Koninklijke Bibliotheek
Nationale bibliotheek van Nederland

The harvest of the Dutch digital fields: the landscape of webarchiving in The Netherlands

Dr. Kees Teszelszky
@keesone

www.kb.nl

Goal:

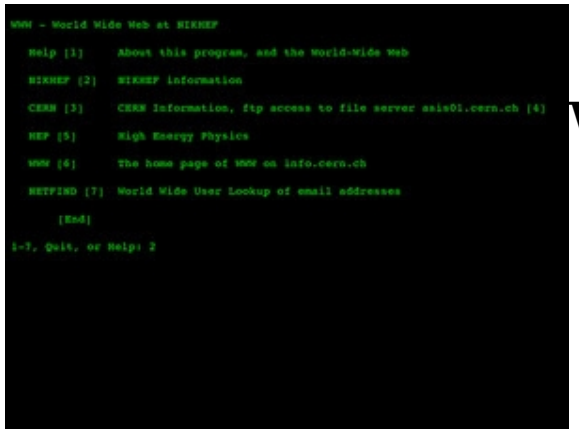
How does the landscape of the web and webarchiving in The Netherlands look like and what is the role of the Koninklijke Bibliotheek – National Library of The Netherlands in webarchiving?

What are the obstacles of webarchiving in The Netherlands and how does the national library deal with this, how can we improve our work?

What can we learn from the Dutch landscape as web archivists and researchers of the web?

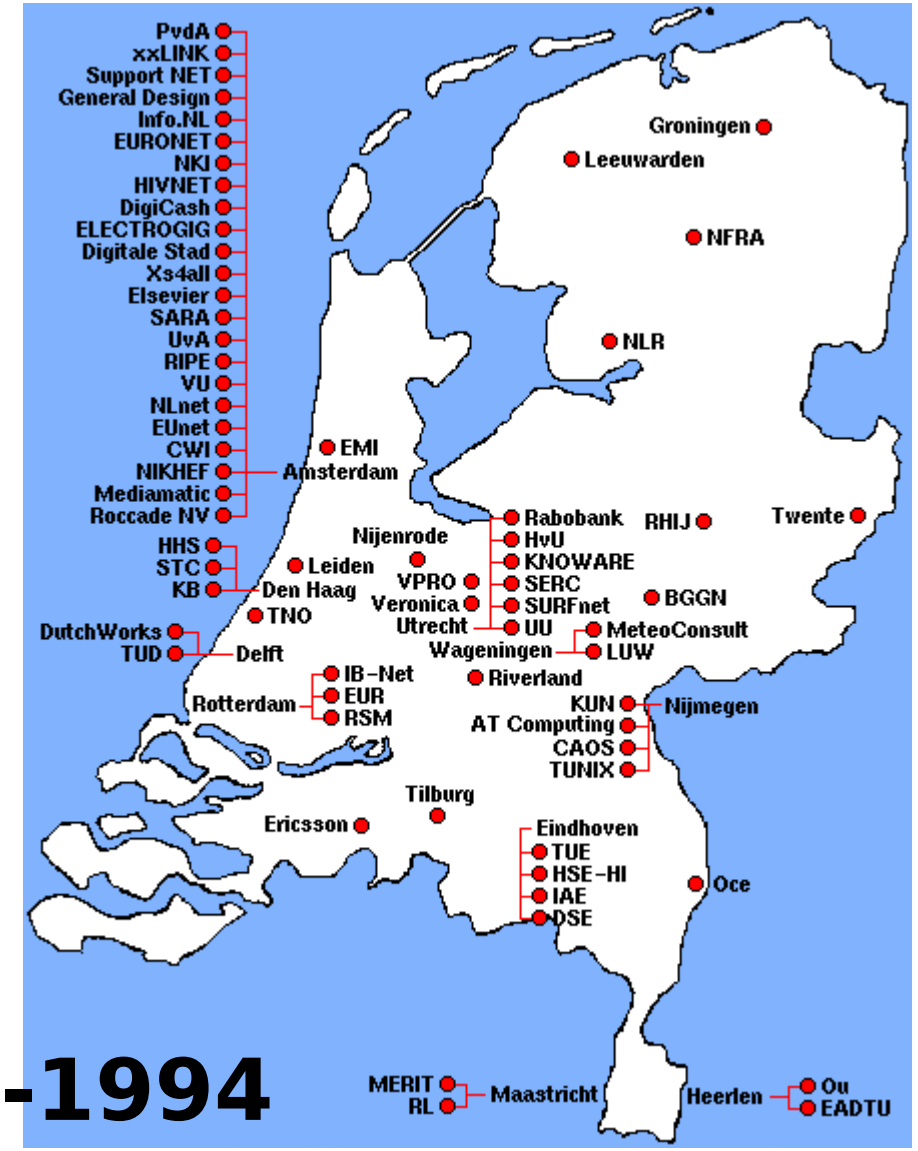
The Dutch national web domain (1992-2017)

.nl country code Top Level Domain: 1986



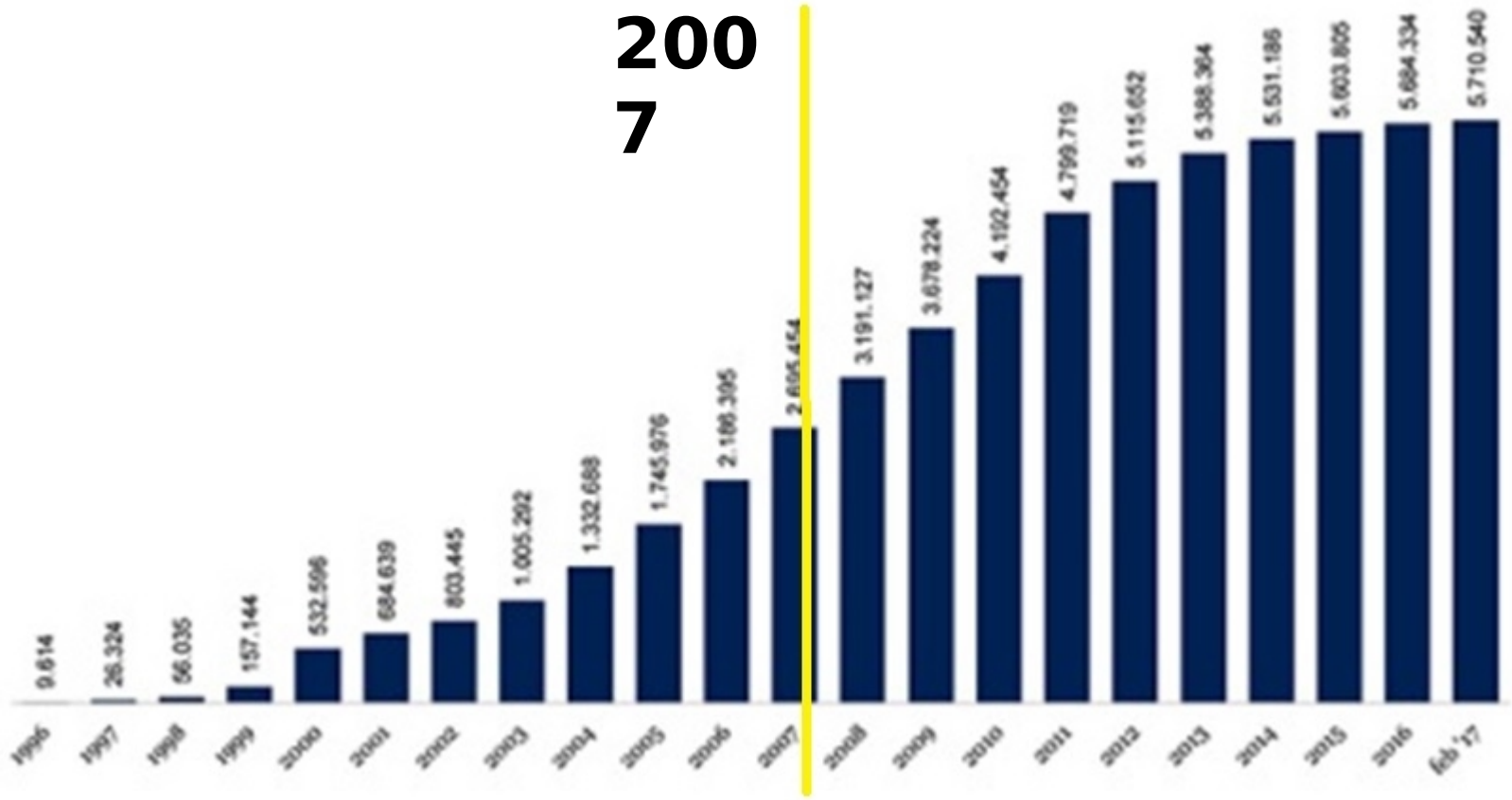
Website of Nikhef, 1992, 3th website in the world

General characteristics:
1. early and innovative, fast-growing
2. local or regional, neither centralised, nor one center
3. less attention to heritage (similar to other institutions…)



Dutch web, mid-1994

.NL-domain names: 5.777.777 million



Dutch national domain: +/- 8 million sites

- .NL Web Archive
  - 13,000 sites = 0.16 %

Selectie


The online world
NOMINET
Did you know?
.ca
.us
.uk
.nl
.de
.be
.fr
.ch
.es
.pt
.pl
.ru
.cn
.in
.jp
.kr
.tk
.au
.br
.co

Web archive of Belgium

The Belgian web is not currently systematically archived

HISTORY

In **June 1988** EARN and Eunet introduce the .be domain. Professor Pierre Verbaeten of the Leuven Catholic University is assigned the responsibility of it.

In **March 1993** the university gets IP connectivity. The name server for .be is installed on the computers of the Computer Science department of the university. The first .be domain names are registered manually in December 1993.
Interest in domain names grows; only names corresponding to the official organisation or company name, a commercial denomination or a trademark are being accepted. Each request has to be justified with official documents.
In **1994**, 129 .be domain names are registered.

TOTAL DOMAIN NAMES

1 5 7 0 8 7 4
.be

7 4 1 2
.brussels

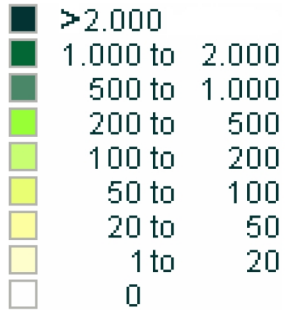
6 6 5 5
.vlaanderen

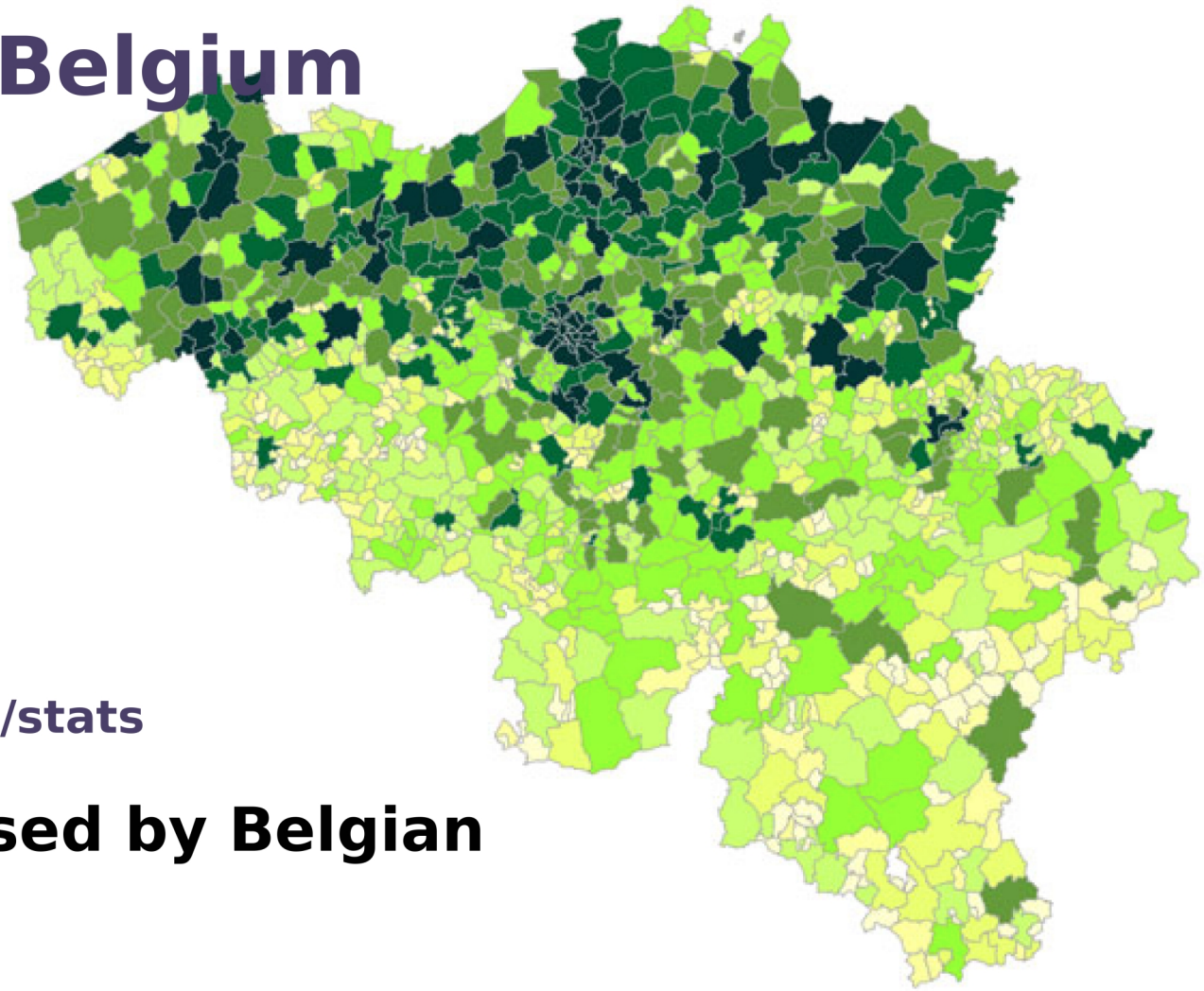
Source: https://www.dnsbelgium.be/en

dnsbelgium

National domain Belgium

Geographic Distribution





Source: https://www.dnsbelgium.be/whois/stats

2.5 % of .nl domain names is used by Belgian citizens: overlap

Languages: Flemish, French, German

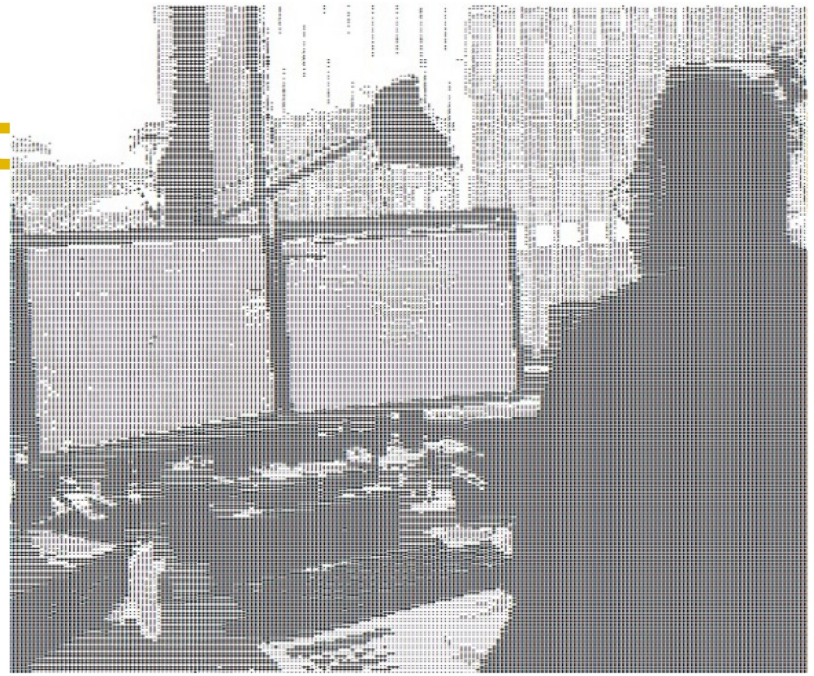
Comparison of web archives

Institution (Figures from early 2016)	Start web archiving	Domain crawl Yes/No	Sites crawled selectively	Size archive in TB	**Size of ccTLD: domain | number of sites**		Persons involved	**FTE**	**Legal deposit**
Koninklijke Bibliotheek (The Netherlands)	**2007**	**no**	**10.000**	**22,5**	**.nl**	**5.623.823**	**2**	**1,3**	**No**
British Library	2005	yes	15.102	27,8	.uk	10.000.000	8	8,0	Yes
Bibliothèque Nationale de France	2004	yes	273.416	567,0	.fr	2.500.000	90	2,5	Yes
Netarchive.dk	2005	yes	50.000	42.7	.dk	1.314.058	20	4,5	Yes
Bibliothèque nationale de Luxembourg	2016	yes	100	14	.lu	90.000	2	1,5	Yes
Belgium: PROMISE project	**2017-8?**	**t.b.c.**	**t.b.c**	**t.b.c**	**.be**	**1.573.331**	**5**	**2.5**	**Yes**

- **Koninklijke Bibliotheek, National Library of The Netherlands**
- **Since 1798, former royal collection, The Hague**
- **Academic library, national library only since 1974**
- **Digital collection since ‘90-s, start web archive in 2007.**

The aim of the KB-NL web collect

To select, preserve and make accessible a representative set of Dutch websites of the Dutch national domain



Web archiving @ KB in numbers

- 13.000 websites in total webcollectio
- 30 Terabyte (one of our biggest digita
- Annual growth: +/- 1.000 sites
- 300 million (hyper)links
- Only selective harvests: no legal dep
- 1,5 full-time equivalent workload
- External hosting
- Restricted access: on site use, Wayback Machine
- One academic research project on data (Web Archive Retrieval Tools, 2011-2016)
- One internal research project on process of selection, harvest and usability (2016-2017)
- No coöperation with National Archive: separate selection and

PROMISE: PReserving Online Multiple Information: towards a Belgian StratEgy

Belgian Science Policy Office
belspo

24 month project financed by Belspo
Start Date: 1 June 2017

Royal Library of Belgium
(Project Coordinator)

Koninklijke Bibliotheek van België
Bibliothèque royale de Belgique

State Archives Belgium

Archives de l'État
Rijksarchief
Staatsarchiv
www.arch.be

Research Group for Media and ICT *and*
Ghent Centre for Digital Humanities

GHENT UNIVERSITY

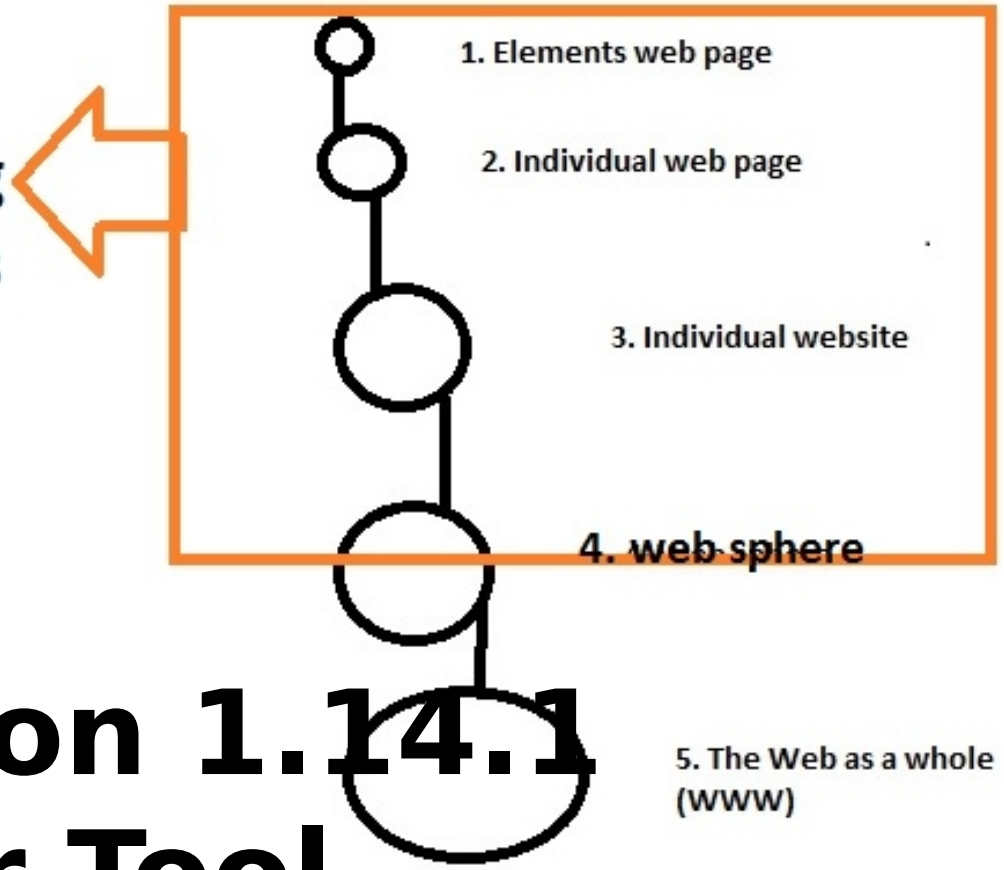
Research Centre on Information, Law & Society

UNIVERSITÉ DE NAMUR

Unité de Recherche et de Formation en Sciences de l'Information et de la Documentation (URF-SID)

HE2B IESSID

Harvesting: what do we preserve?



Web strata after Niels Brügger (2010)

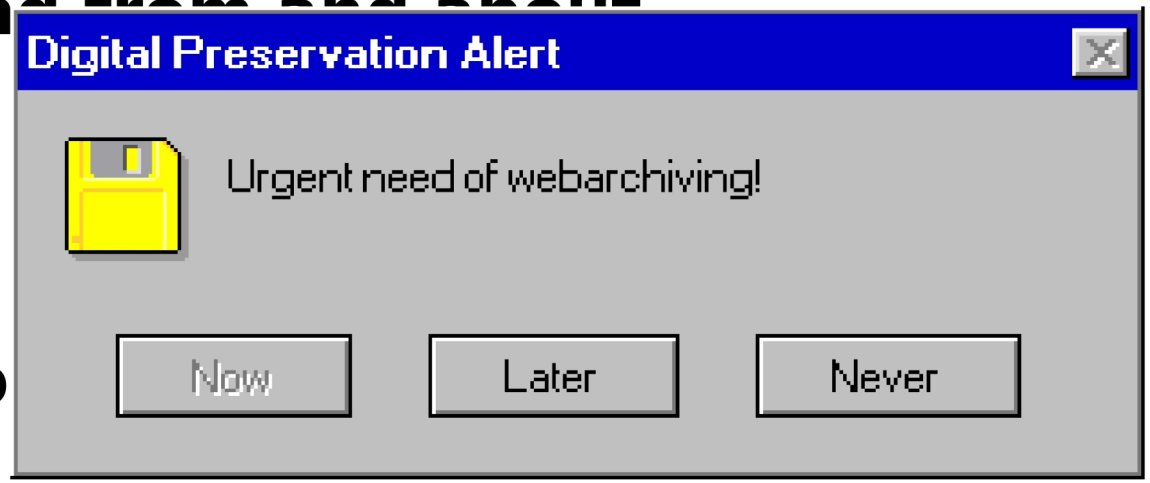
Heritrix version 1.14.1
+ Webcurator Tool
+ opt-out mailer (together hard to upgrade)

Harvesting the Dutch landscape

No legal deposit, therefore no domain crawl of the Dutch national web

Instead: selective harvest

- **By collection specialists (everything from and about The Netherlands)**
- **On request of owners**
- **Endangered digital heritage**
- **Actual sites based on trends and p**
- **Special web collections**
- **National coöperation**
- **International coöperation**



- **Special webcollections**

Some collections:

- Netherlands in WW I
- Premier league football
- Dutch Santa Claus
- Plane crash MH17
- 500 years Reformation
- (Former) monastries
- Frisian websites (Frisian language, Frisian territory)

IIPC– collections (legal issues!)

Other Dutch webcollect

Stadsarchief Rotterdam
mensen locaties gebeurtenissen

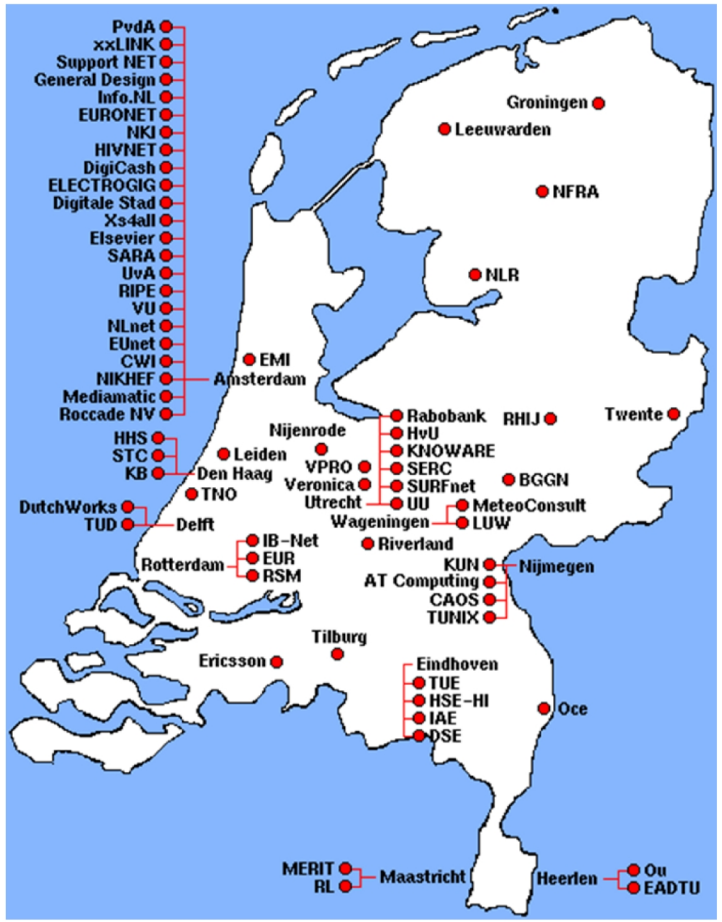
WEBARCHIEF OMROEPEN BEELD EN GELUID

na nationaal archief

archi POL PROJECT ARCHIEF DNPP LINKS ZOEKEN

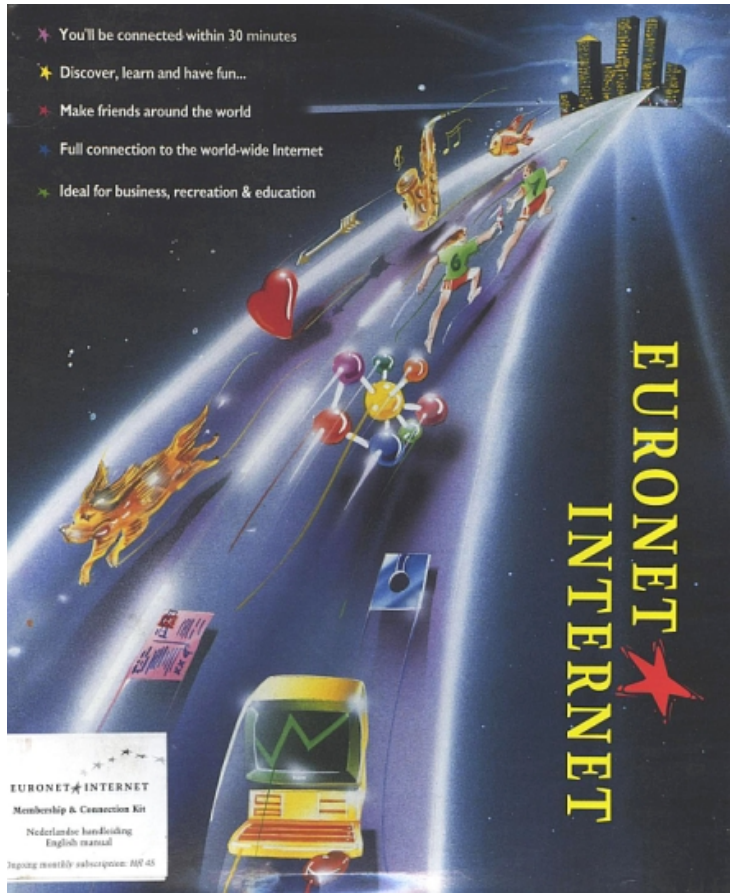
Thematic or regional collections, no national collection (only Frisian)
In total 3,000 archived websites (KB: 13,000, 16,000 country wide)
Few resources per collection
Different crawl strategies, techniques

Dutch web (1994) – Dutch webarchives (2017)



How to improve? Special webcollection Dutch webarchaeology: find the pearls before 2007, esp. 90's

Casus: Euronet provider user sites



Total number of hits: 50198117
Total number of megabytes: 410168

Last three months: 1 |2 |3

Unique find: data and statistics from 1997, 1998 and 2005. (Almost like a domain crawl of euronet.nl)

User	Requests	Mbytes	
jharmsma	14542819	49540	index
darkside	455164	36812	File Not found
dds	1483859	34705	Europian Phil Ochs pages
slf	2131068	22680	EURONET SEX PART
m e	669746	20160	http://www.euronet.
sjoerdb	10559	11282	Beginners MP3 Site,
siri2	763036	8812	
mvda	377015	8606	Marco's Armageddor
absolute	382761	7644	??
jbc	2160569	7242	Dutchbear is moving
elaine	3927363	6120	??
janenman	1123592	5833	Dutchbear is moving
elmar	554041	5798	MagnetTheatre Jumps
w solarz	1026732	5554	freeware shareware
ybpalace	171625	4836	??
rollo	1223016	4659	Amsterdam GAY! DANC
ilm	489476	4411	Male Nude Celebs! F
ceesa	613127	4408	Cees' Original Male
playsoft	301413	4203	Home Page
thieu	280702	4016	Thieu's testing pag
elfrs mp	1109	3740	??
satan	529831	3706	Female Spanking Mal

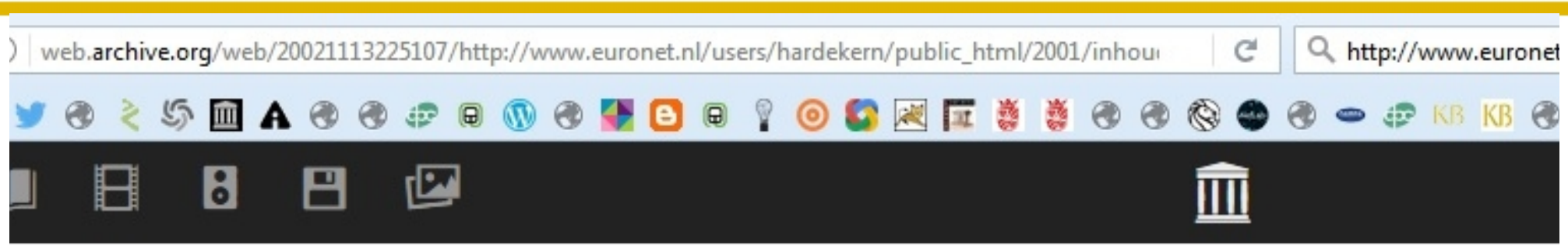
User Home Page stats for 2005-02-02

Total number of hits: 920405
Total number of kilobytes: 12461684

Last five days: 1 (Yesterday) |2 |3 |4 |5

User	Requests	Kbytes	
wilfried	246174	2213570	Wilfried Braakhuis's Website
kgielen	10782	776859	??
pbstg mp	375	554941	??
marij	35090	443390	De Nederlandse Harry Potter site
in002561	753	416314	Arjen's Homepage
nh	112634	387006	ADDRESS WAREHOUSE - ROTTERDAM THE NETHERLANDS
erikje	507	342502	??
citron	7471	287535	LE JOURNAL CITROËN
keesree	23905	258672	TITANIC - A Voyage of Discovery (index)
marbak	25847	247685	MB Music about Faithless, Rollo, Kristine W and more...
warnar	13156	243476	Biosciences
Antonie van Leeuwenhoek Antony van Leeuwenhoek Engravings Images Animalcules Microbiology Statistics Statistiek Methodologie Amsterdam			
cindyx	6758	229325	Torture video's, Extreme SM picture gallery, Bizar. Erotic, Sex, Bondage
e wesker	12236	211849	Evert's Homepage
w solarz	22514	198716	freeware shareware download software windows 98 games and Entertainment
lindboom	2864	156773	??

- Amount of user sites in 1998, 2005, 2006;
- Description of content and data;
- User statistics;
- Exact URL, user name.



Hard to crawl due to bad construction of sites

Legal issues of web archaeology

Problems:

- **No contact address: opt-out**
- **"Digital dementia": owner do**
  want to be associated with pa
- **No legal means to obtain mat**
- **Neither owners, nor provider**
  interested in preserving herit

.

Digital incunables of the Dutch web

Positive points: finding unique and missing parts of other collections

Broadcast sites, political parties, local si

Journalistieke en andere vertrekpunten

Journalistieke adressen

- ASNE American Society of Newspaper Editors
- Awesome Lists Met mooie doorverwijzingen
- Beat Page Lijst met journalistieke links van reporter.org
- Committee to Protect Journalists (CPJ) Laatste nieuws over problemen rond persvrijheid in de wereld
- Eurojournalism Europese adressen voor de journalist
- European Institute for the Media Persberichten, agenda, publicaties
- European Union Newsroom Nieuwsoverzicht, agenda, persberichten
- Finding Data on the Internet A Journalist's Guide
- Hotzones Informatie en tips voor journalisten die naar gevaarlijke gebieden reizen
- IFJ International Federation of Journalists
- IFEX Schendingen van persvrijheid
- Index Persverwijzingen van collega Ren de Vree van het AD ***Aanbevolen!***
- Internet Exercises for Journalists Het Net als Uitdaging...
- Internet Newsroom Begin hier de jacht op informatie
- Investigative Reporters
- Journalism College Florida Met doorverwijzingen plus instrumenten om op het Internet informatie te vinden
- Journalism Ethics Een index met theorie en praktijk
- Journalism List Pagina met fraaie doorverwijzingen
- Journalism SourceFinder Met be- en onbekende adressen
- Journalist's Toolbox Veel mooie doorverwijzingen
- Megasources Mega, dus zeer veel links
- National Press Club in Washington Persberichten, adressen
- Newseum Museum over het nieuws
- NewsPlace News, resources
- Nipkow Van de schijven
- Online Journalist Bronnen, links, een Webtour
- People's Communication Charter Handvest van de Amsterdamse hoogleraar Hamelink
- Press News Site Current events, press room, tools
- Pulitzer De Amerikaanse prijs voor journalisten
- Reference Guide Gopher-overzicht met veel naslagwerken
- Reporters Network Met Journalism Resources
- Research on the Net Met behulp van Wired's editor guide
- Scoop Internet Rescource Guide voor m.n. journalisten
- Smoking Gun Site met materiaal over schandalen
- WebOvision Journalism, dus met nuttige links
- World Wide Events Agenda van aanstaande evenementen

Andere vertrekpunten

- WWW Virtual Library From A to Z, hier komt geen eind aan, van Aboriginal Studies tot Zoos
- Yahoo Met heel veel mooie doorverwijzingen en zoekmogelijkheden (zie ook 'Zoekinstrumenten')
- Yanoff's List Zeer volledige List van een veteraan op het Internet ***AANBEVOLEN!***

De visie van | Meer informa | Opinie- | D66 Interact | Wat staat er | Verwijzir

Chatten met Kamerleden!

Een partijcongres is altijd een goede gelegenheid om eens een babbeltje te maken met die mensen die midden in het politieke centrum staan. Heeft u gerichte vragen of wilt u gewoon eens een babbeltje maken, dan bent u zaterdag en zondag van harte welkom op het IRC-kanaal van D66. U heeft hiervoor wel een apart programma nodig. Hoe u dit kunt downloaden, installeren en hoe dit verder werkt (het is NIET ingewikkeld!) dat leest u hier. De volgende mensen kunt u bereiken op de volgende tijden:

Zaterdag:
10.30 Bert Bakker
12.30 Roger van Boxtel
14.30 Nicky van 't Riet
16.30 Pieter ter Veer

Verkiezingscongres via Internet

Zaterdag 14 en Zondag 15 Februari zal in Noordwijk het D66 congres plaatsvinden. Ook dit congres zal weer via internet te volgen zijn. Met behulp van teksten, audio en foto's zal iedereen in staat zijn om het congres zo goed mogelijk te volgen. In Emmen zette D66 de trend om congressen live te verslaan en deze keer belooft het nog weer beter te worden. Maar wie het congres echt optimaal wil meemaken kan hier een omschrijving van het congres vinden en hier een routebeschrijving...

D66 vindt sport

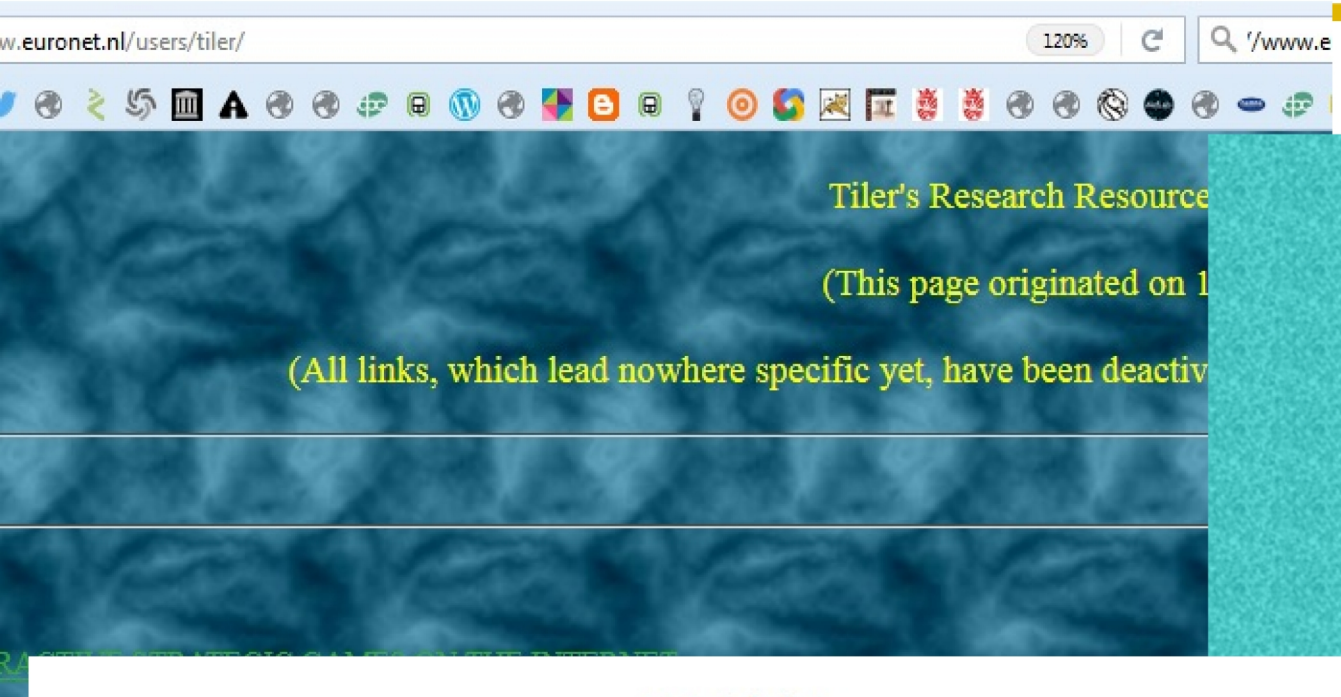
Virtuele wandelgang

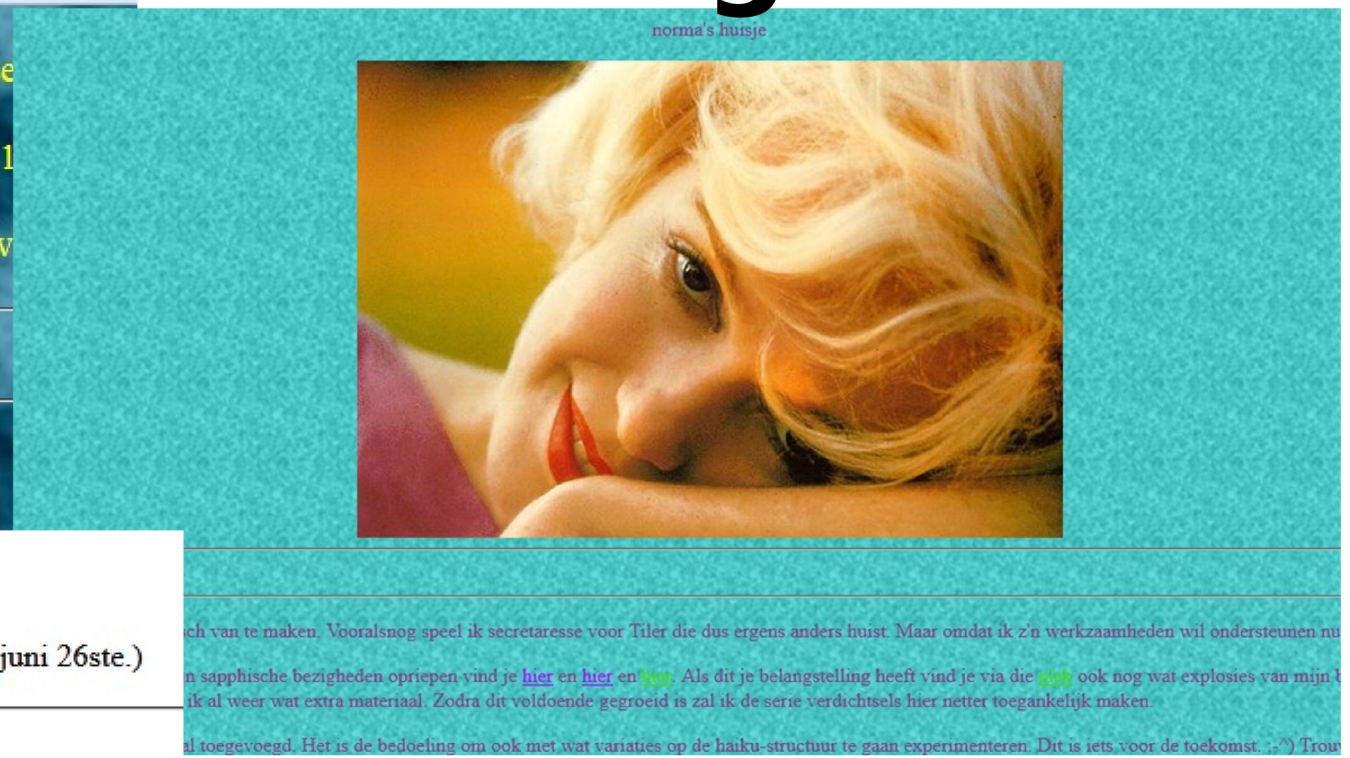
Op het komende D66-congres lopen uiteraard ook alle D66-prominenten rond. De beste plaats dus om even een kamerlid of een minister aan te schieten. Wie niet naar het congres in Noordwijk kan komen kan toch van de gelegenheid gebruik maken. Mail u vraag naar webmaster@d66.nl en vermelt aan wie u de vraag wilt stellen. Het D66 Nieuwe Media Team loopt voordurend door de wandelgangen, dus als de gelegheid zich voordoet zullen wij u vraag aan de desbetreffende persoon voorleggen. Alle antwoorden op uw vragen vindt u dan dit weekend terug op de site - in Realaudio of Realvideo.

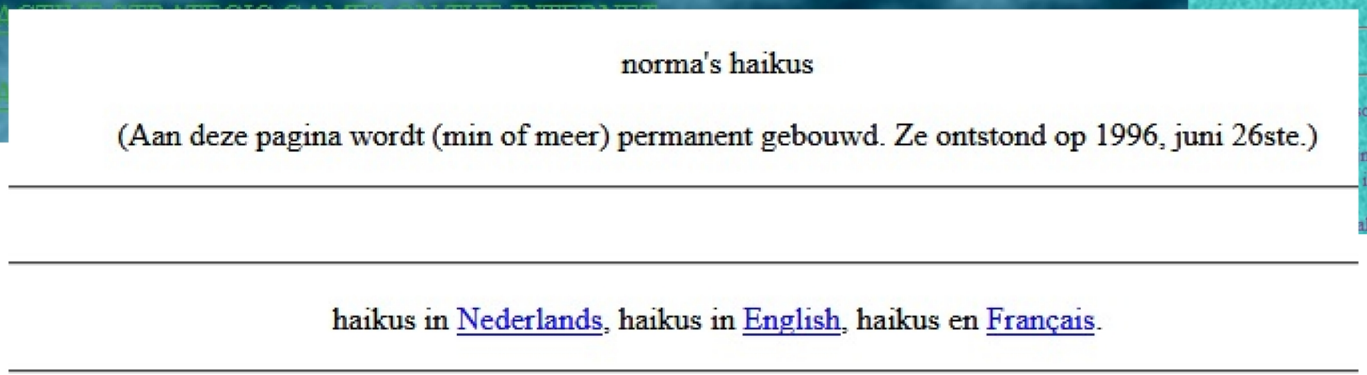
Stuur email aan Els Borst

Koninklijke Bibliotheek
Nationale bibliotheek van Nederland

Born digital mate



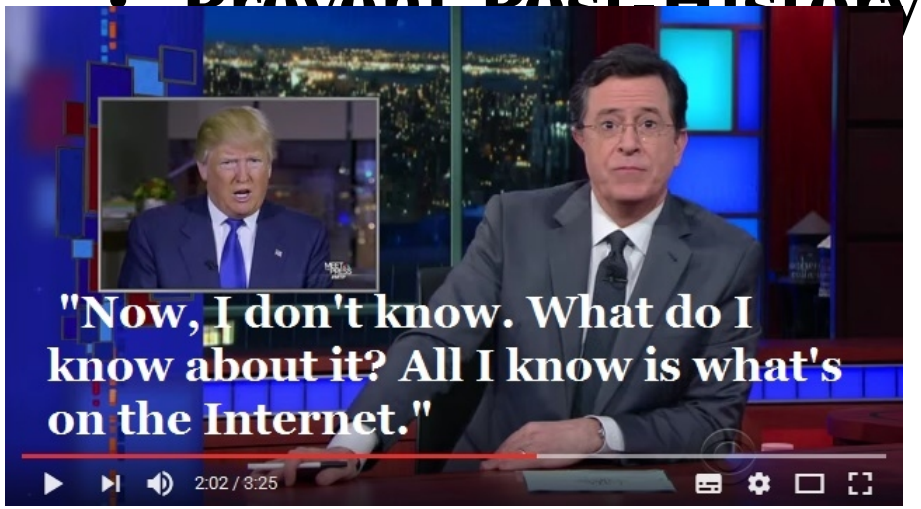




First research sites with born digital scientific publications (and hidden lite
Web archaeology: layers.

Post-truth web collection

- **Context of “truth” (internal / external link structure), fakenews**
- **Historic sources as building blocks for academic**
- **Archived website or at least data (web sphere!)**
- **Coöperation with academics:**
- **Prevent Post-History period**



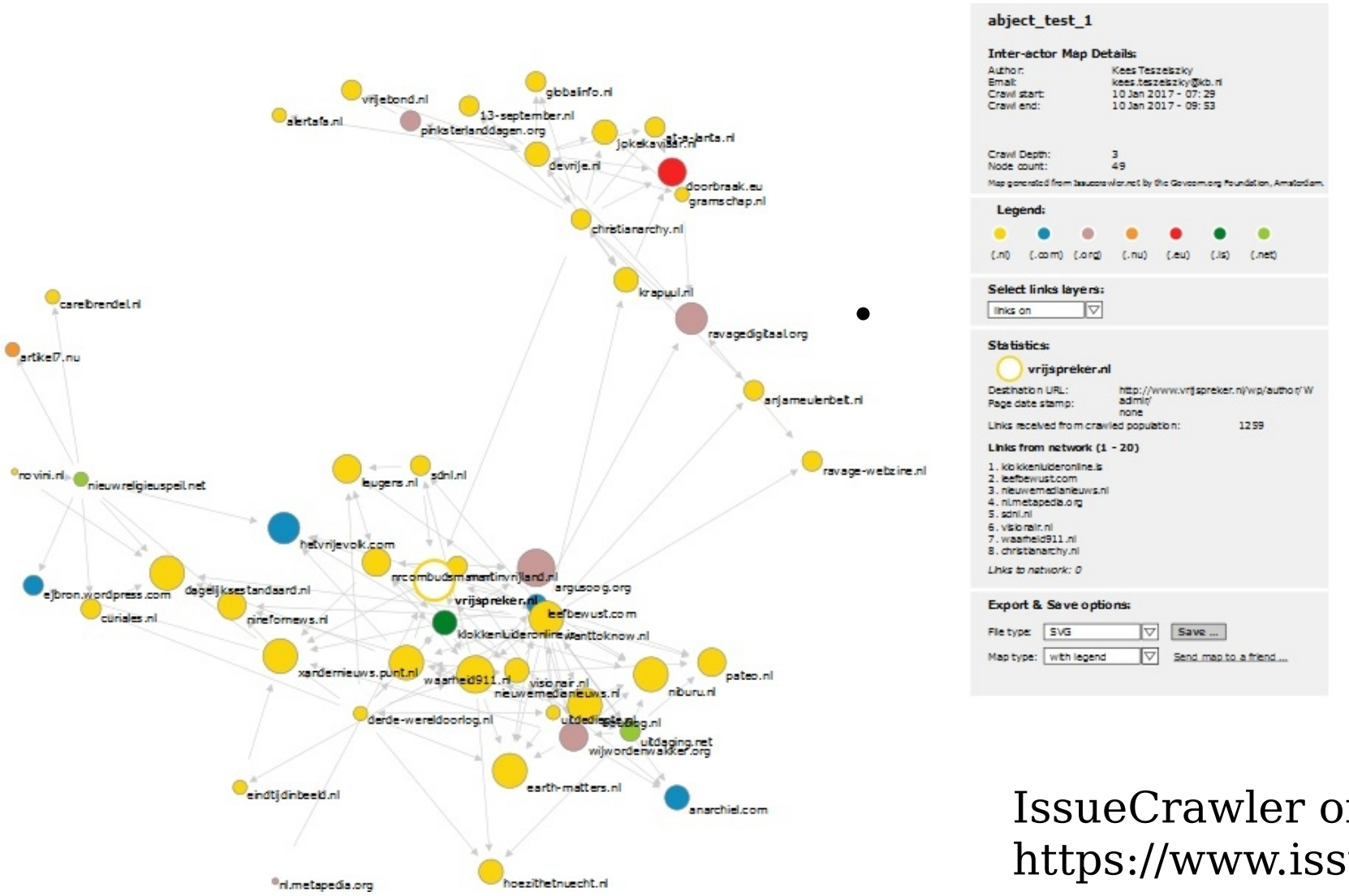

LEIDS UNIVERSITAIR WEEKBLAD

Mare

Lees online
Download Mare

071-527... Internetarcheoloog

Kees Teszelszky doet onderzoek naar webarchivering bij de Koninklijke Bibliotheek. Samen met Leidse onderzoekers spreekt



abject_test_1

Inter-actor Map Details:

Author:	Kees Teszelszky
Email:	kees.teszelszky@kb.nl
Crawl start:	10 Jan 2017 - 07:29
Crawl end:	10 Jan 2017 - 09:53
Crawl Depth:	3
Node count:	49

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

Legend:

(.nl) (.com) (.org) (.nu) (.eu) (.is) (.net)

Select links layers:

links on

Statistics:

vrijspreker.nl

Destination URL:	http://www.vrijspreker.nl/wp/author/'W admin/'
Page date stamp:	none

Links received from crawled population: 1259

Links from network (1 - 20)

1. klokkenluideronline.is
2. leefbewust.com
3. nieuwemedianieuws.nl
4. nl.metapedia.org
5. sdnl.nl
6. visionair.nl
7. waarheid911.nl
8. christianarchy.nl

Links to network: 0

Export & Save options:

File type: SVG Save ...

Map type: with legend Send map to a friend ...

Link analysis as important as webarchiving

IssueCrawler of Digital Methods Initiative
https://www.issuecrawler.net/

maatschappijkritisch_4_rechts

Co-link Map Details:

Author:	Kees Teszelszky
Email:	kees.teszelszky@kb.nl
Crawl start:	21 Sep 2017 - 06:46
Crawl end:	21 Sep 2017 - 09:17
Privilege starting points:	off
Co-link Analysis Mode:	page
Iterations:	3
Crawl Depth:	3
Node count:	43

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

Legend:

(.com) (.nu) (.nl) (.org) (.net) (.be) (.eu)

Select links layers:

links on

Statistics:

jihadwatch.org

Destination URL:	http://www.jihadwatch.org/
Page date stamp:	none
Links received from crawled population:	4133

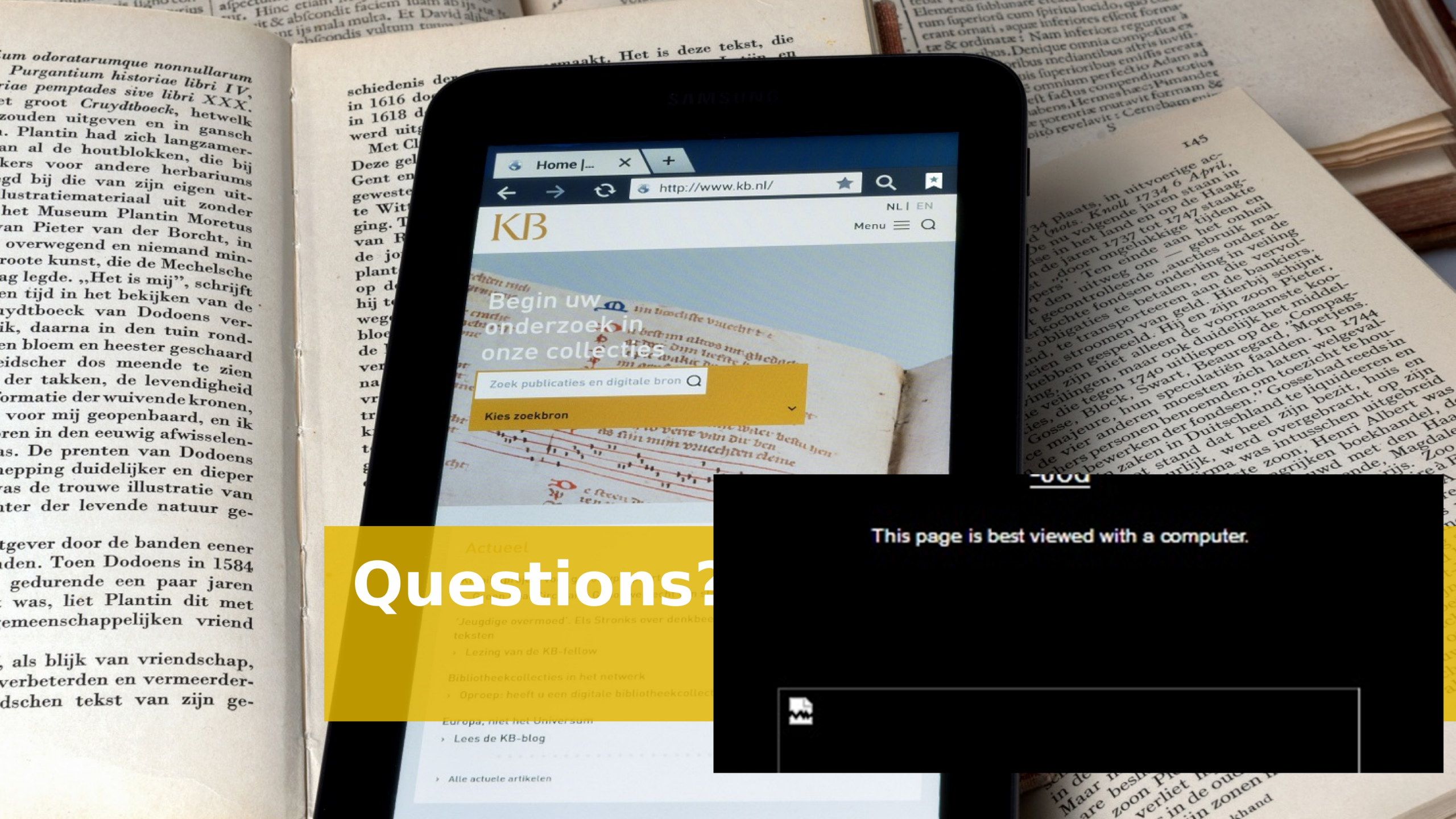
Links from network (1 - 20)

1. dry bonesblog.blogspot.com
2. elderoftziyon.blogspot.com
3. fjordman.blogspot.com
4. gatesofvienna.blogspot.com
5. islamofobie.nl
6. nieuwreligieuspeil.net
7. thereligionofpeace.com

Links to network: 0

Conclusion

- Webarchiving differs in each country due to local culture and legal circumstances (similar to libraries and archives): it is important to take this phenomen in account when web archiving and doing research
- The Netherlands: locally organised, all web archives are in fact special collections, no central collection, therefore national coöperation is needed
- All relatively small collections, but with much local expertise and devotion.
- Selection policy have to be reviewed every 5 years: but also in retrospective: permanent web archaeology
- Special collections are good to unite local efforts nationally and to focus selective crawls for past, present and future webarchiving.


Home |...
http://www.kb.nl/
KB
NL | EN
Menu
Begin uw onderzoek in onze collecties
Zoek publicaties en digitale bron
Kies zoekbron
Actueel
Lees de KB-blog
Alle actuele artikelen
This page is best viewed with a computer.
Questions?