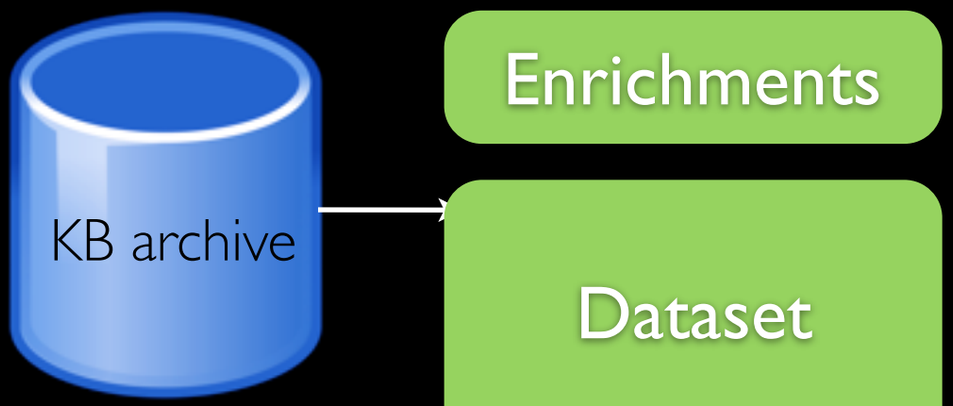
(Re)search

- From Wayback Machine to searchable archive
  - Big step forward.
    - from URL browsing to ‘full-text’ search: new research opportunities
  - Potential support various steps research process, e.g. exploration and corpus creation
- But, also limitations…

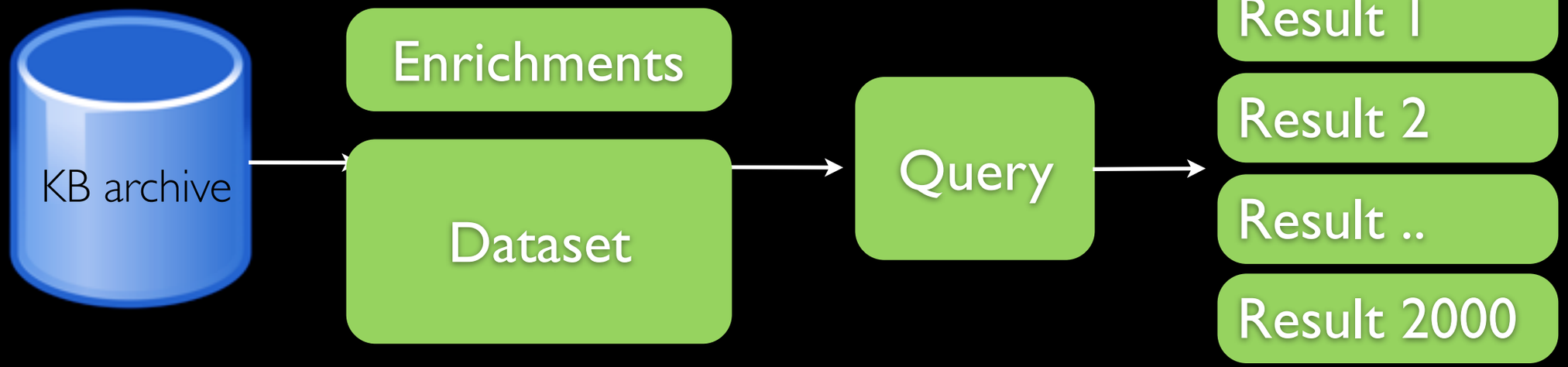
WebARTist zoekmachine



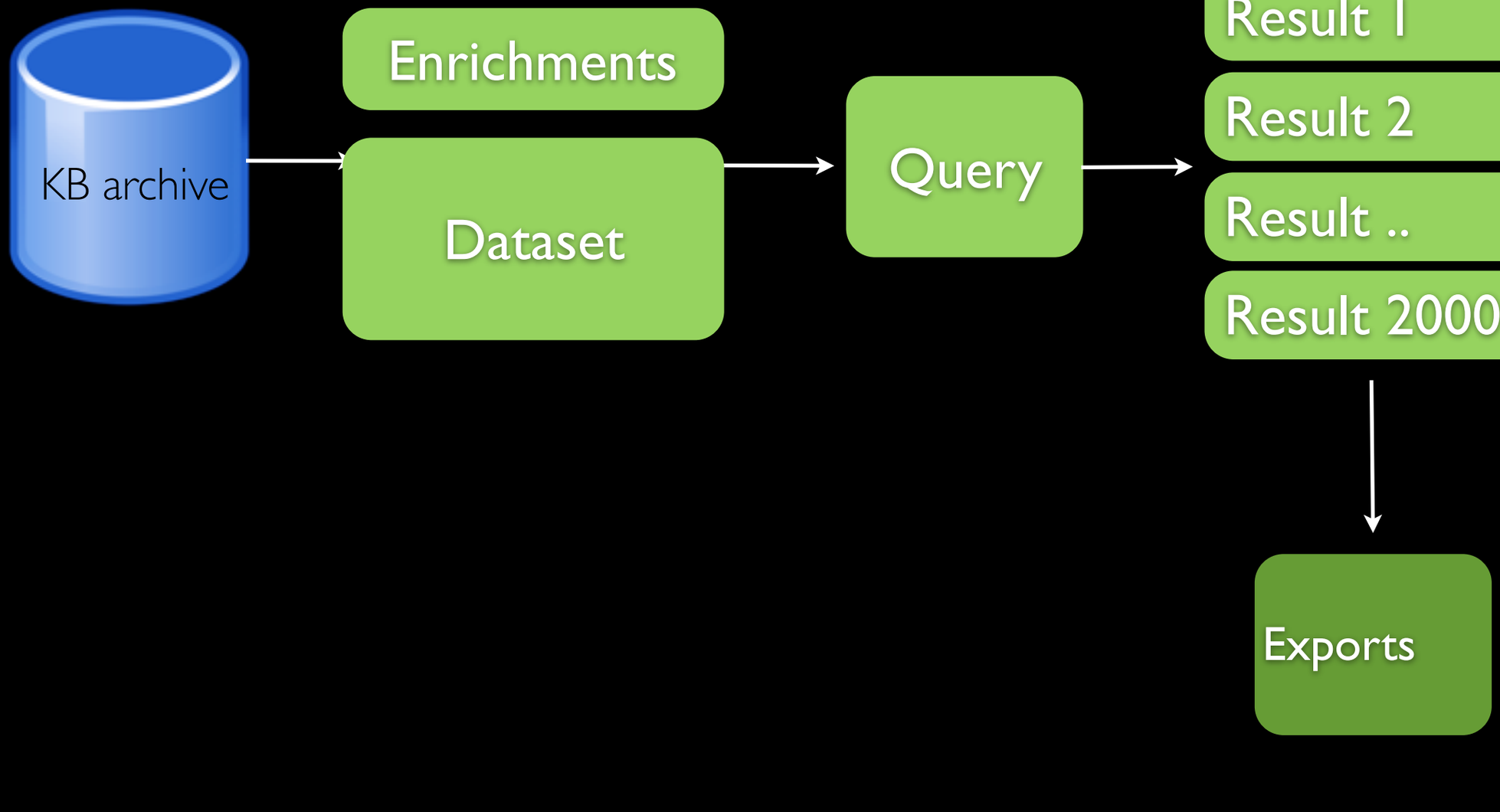
WebARTist zoekmachine



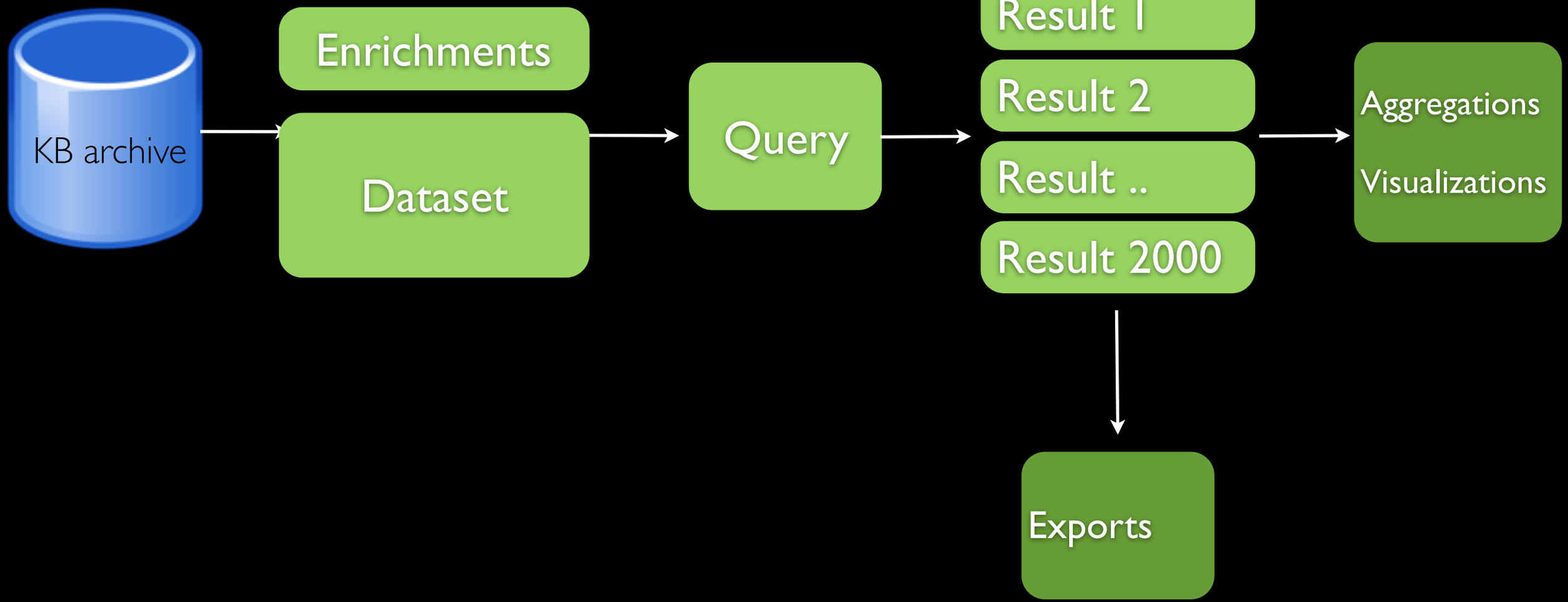
WebARTist zoekmachine



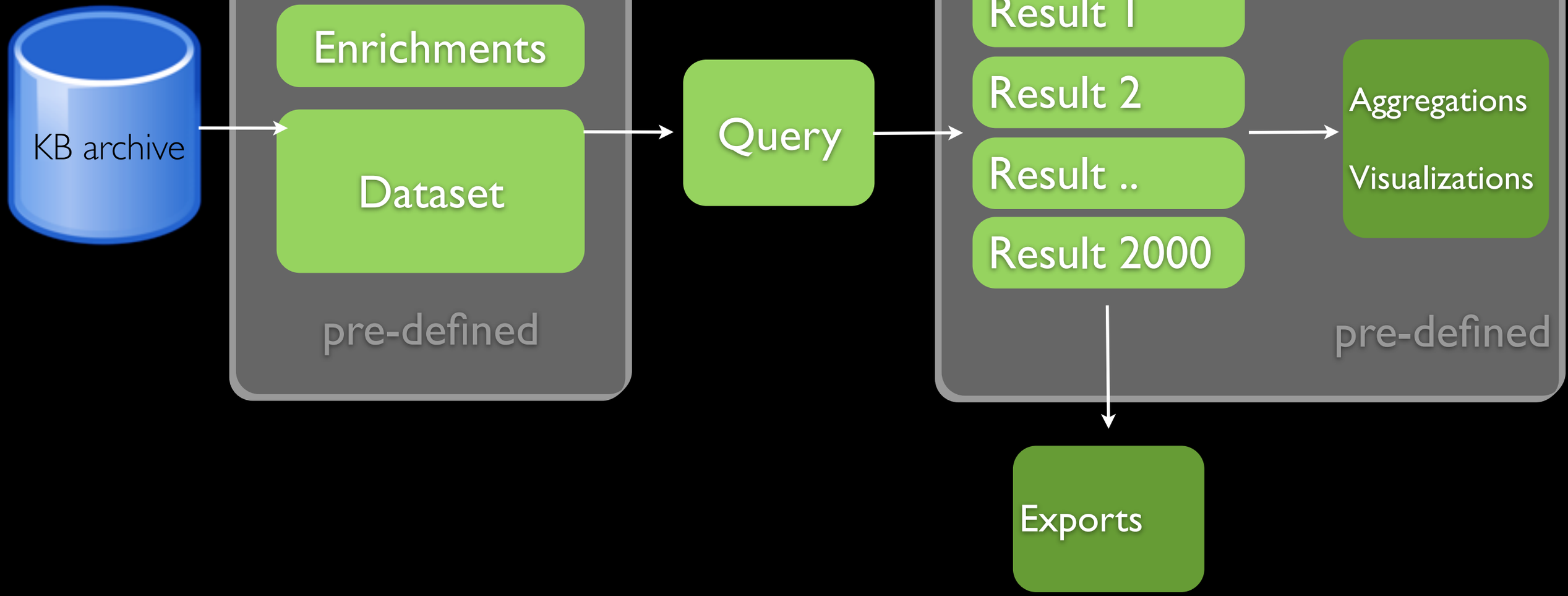
WebARTist zoekmachine



WebARTist zoekmachine



WebARTist zoekmachine



Limitations search tools

1. Datasets & enrichments **predefined**
2. Search engines and functionality **predefined**

- "Trade-off": support *universal* versus *specific* tasks and analyses
- A search system cannot support áll specific tasks

Limitations search tools

1. Datasets & enrichments **predefined**
2. Search engines and functionality **predefined**

- "Trade-off": support *universal* versus *specific* tasks and analyses
- A search system cannot support áll specific tasks

Limitations search tools

1. Datasets & enrichments **predefined**
2. Search engines and functionality **predefined**
    - "Trade-off": support *universal* versus *specific* tasks and analyses
    - A search system cannot support áll specific tasks

- We need more **flexibility** to support the research process!

Limitations search tools

1. Datasets & enrichments **predefined**
2. Search engines and functionality **predefined**
    - "Trade-off": support *universal* versus *specific* tasks and analyses
    - A search system cannot support áll specific tasks

- We need more **flexibility** to support the research process!
    - in terms of *data processing*, *underlying systems* & *access interfaces*

6.3.1 More flexibility: data



use "Hadoop" computing power to build datasets, for enrichments, and high-level analysis

6.3.1 Example

Show what is (not) in the archive: Introduction

- Due to restrictions in crawling depth, frequency and retrictive selections policies, large parts of the web are unarchived

- Therefore, issues may arise when performing Web archive research

- Would it be possible to uncover and generate representations of unarchived web contents?



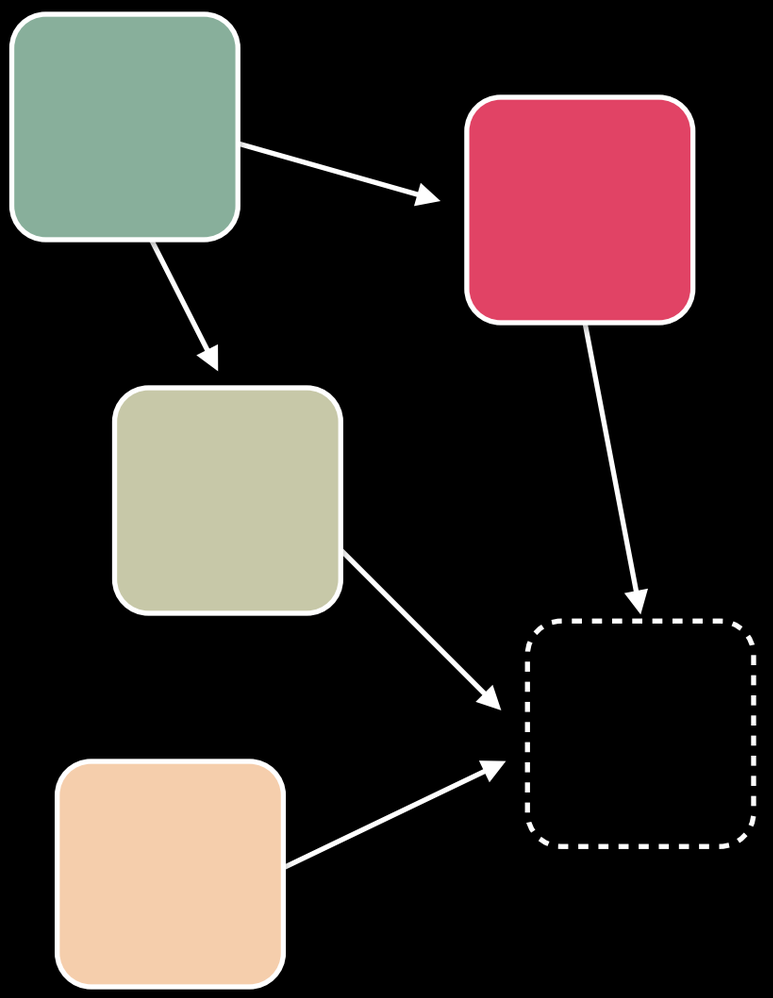
6.3.1 Example

Show what is (not) in the archive: Approach

- Defining property of the Web: hyperlinked structure
  - links: source, destination & anchor text

- We use the information from the link structure to generate representations of *unarchived* pages and sites

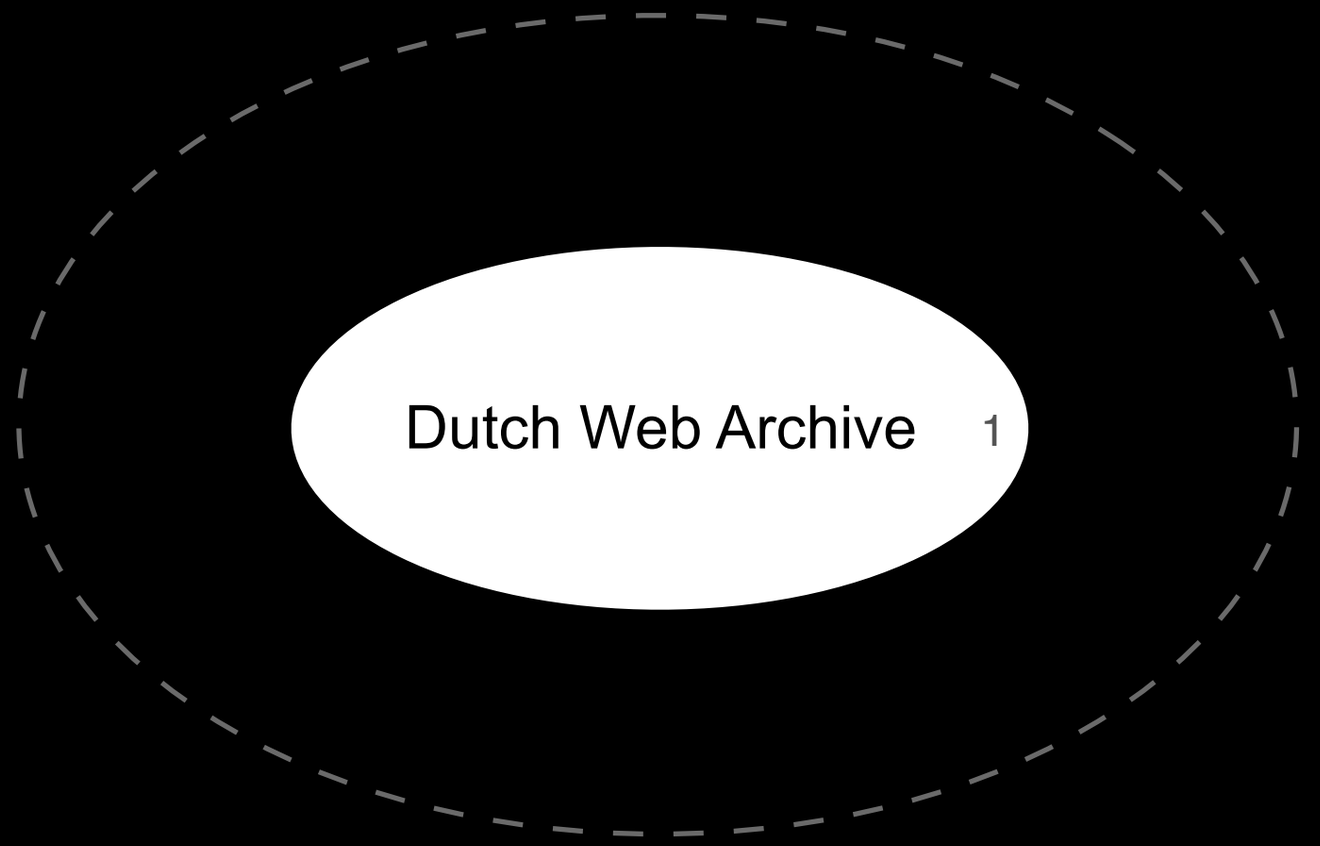
- Extraction and analysis via Hadoop computer cluster at SURFsara



6.3.1 Example

Archived content (2012)

1. Contents in seedlist
   - 10.2M unique pages
   - 3,413 unique domains

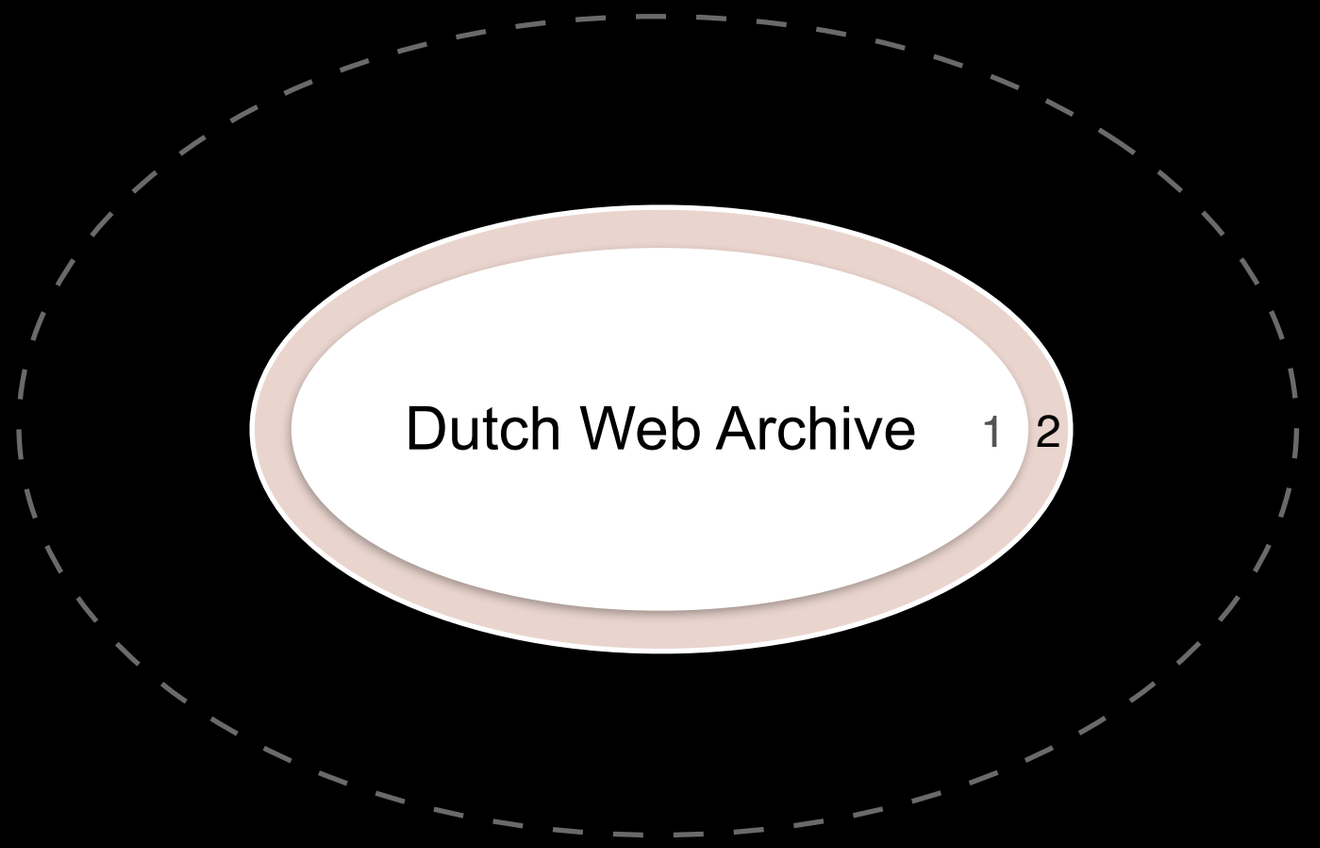


6.3.1 Example

Archived content (2012)

1. Contents in seedlist
   - 10.2M unique pages
   - 3,413 unique domains

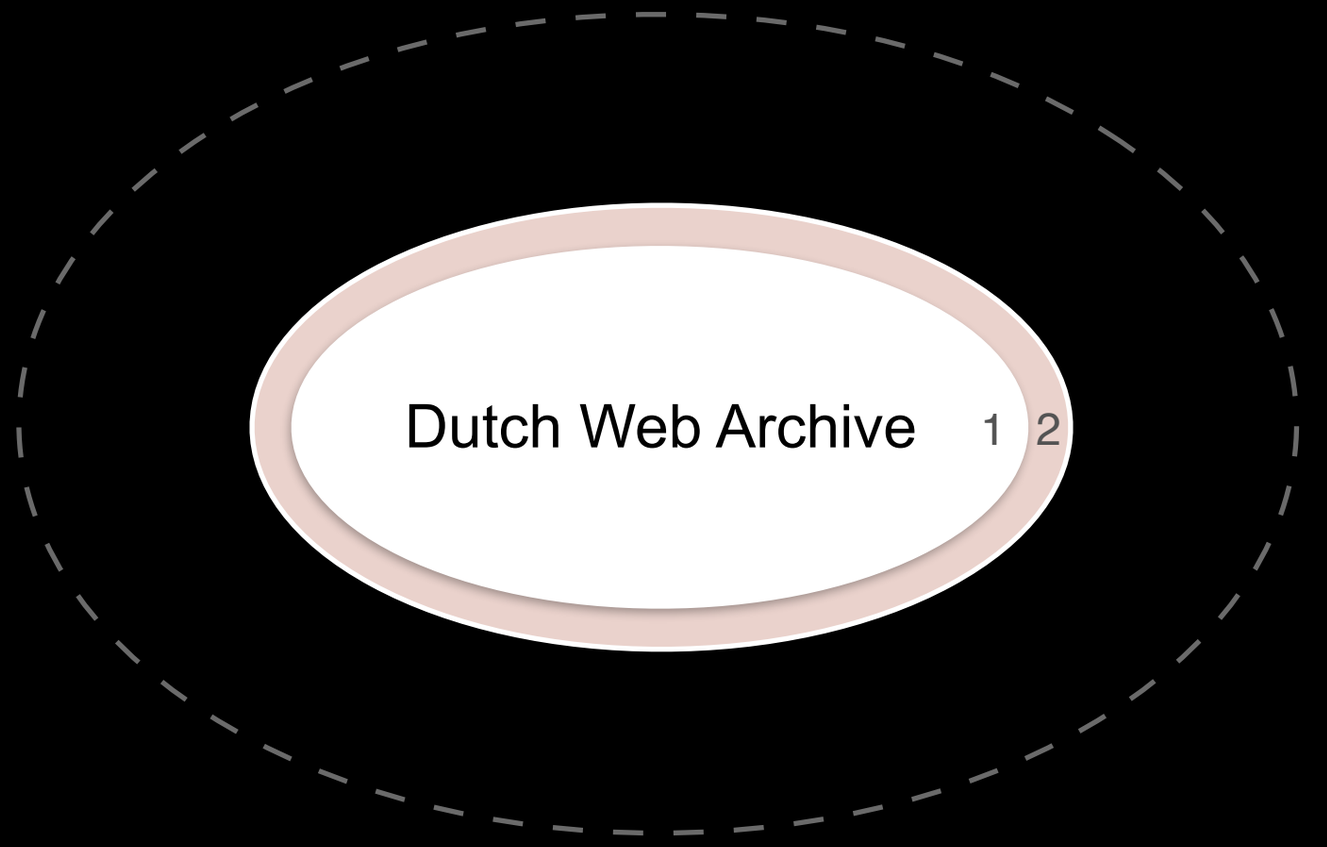
2. Contents **not** in seedlist
   - 0.9M unique pages
   - 30,367 unique domains



6.3.1 Example

*Un*archived content (2012)

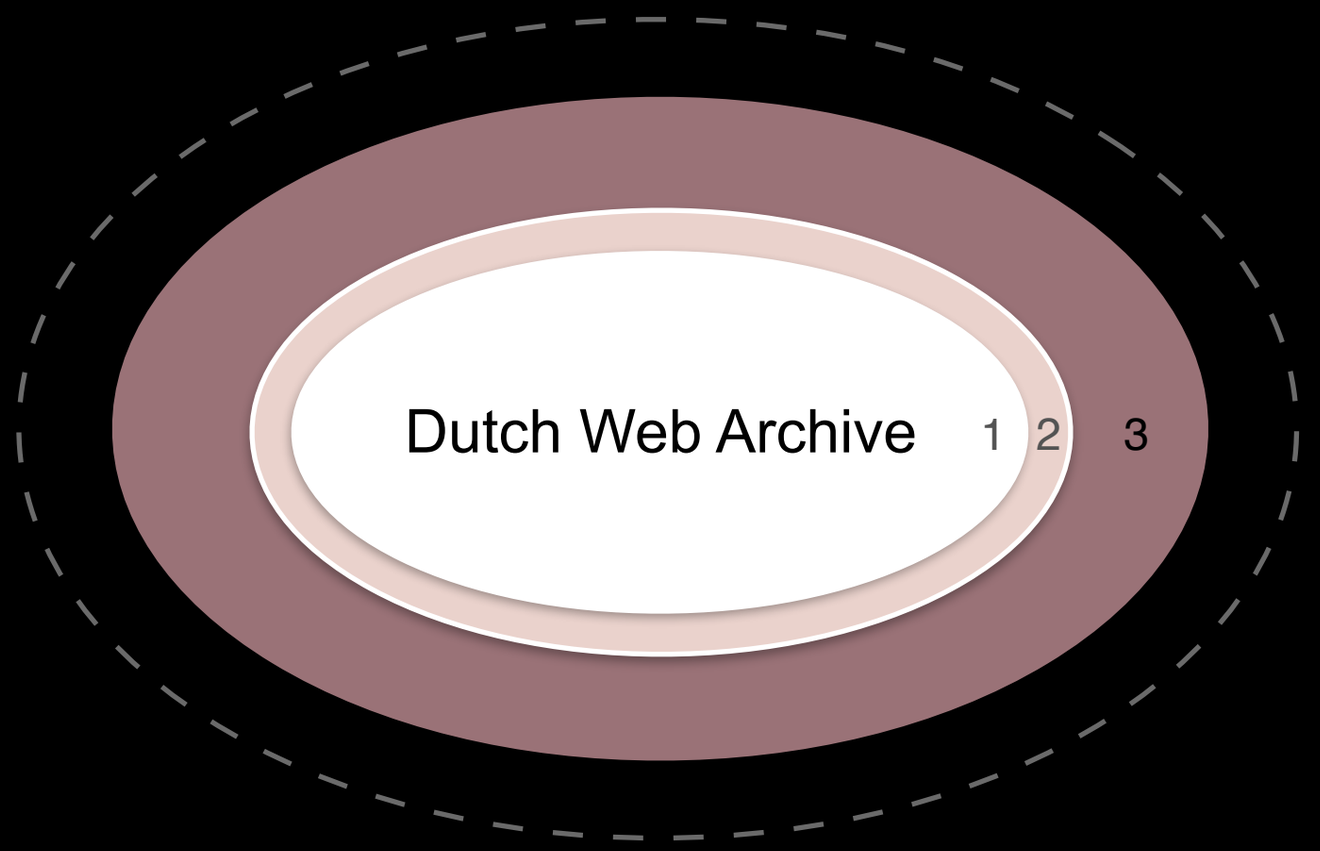
3. Inner aura
   - 5.5M unique pages
   - 3,019 unique domains



6.3.1 Example

*Un*archived content (2012)

3. Inner aura
   - 5.5M unique pages
   - 3,019 unique domains

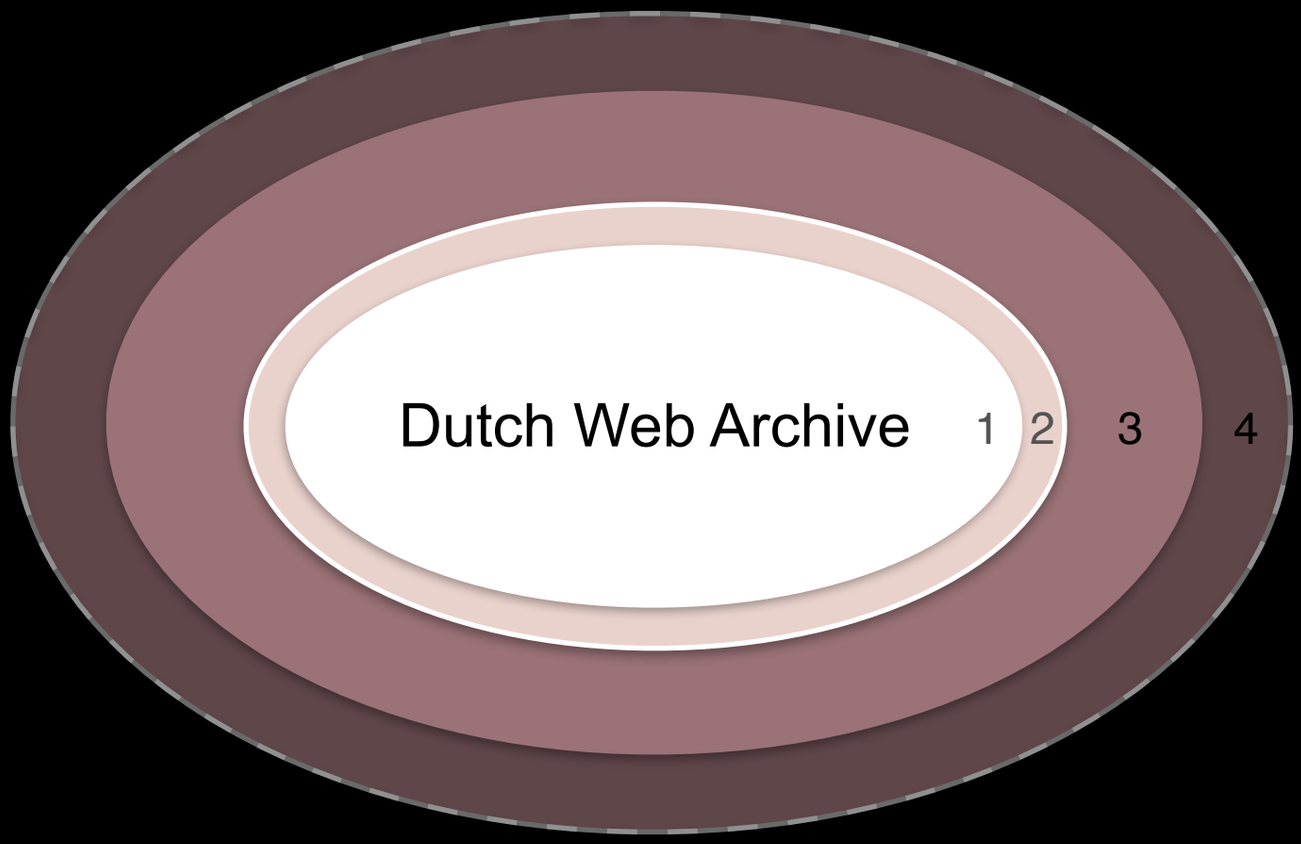


6.3.1 Example

*Un*archived content (2012)

3. Inner aura
   - 5.5M unique pages
   - 3,019 unique domains

4. Outer aura
   - 5.2M unique pages
   - 369,721 unique domains



6.3.1 Example

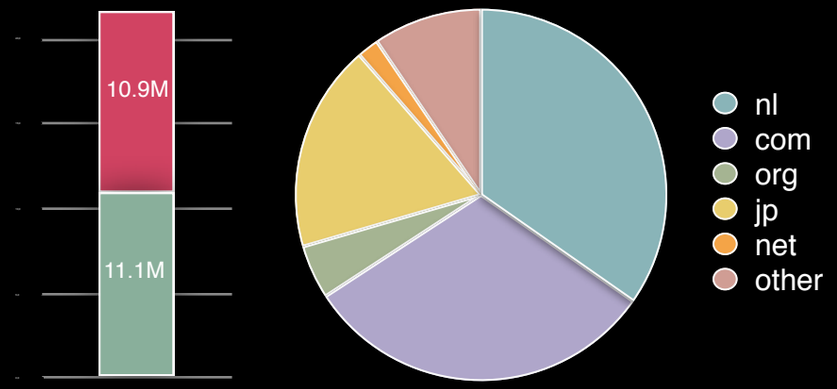
Show what is (not) in the archive: Conclusion

- Read more in [Huurdeman et al, 2015]

6.3.1 Example

Show what is (not) in the archive: Conclusion

- Evidence high number unarchived pages
  - dramatically increasing archival coverage

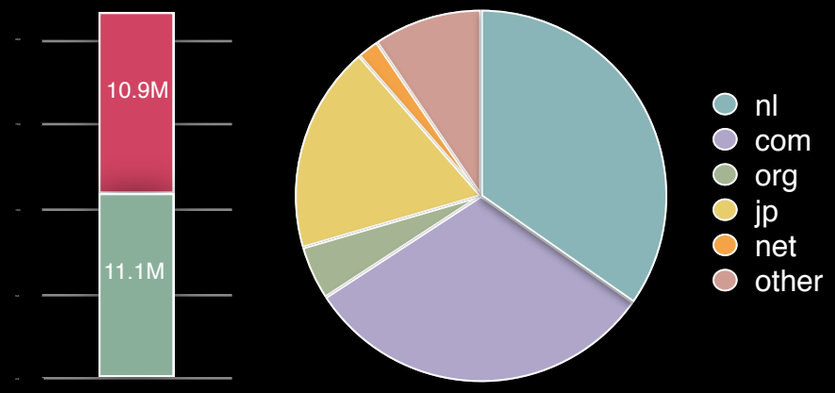


- Read more in [Huurdeman et al, 2015]

6.3.1 Example

Show what is (not) in the archive: Conclusion

- Evidence high number unarchived pages
  - dramatically increasing archival coverage



- Skewed distribution of generated descriptions
  - popular pages have more terms
  - richness tapers off quickly

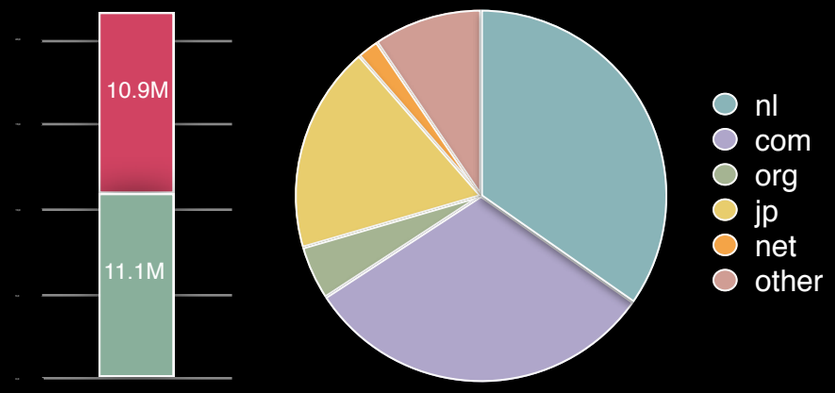


- Read more in [Huurdeman et al, 2015]

6.3.1 Example

Show what is (not) in the archive: Conclusion

- Evidence high number unarchived pages
  - dramatically increasing archival coverage



- Skewed distribution of generated descriptions
  - popular pages have more terms
  - richness tapers off quickly



- Succint representation generally rich enough to identify pages
  - in a known item search setting
  - can be used to contextualize search tools



- Read more in [Huurdeman et al, 2015]

WebARTist

Web Archive Temporal Information Search Tools

onderzoeksdata Search

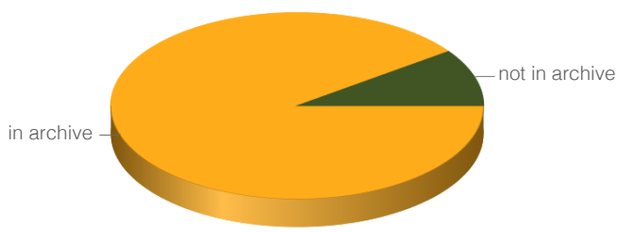
Results | Timeline | Aggregations | Visualizations

347 archived results for onderzoeksdata, 33 unarchived results. *Edit settings* *Help*

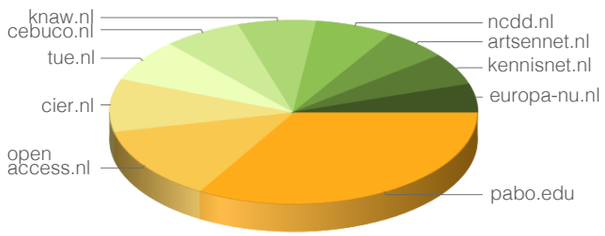
occurring words:

aanvragen access afgelopen artikel belang beschikbaar bijdragen binnen buning burgers centraal cier cock creativiteit data december dient digital domein duurzaam duurzame eerste eijsvogels europese for forum gericht groot ict ier informatie informatiemaatschappij infrastructuren innovatie intellectuele jubileumcongres knowledge maatschappij media moeten mogelijk moment nationale nederlandse nieuwe noot novay nwo onder onderzoeksdata onderzoek ontwikkeling open praktijkgericht private proces professor project publiek publieke read research ringnalda steeds stimuleren stond that toegang tweede verband

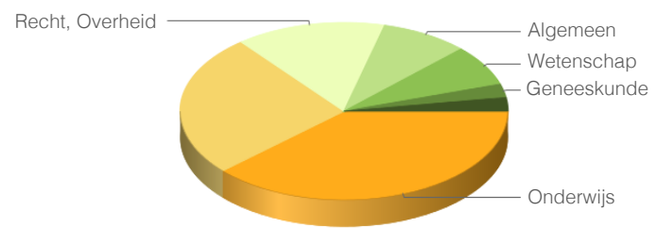
statistics:



archived results

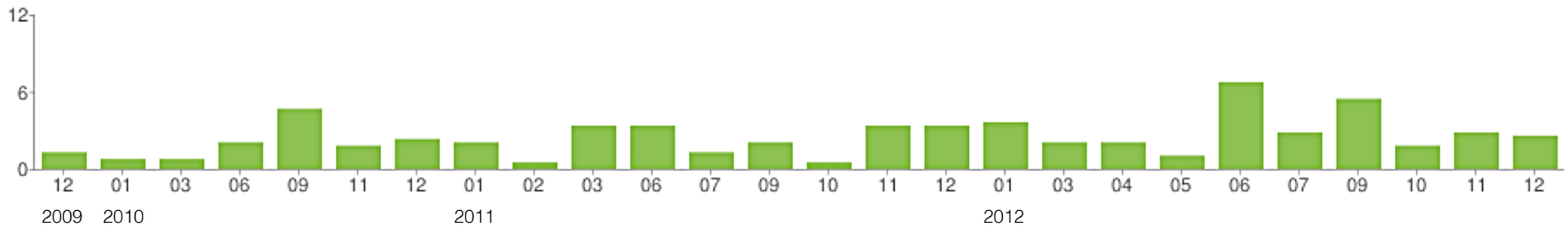


websites



unesco categories

temporal distribution:

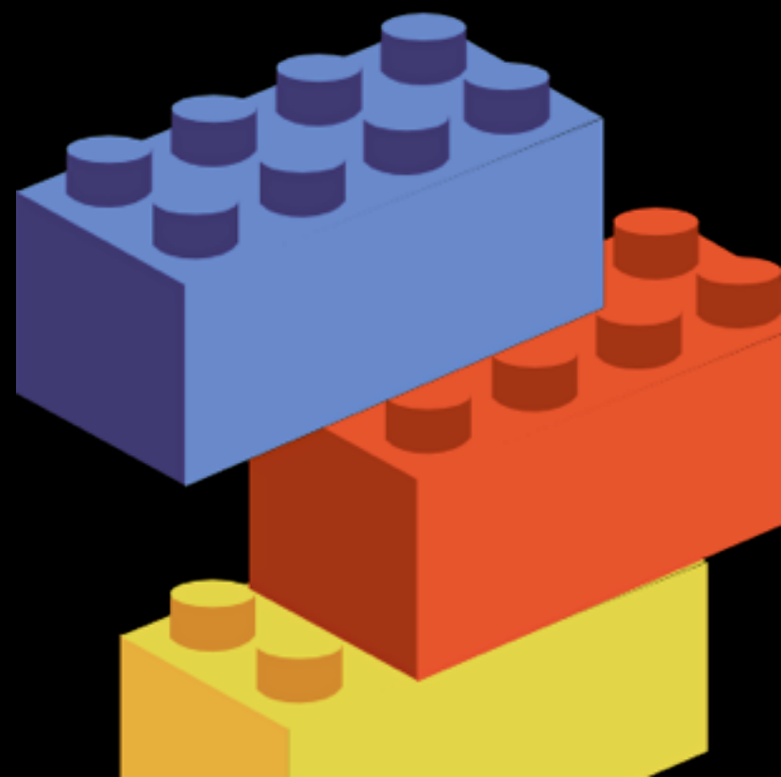


Potential integration into access interfaces

6.3.2 More flexibility: system

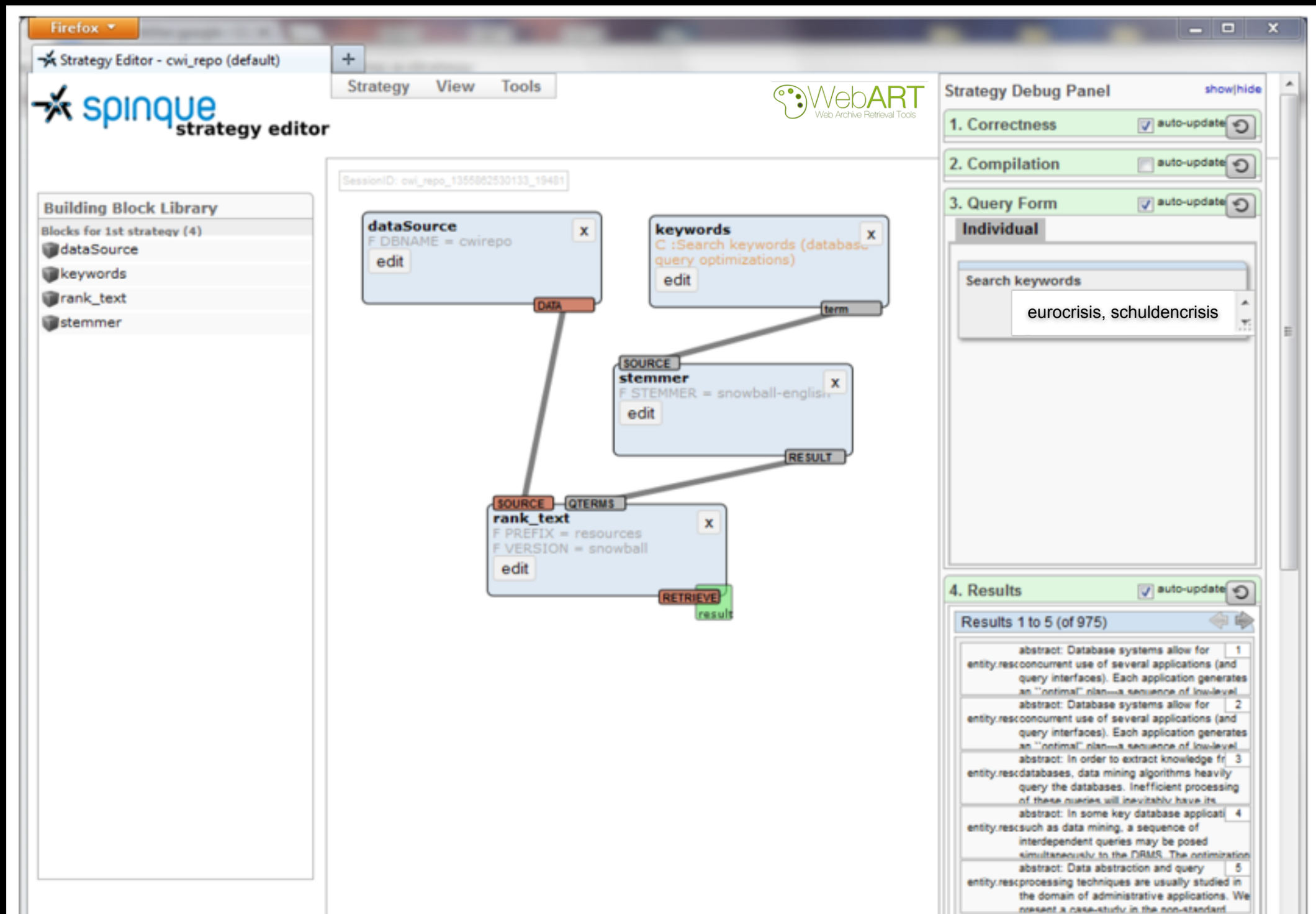
- Creating *customizable* systems, or *toolmakers' tools*
- Offer 'building blocks'
  - *custom* search engines
    - search strategies
    - complex queries

[Kamps '11, de Vries e.a. '10]



6.3.2 Example

Spinque Search Strategies



6.3.3 More flexibility: interface

- How to integrate all proposed (research) functionality in search systems?
  - *browse, search, select, analyze, visualize, …*
  - *usability* problems
  - cognitive overload

edu.umd.cs.hcil.iopener 201007282301

File Edit View Navigate Tools Window Help

References

File Edit View BibTeX Tools Web search Plugins Options Help

Groups

- CoNLL 2007 shared task
  - false positives @99
  - false negatives @99
- Evolution

Settings

Search

author=eisner

Search Specified Field(s)

Clear

- Incremental
- Float
- Filter
- Show results in dialog
- Global search

Settings

DP.meta.bib

#	Acl-id	Author	Title	Year	Booktitle
1	W07-2220	Nivre	Data-Driven Dependency Parsing across Languages an...	2007	Tenth International Conferenc...
2	D07-1096	Nivre, Hall, Kubler, McDonald, Nilsson, ...	The CoNLL 2007 Shared Task on Dependency Parsing	2007	2007 Joint Conference on Em...
3	D07-1098	Duan, Zhao and Xu	Probabilistic Parsing Action Models for Multi-Lingual De...	2007	2007 Joint Conference on Em...
4	D07-1101	Carreras	Experiments with a Higher-Order Projective Dependency...	2007	2007 Joint Conference on Em...
5	D07-1111	Sagae and Tsujii	Dependency Parsing and Domain Adaptation with LR M...	2007	2007 Joint Conference on Em...
6	D07-1112	Dredze, Blitzer, Talukdar, Ganchev, Grac...	Frustratingly Hard Domain Adaptation for Dependency P...	2007	2007 Joint Conference on Em...
7	D07-1119	Attardi, Dell'Orletta, Simi, Chanev and Ci...	Multilingual Dependency Parsing and Domain Adaptatio...	2007	2007 Joint Conference on Em...
8	D07-1120	Bick	Hybrid Ways to Improve Domain Independence in an ML...	2007	2007 Joint Conference on Em...
9	D07-1121	Canisius and Sang	A Constraint Satisfaction Approach to Dependency Parsi...	2007	2007 Joint Conference on Em...
10	D07-1122	Chen, Zhang and Isahara	A Two-Stage Parser for Multilingual Dependency Parsing	2007	2007 Joint Conference on Em...
11	D07-1129	Shimizu and Nakagawa	Structural Correspondence Learning for Dependency Pa...	2007	2007 Joint Conference on Em...
12	L08-1337	Nilsson and Nivre	MaltEval: an Evaluation and Visualization Tool for Depen...	2008	LREC
13	I08-2097	Kawahara and Uchimoto	Learning Reliability of Parses for Domain Adaptation of ...	2008	Proceedings of the Third Intern...
14	C08-1015	Chen, Wu and Isahara	Learning Reliable Information for Dependency Parsing A...	2008	International Conference On C...
15	C69-0201	Shank and Tesler	A Conceptual Dependency Parser For Natural Language	1969	International Conference On C...
16	C80-1003	Hitaka and Yoshida	A Syntax Parser Based On The Case Dependency Gram...	1980	International Conference On C...
17	P83-1003	Thompson	Crossed Serial Dependencies: A Low-Power Parseable ...	1983	Annual Meeting Of The Associ...
18	P84-1080	Nelimarkka, Jappinen and Lehtola	Two-Way Finite Automata And Dependency Grammar: A ...	1984	International Conference On C...
19	C86-1109	Jappinen, Lehtola and Valkonen	Functional Structures For Parsing Dependency Constrai...	1986	International Conference On C...

Status: Updated group selection.

Out-Cite Text (D07-1098: Inproceedings:Duan07Prob...

time and comparable performances for multi-lingual dependency parsing (Nivre et al., 2006). Deterministic parser regards parsing as a sequence of parsing actions that are taken step by step on the input sentence. Parsing actions construct dependency relations between words. Deterministic dependency parser does not score the entire dependency tree as most of state-of-theart parsers. They only stepwisely choose the most probable parsing action. In this paper, to globally model parsing actions of all steps that are taken on the input sentence, we propose two kinds of probabilistic parsing action models that can compute the entire dependency trees probability. Experiments are evaluated on diverse data set of 10 languages provided by CoNLL 2007 shared-task (Nivre et al., 2007). Results show that our probabilistic parsing action models outperform the original deterministic dependency parser. We also present a general error analysis across a wide set of languages plus a detailed error analysis of Chinese. Next we briefly introduce the original deterministic dependency parsing algorithm that is a basic component of our models. 2 Introduction of Deterministic Dependency Parsing There are mainly two representative deterministic dependency parsing algorithms proposed respectively by Nivre (2003), Yamada and Matsumoto (2003). Here we briefly introduce Yamada and Matsumotos algorithm, which is adopted by our models to illustrate deterministic dependency

In-Cite Text (14 entries)

important differences in the distribution of errors associated with the two models.

Experiments are evaluated on diverse data set of 10 languages provided by CoNLL 2007 shared-task (Nivre et al., 2007).

1 Introduction The target of CoNLL 2007 shared task (Nivre et al., 2007) is to parse texts in multiple languages by using a single dependency parser that has the capacity to learn from treebank data.

4 Experiments and Results Experiments are carried on 10 languages provided by CoNLL 2007 shared-task organizers (Nivre et al., 2007).

1 Introduction Like the CoNLL-2006 shared task, the 2007

In-Cite Summary: 'DP.B_CS-michael_summ4' (# 99-1...

Data-driven grammar-based parsers, such as Bick (2007), Schneider et al.

The systems by Sagae and Tsujii (2007), Attardi et al.

Sagae and Tsujii (2007) presented a procedure similar to a single iteration of co-training.

Compared with the adaptation methods of Sagae and Tsujii (2007) and Reichart and Rappoport (2007), our approach uses the information on word pairs in auto-parsed data instead of using the whole sentences as newly labeled data for training new parsers.

Carreras (2007) employs his own extension of Eisners algorithm for the case of projective trees and second-order models that include headgrandparent relations

Citation Network

Main menu | Overview | Rank Nodes | Attribute Nodes | Rank Edges | Plot Nodes | Find Communities | Edge Types | Graph Readability | Details | Settings

Rank Nodes

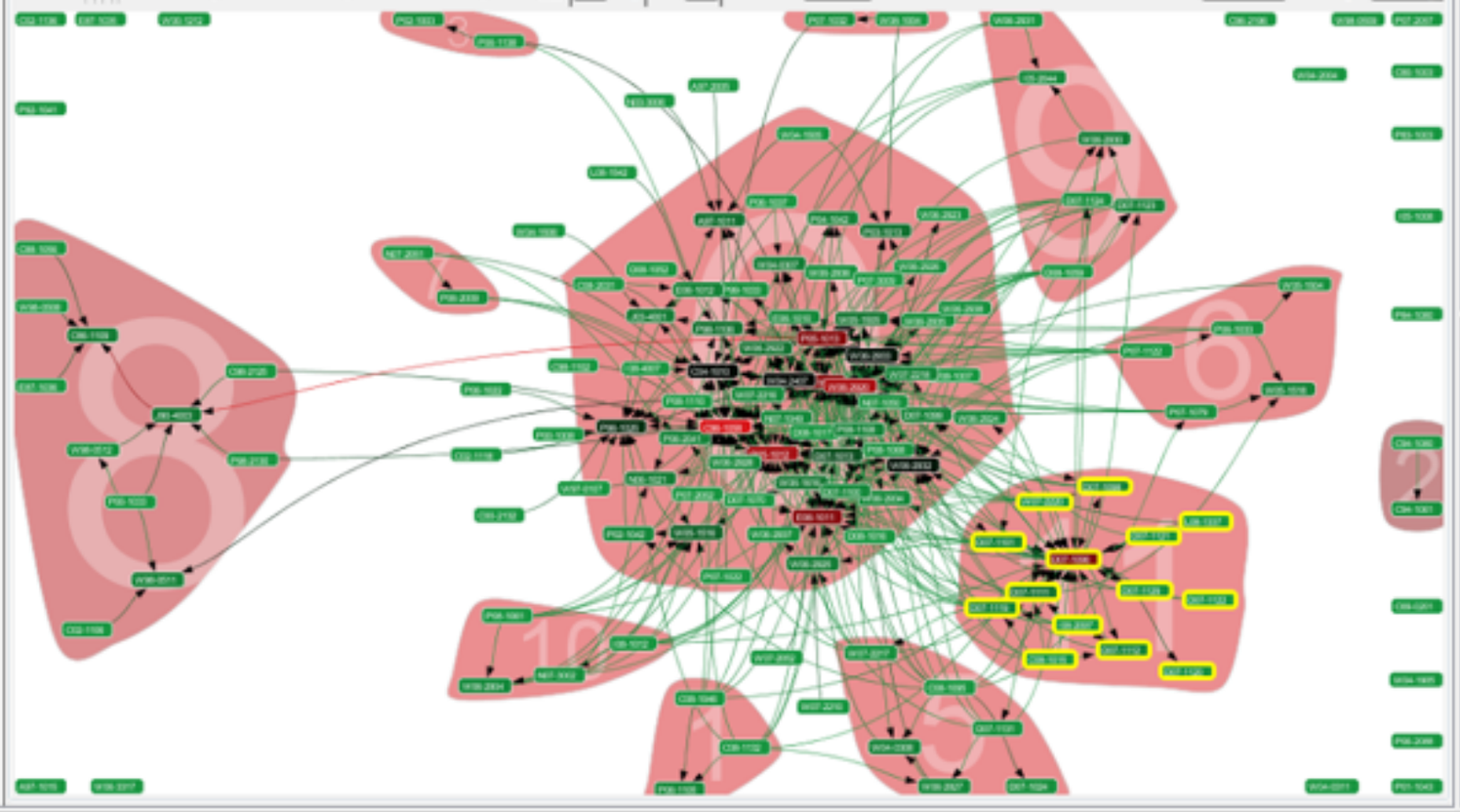
In-degree

The number of inbound edges connected to the node

Rank	Label
43.00000	C96-1058
34.00000	W06-2920
33.00000	P05-1012
27.00000	P05-1013
27.00000	E06-1011
27.00000	H05-1066
25.00000	D07-1096
19.00000	C04-1010
19.00000	W04-2407
18.00000	W06-2932
18.00000	W06-2933
14.00000	P96-1025
10.00000	D07-1013
7.00000	W05-1516
6.00000	P98-1106
6.00000	P03-1013
6.00000	A97-1011
6.00000	W06-2922
6.00000	D07-1111
5.00000	J90-4003
5.00000	W05-1505
5.00000	W06-1616
4.00000	J03-4001
4.00000	C86-1109

Network Visualization

Layout: | Font Size: | Edge Size: | Label Size: | ForceDir | SocNet | MidiComm | MidiCommNbr | Node | Node | Edge | Edge



6.3.3 More flexibility: interface

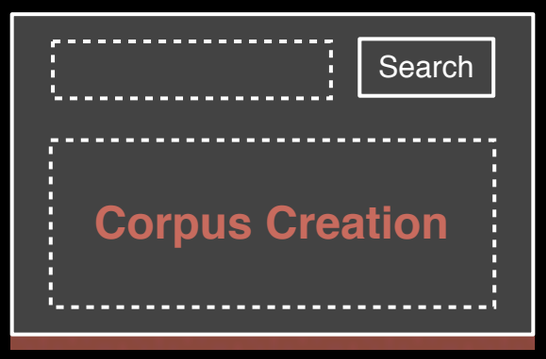
- **Traditional approach:** "Complex" interface integrating all functionality

6.3.3 More flexibility: interface

- **Traditional approach:** "Complex" interface integrating all functionality
- **Our Approach:** Divide functionality per (research) stage (Huurdeman & Kamps, 2014)

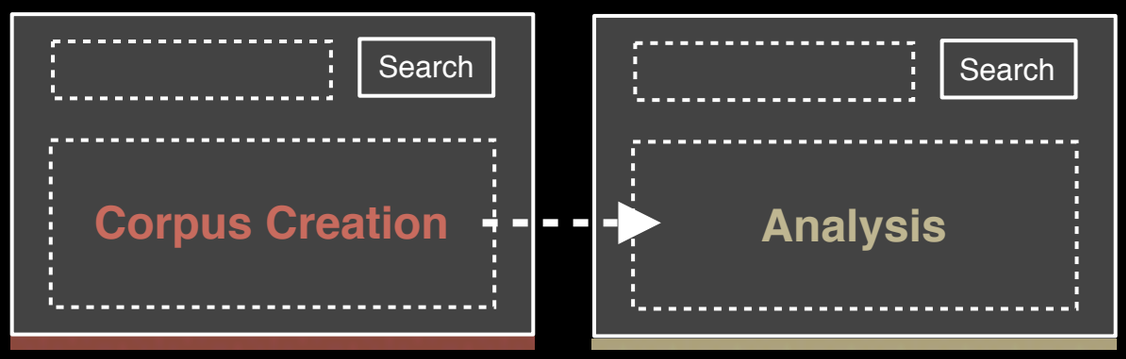
6.3.3 More flexibility: interface

- **Traditional approach:** "Complex" interface integrating all functionality
- **Our Approach:** Divide functionality per (research) stage (Huurdeman & Kamps, 2014)



6.3.3 More flexibility: interface

- **Traditional approach:** "Complex" interface integrating all functionality
- **Our Approach:** Divide functionality per (research) stage (Huurdeman & Kamps, 2014)



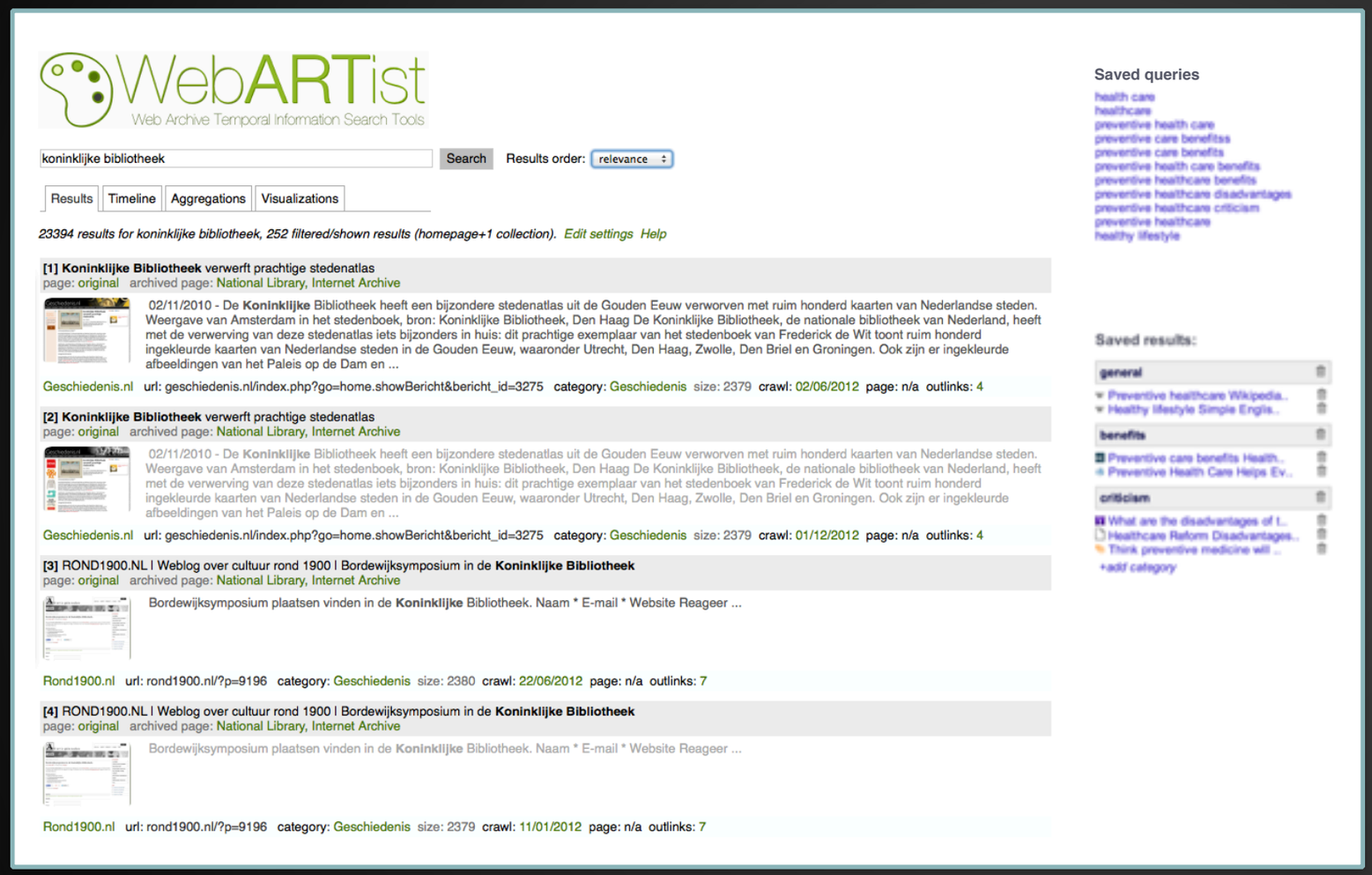
6.3.3 More flexibility: interface

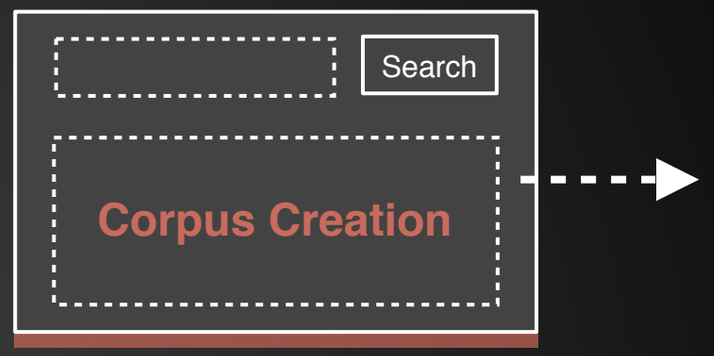
- **Traditional approach:** "Complex" interface integrating all functionality
- **Our Approach:** Divide functionality per (research) stage (Huurdeman & Kamps, 2014)



6.3.3 Example

Supporting research phases: *corpus creation*



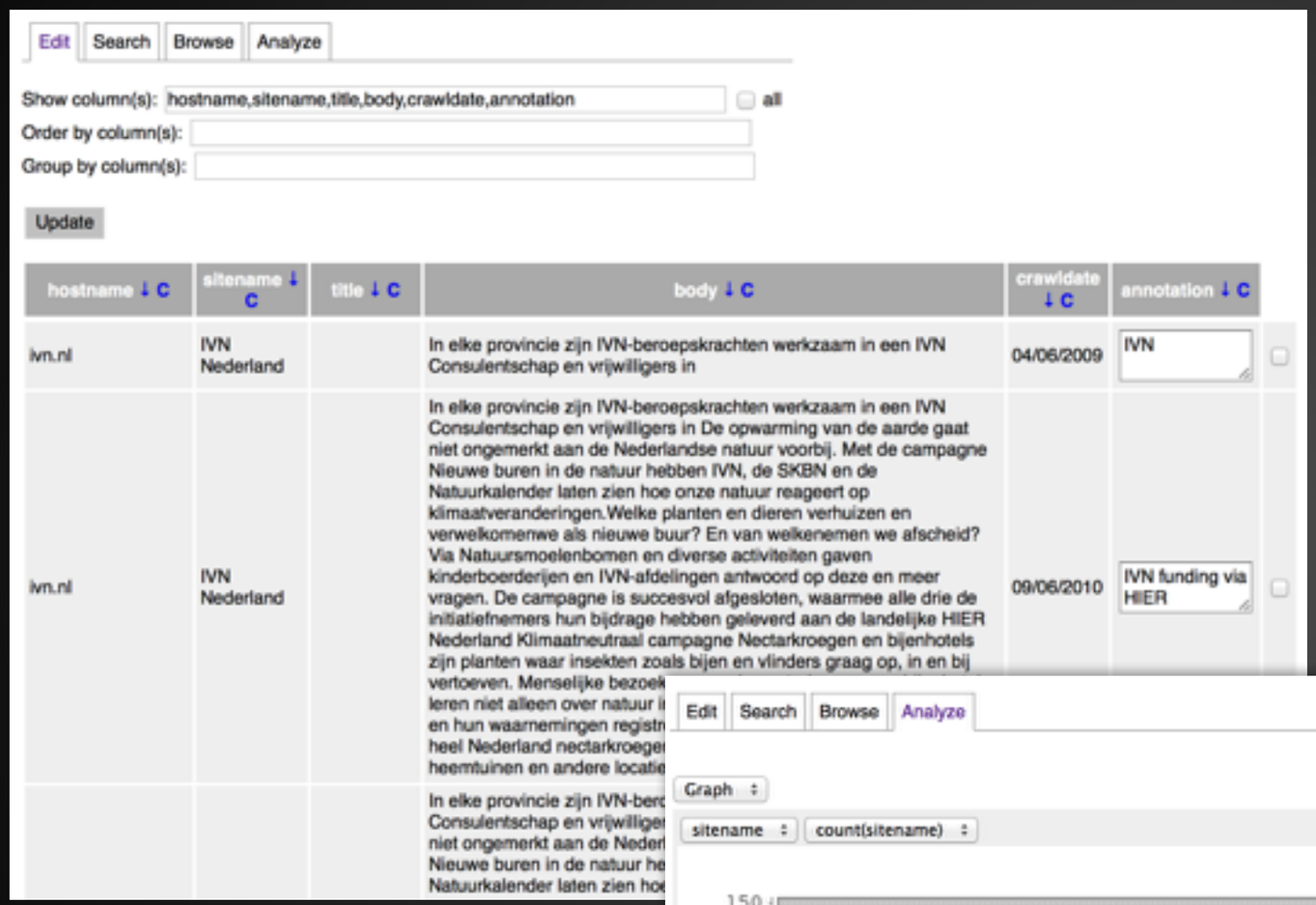


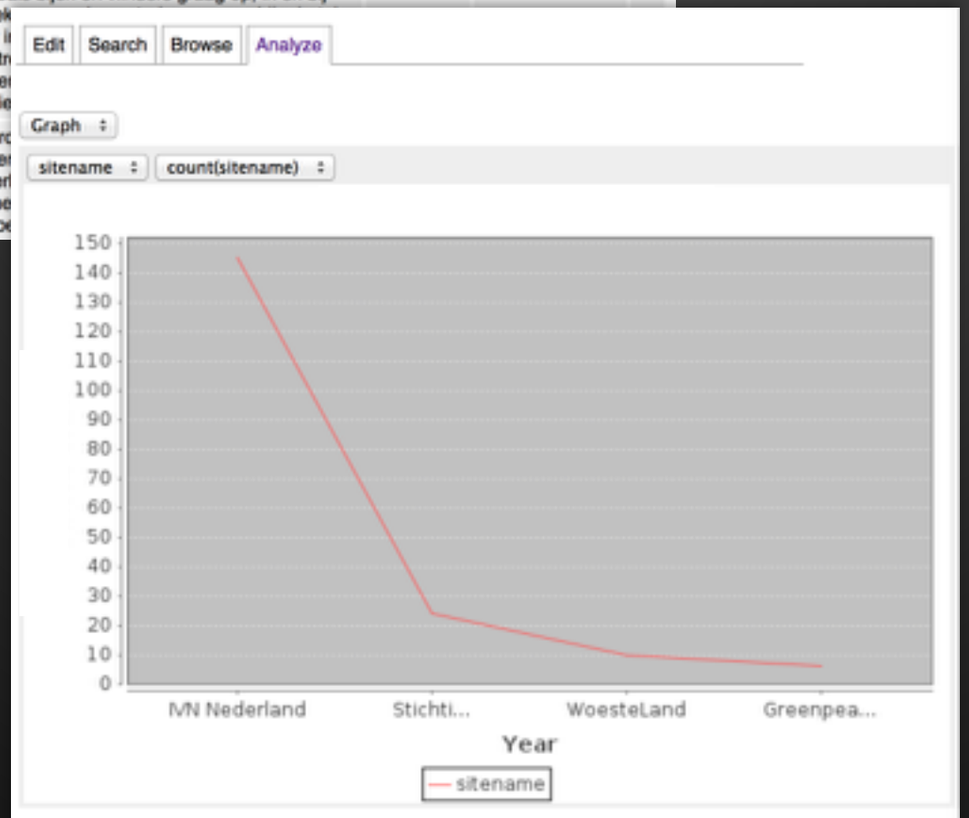
- *faceted* search interface
  - different modalities to explore results

- save complex queries (e.g. `klimaatverandering unesco:17 outlink:postcodeloterij.nl,staatsloterij.nl crawldate:2011 depth:11 ..`)
- save & categorize results

6.3.3 Example

Supporting research phases: *analysis*

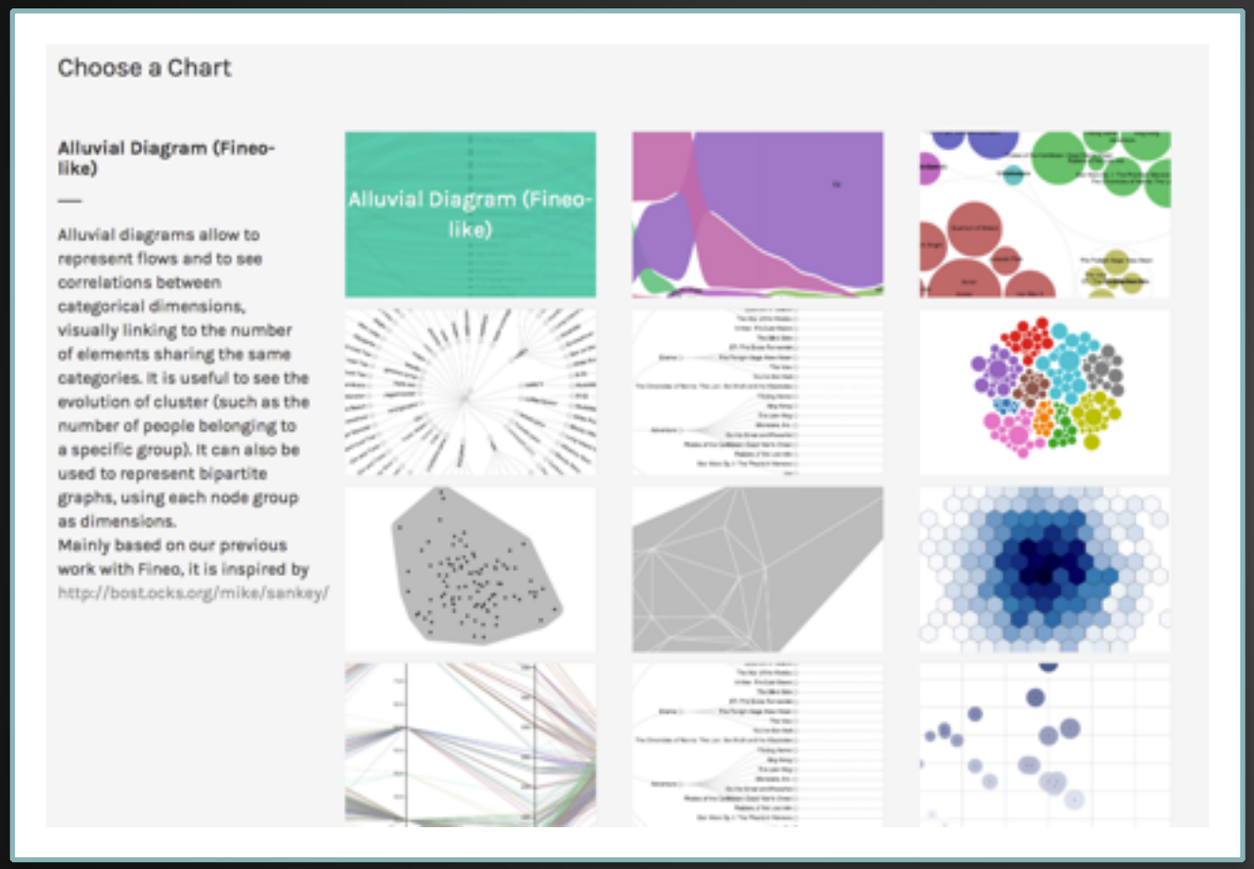






- Analysis interface
  - edit/annotate dataset
  - search & browse dataset
  - analyze

6.3.3 *Example*
Supporting research phases: *dissemination*





- Visualization interface
  - based on **RAW** (*raw.densitydesign.org*)
  - visualize datasets (graphs and visualizations)

From *search* to *research* tools

Wayback Machine → Search engine → “Research” engine

supporting full research task


Web
data
analysis
content
site
study
results
research
bloggers
sites
social
information
web
pages
events
political
public
article
paper
blogs
use
Internet
London
communication
provide
timelines
changes
feeds
media
sharing
evidence
potential
Twitter
Web-based
UIG
validity
links
link
terms
party
seems
attacks
published
development
technologies
persistence
measures
memorializing
networking
Palestinian
election
various
approach
context
may
future
people
linking
longitudinal
MySpace
associated
positive
case
importance
sentiment
related
set
personal
engines
using
indicate
coverage
RSS
whether
produced
issues
gender
users
also
two
discussions
approximately
male
WAB
major
relationship
important
insights
three
news
characteristics
negative
debates
number
words
blog
July
digital
issue
examine
studies
methods
accessibility
network
identify
increase
show
large
U.S
found
availability
however
differences
members
method
examines
change
online
broad
word
science
well
particular
popular
blogosphere
search
postings
practices
years
significant
term
source
time
new
existing
resources
swearing
probably
first
used
national
blogging
contemporary
objects
researchers
collection
borders
sources

Conclusion: Beyond the 'Eternal Now'

- Preserving and accessing our past in Web archives
  - definitions, actors, data issues & use
- The WebART project
  - Workshops & search tools
  - From 'Search' to 'Research' engines

E-DATA & RESEARCH

Jaargang 10 | nummer 2

Nieuwsbrief over data en onderzoek in de alfa- en gamma-wetenschappen.

E-data & Research verschijnt drie keer per jaar en wordt mogelijk gemaakt door: CentERdata, CLARIAH, DANS, Huygens ING, de Koninklijke Bibliotheek en het RIVM.

INHOUD

2
Verslagen van events in Gehoord en bijgewoond

3
Landelijk Coördinatiepunt gaat voor samenhang

3
De Open Universiteit vertelt over RDM-aanpak

4
Mary Vardigan trots op

Hoe maak je gearchiveerde websites bruikbaar voor de wetenschap?

Nationale webarchief onderzocht door WebART

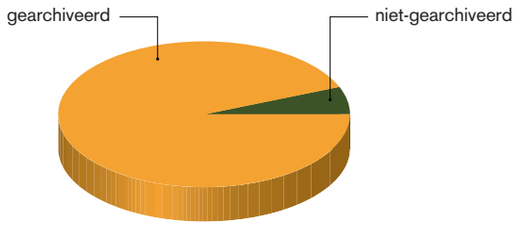
Het eerste grote onderzoeksproject in Nederland naar gebruik van gearchiveerde Nederlandse websites als primaire bron voor onderzoek sluit binnenkort de boeken. WebART-promovendus Hugo Huurdeman blikt terug.

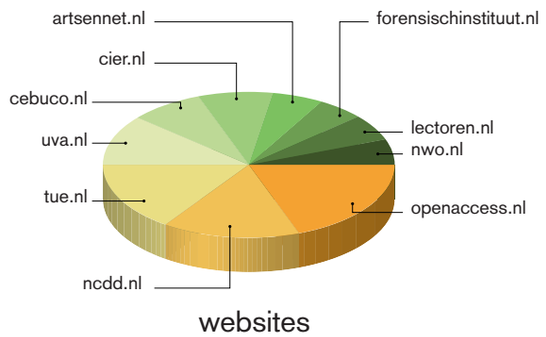
Steven Claeyssens

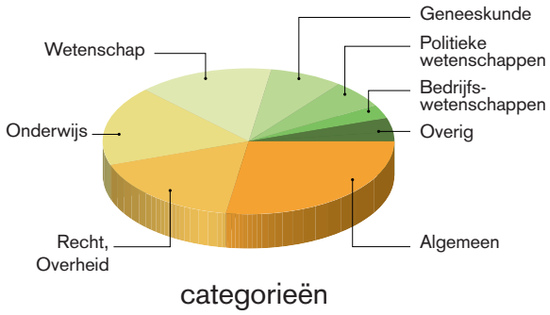
Dit jaar ronden de laatste CATCH-projecten (Continuous Access to Cultural Heritage) hun werkzaamheden af en dus zet ook WebART (Web Archive Retrieval Tools) er een punt achter. WebART was een samenwerking tussen de Universiteit van Amsterdam (UvA), het Centrum Wiskunde en Informatica (CWI) en de Koninklijke Bibliotheek (KB). Het WebART-team lichtte als eerste het Nederlandse nationale webarchief grondig door. Ze gingen daarbij na hoe zo'n heterogeen en omvangrijk *born-digital* archief voor onderzoeksdoeleinden bruikbaar kan zijn en bruikbaarder kan worden gemaakt.

10.000 websites

De KB archiveert sinds 2007 een immer groeiende selectie van Nederlandse websites. Op 1 januari







Op basis van de zoekterr
toont WebARTist verschi
waaronder deze grafieke
grafiek laat de verhouding tussen de gearchi-

gekeken naar de onderzoeksvragen die wetenschappers aan webarchieven zouden willen stellen. Via een intensieve samenwerking met nieuwe media-onderzoekers hebben we vervolgens zoek- en onderzoekstools ontwikkeld die complexe onderzoekstaken kunnen ondersteunen. Denk bijvoorbeeld aan de initiële exploratie van het archief, het definiëren van een dataset en de analyse daarvan. Hiervoor was onderzoek nodig naar schaalbare extractie- en analysemethoden en naar bruikbare interfaces voor verschillende zoekstadia." Zo bouwde het team onder meer WebARTist, een interface waarmee onderzoekers op verschillende manieren het webarchief kunnen verkennen en bevragen.

Ongearchiveerde websites

"Doordat webarchieven van nature incompleet zijn, vroegen wetenschappers ook om contextualisatie over wat er wel en niet in het archief zit. Dit heeft geleid tot verder onderzoek waarin we niet-gearchiveerde webinhoud hebben blootgelegd en gereconstrueerd." Zo slaagden Huurdeman en zijn mede-onderzoekers erin een fors aantal niet-gearchiveerde sites te identificeren

Read more at edata.nl

ze dit niet-gearchiveerde deel van het web tot op

Documentary: Digital Amnesia



- VPRO Tegenlicht (Backlight)
- Dutch version
  - http://tegenlicht.vpro.nl/afleveringen/2014-2015/digitaal-geheugenverlies.html
- English version (Sadly only available *outside* NL)
  - https://www.youtube.com/watch?v=NdZxI3nFVJs

WebART
Web Archive Retrieval Tools

WebART: Web Archive Retrieval Tools

The WebART project aims to critically assess the value of Web archives for realistic research scenarios, and develop information access tools and methods that maximize the archive's utility for research. Our approach is to conduct actual Web archive research hand-in-hand with the development of Web archive access tools tailored to the realistic research scenarios. Within the project, our focus is on the use-case of a humanities researcher, although tools will no doubt be useful for other use-cases as well.

Based in the Netherlands, WebART is an interdisciplinary research project funded by NWO-CATCH. Its team members from the University of Amsterdam and the Centrum for Wiskunde and Informatica combine expertise in Computer Science, Information Science and New Media research. Our cultural heritage partner is the National Library of the Netherlands.



Past events

CATCH Meeting - Archiving the Web: How to Support Research of Future Heritage?
April 19, 2013, National Library of the Netherlands

The WebART project hosted the CATCH meeting at the National Library of the Netherlands, with keynotes by Helen Hockx-Yu, Bill LeFurgy and Bernhard Rieder. Read more.

webarchiving.nl

@webart12

WebART
Web Archive Retrieval Tools

WebART: Web Archive Retrieval Tools

The WebART project aims to critically assess the value of Web archives for realistic research scenarios, and develop information access tools and methods that maximize the archive's utility for research. Our approach is to conduct actual Web archive research hand-in-hand with the development of Web archive access tools tailored to the realistic research scenarios. Within the project, our focus is on the use-case of a humanities researcher, although tools will no doubt be useful for other use-cases as well.

Based in the Netherlands, WebART is an interdisciplinary research project funded by NWO-CATCH. Its team members from the University of Amsterdam and the Centrum for Wiskunde and Informatica combine expertise in Computer Science, Information Science and New Media research. Our cultural heritage partner is the National Library of the Netherlands.



Past events

CATCH Meeting - Archiving the Web: How to Support Research of Future Heritage?
April 19, 2013, National Library of the Netherlands

The WebART project hosted the CATCH meeting at the National Library of the Netherlands, with keynotes by Helen Hockx-Yu, Bill LeFurgy and Bernhard Rieder. Read more.

Sources (1)

- Ball (2010), Webarchiving. State of the art report, DCC Curation Centre.
- Ben-David A. & Huurdeman H. (2014). Web Archive Search as Research: Methodological and Theoretical Implications. Alexandria Journal, Volume 25, No. 1 (2014)
- Brügger N. Website history and the website as an object of study. New Media & Society, 11 (115), Mar. 2009.
- Brügger N. Web Archiving – Between Past, Present, and Future. The Handbook of Internet Studies, 2011.
- Dougherty M. and Meyer E. (2014). Community, tools, and practices in web archiving: The state-of- the-art in relation to social science and humanities research needs. Journal of the Association for Information Science and Technology, 65(11):2195–2209.

Sources (2)

- Huurdeman H., Ben-David A., Samar T. (2013). Sprint Methods for Web Archive Research. Proc. Web Science 2013.
- Huurdeman H., Ben-David A., Kamps J., Samar T., de Vries A., Rogers R. (2015). Lost but Not Forgotten - Finding Pages in the Unarchived Web. Int. Journal on Digital Libraries. Volume 16, pages 247–265, 2015.
- Huurdeman H., Kamps J., From Multistage Information Seeking Models to Multistage Search Systems. Proceedings Fifth Information Interaction in Context Symposium (IIiX 2014).
- International Internet Preservation Consortium (IIPC), http://www.netpreserve.org
- Kamps J. (2011). Toward a model of interaction for complex search tasks. Proceedings ESAIR 2011
- Rogers R. (2013). Digital Methods. MIT Press 2013
- de Vries A., Alink W., Cornacchia R. (2010). Search by Strategy. Proc. ESAIR '10

Beyond the ‘Eternal Now’

preserving & accessing our past via Web archives

Hugo Huurdeman huurdeman@uva.nl

Digital Heritage - 8 March 2016

