

# Discovering and Archiving the Frisian Web. Preparing for a National Domain Crawl



Susanne van den Eijkel and Iris Geldermans  
IIPC WAC2023, Hilversum – 12 May 2023

**KB** } nationale  
bibliotheek

# About the KB team

## Team Web Archiving

- IT support en programming (4 FTE)
- Preservation specialists (1 FTE)
- Testers (0.5 FTE)
- Coordination (2 FTE)
- Researcher (0.5 FTE)

Team members: 13

## Team Collections

- Collection specialists & quality assurance (2-3 FTE)
- Curator Digital Collections
- Internet archaeology (0.1 FTE)

Team members: 5

# About the KB Web Collection

- Selective web collection
- Special subcollections (XS4ALL, COVID-19.)
  
- Access for public: reading room,
- Access for researchers: Researcher-in-Residence and Twi-XL
- Access through internships
  
- Web Curator Tool (WCT)
- Heritrix

**KBLAB**

Join us and explore the KB's digital treasure trove

The KB Lab hosts all experimental tools and data sets based on the KB's digitised collection.

**twixl**

an infrastructure for cross-media research

The Twixl logo features the word 'twixl' in a bold, lowercase, sans-serif font. Below it, the tagline 'an infrastructure for cross-media research' is written in a smaller, lowercase font. The bottom portion of the logo is a teal-colored banner containing several blue gears of varying sizes, with a magnifying glass icon integrated into one of the gears.The Web Curator Tool logo consists of a dark red banner. On the left side, there is a stylized globe with white latitude and longitude lines. To the right of the globe, the words 'WEB CURATOR TOOL' are written in a white, uppercase, serif font.

# Legal issues

- No legal deposit
  - KB wants to preserve everything that was published in or about the Netherlands, including websites!
- Opt-out principle
- Ongoing joint lobby for new legislation in the Netherlands

# National Domain

- Library wishes to crawl the Dutch web (.nl)
- About 6 million Dutch websites, yearly crawled
- About 100 TB per year
  
- Frisian domain crawl was a pilot for national domain



# Fryslân

*A Dutch province with its own national language.*

# Selecting a Domain

## Must Have:

- I. Top level domain
- IV. Culture
- V. Structure (network)

## Must have in future:

- VI. History (time)

## Nice to have:

II. Geography

III. Language

## Must Have:

- I. .frl websites
- IV. Ask experts
- V. Frisian Wikipedia

## Must have in future:

- VI. -

## Nice to have:

II. -

III. -

9.500 domains

800 domains

1 domain

# Scope of the Pilot

## **In scope**

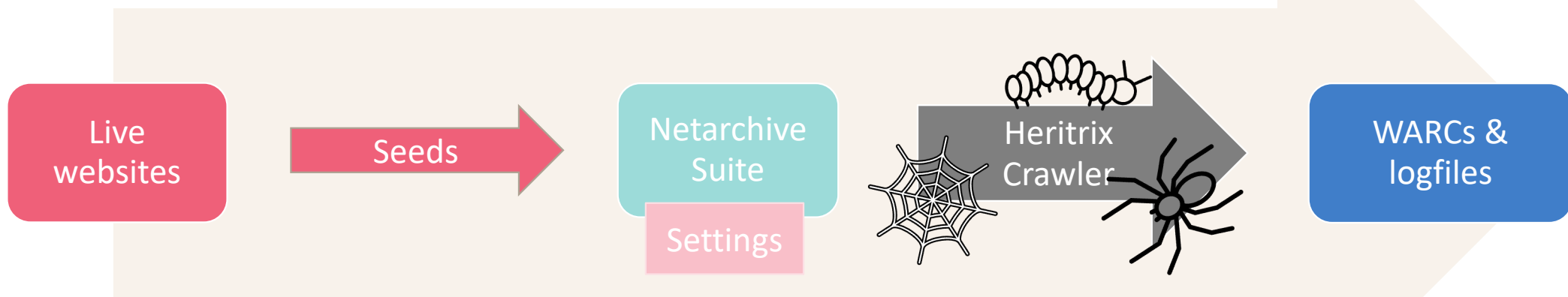
- Tooling (NetarchiveSuite)
- Defining selection criteria
- Profile settings
- Repeat FDC (IT Support)
- Impact analysis


## **Out of scope**

- Long term preservation
- Create a new collection
- Access for broad public



# Domain Crawl: Proces



 Menu

- Definitions
- Harvest status
  - All Jobs**
  - All Running Jobs
  - H3 Remote Access
- Harvest Channels
- Bitpreservation
- Quality Assurance
- Systemstate

Job status

Harvest name

Order  Display  rows per page.

Search results: 5, displaying results 1 to 5.

## Job Status

Job ID	Harvest name	Run number	Start time
5	FDC 2022	1	2022/11/30 10:55:41
4	FDC 2022	1	2022/11/30 10:55:40
3	FDC 2022	1	2022/11/30 10:55:39
2	FDC 2022	1	2022/11/30 10:55:38
1	FDC 2022	1	2022/11/30 10:55:37

# Testing the Settings: Discovery Path in Log Files

- Whole path of crawler 'hops'.
- Different characters for a type of hop.
- Our content strategy was to harvest websites as complete as possible, if they were related to the original seed.
- With local Python scripting we compared the results and tried to filter out irrelevant content.

## Discovery path character legend as found in the log files:

R - Redirect

E - Embedded links necessary to render the page `<img src=...>`

X - Speculative embed (aggressive/Javascript link extraction)

L - Link (normal navigation, like: `<img src=...>` )

P - Prerequisite (as for DNS or robots.txt before another URI)

## Discovery path in log

```
- dns:112fryslan.nl-ams1.upcloudobjects.com RRRLLLEXXP
- dns:112fryslan.nl-ams1.upcloudobjects.com RRRLLLEXXP
- dns:112fryslan.nl-ams1.upcloudobjects.com RRRLLLEXXP
- https://112fryslan.nl-ams1.upcloudobjects.com/uploads
- dns:112fryslan.nl-ams1.upcloudobjects.com RRRLLLEXXP
```

# Testing the Settings: Discovery Path in Log Files

- First test:
  - Five E's and one X permitted (default).
  - Result: too much unwanted content.
- Second test:
  - Two E's and no X permitted
  - Result: relevant PDF-files and stylesheets were no longer harvested.
- Third test:
  - Two E's and no X permitted, unless the URL contained the seed URL.
  - Result: PDF-files that were missing in the second test, were available now but with minimal other unwanted content.

## Settings

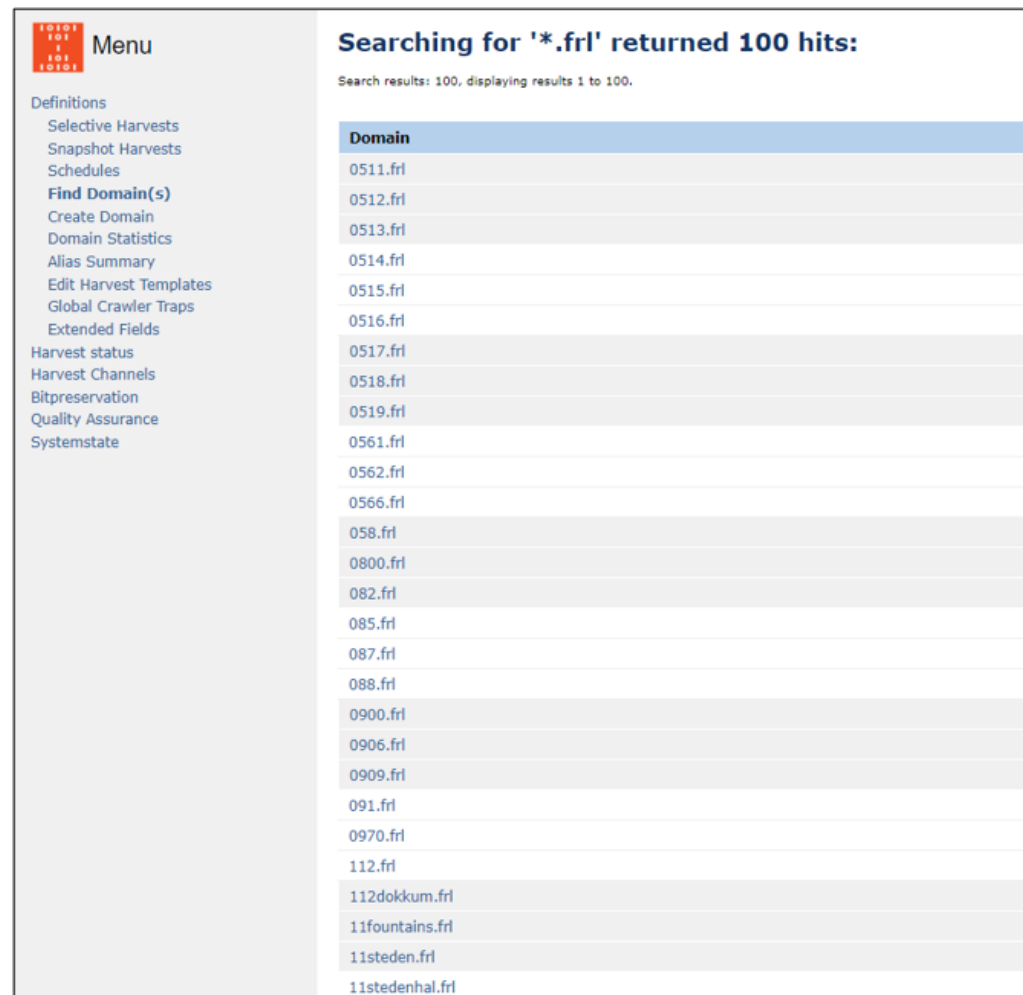
```
## Override properties for  
org.archive.modules.deciderules.TransclusionDecideRule  
  scope.rules[3].maxTransHops=2  
  scope.rules[3].maxSpeculativeHops=0
```

TransHops is default 5 en maxSpeculativeHops 1.

```
<!-- Begin by REJECTing all... -->  
<bean class="org.archive.modules.deciderules.RejectDecideRule" /> </bean>  
<!-- ...then ACCEPT those within configured/seed-implied SURT prefixes...  
>  
<bean class="org.archive.modules.deciderules.surt.SurtPrefixedDecideRule">  
  <property name="seedsAsSurtPrefixes" value="true" />  
  <property name="alsoCheckVia" value="true" />  
  <!-- <property name="surtsSourceFile" value="" /> -->  
  <property name="surtsDumpFile" value="surts.dump" />  
</bean>
```

# Crawling the Frisian Web Domain

1. The .frl domain
2. Expert list
3. Wikipedia
4. The .frl domain with fresh data



The screenshot shows a web interface with a menu on the left and search results on the right. The menu includes options like 'Definitions', 'Find Domain(s)', and 'Harvest status'. The search results section is titled 'Searching for '\*.frl' returned 100 hits:' and displays a list of domain names.

**Menu**

- Definitions
- Selective Harvests
- Snapshot Harvests
- Schedules
- Find Domain(s)**
- Create Domain
- Domain Statistics
- Alias Summary
- Edit Harvest Templates
- Global Crawler Traps
- Extended Fields
- Harvest status
- Harvest Channels
- Bitpreservation
- Quality Assurance
- Systemstate

**Searching for '\*.frl' returned 100 hits:**

Search results: 100, displaying results 1 to 100.

Domain
0511.frl
0512.frl
0513.frl
0514.frl
0515.frl
0516.frl
0517.frl
0518.frl
0519.frl
0561.frl
0562.frl
0566.frl
058.frl
0800.frl
082.frl
085.frl
087.frl
088.frl
0900.frl
0906.frl
0909.frl
091.frl
0970.frl
112.frl
112dokkum.frl
11fountains.frl
11steden.frl
11stedenhal.frl

# Crawling Top Level Domain

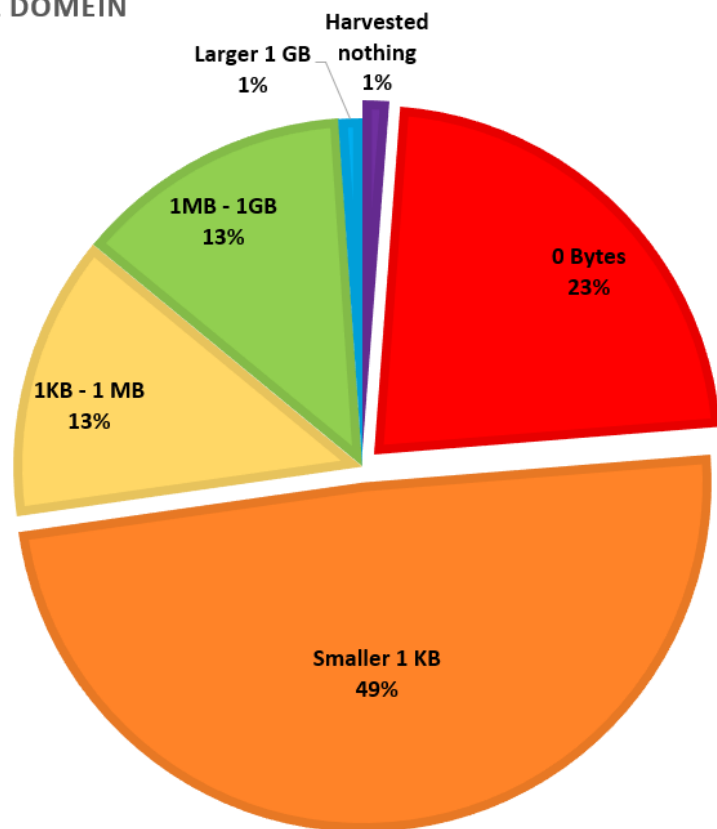
*Harvest results according to NAS metadata*

	Bytes	Documents	Domains
.frl harvest (2022)	296.115.975.995	3.035.059	9592

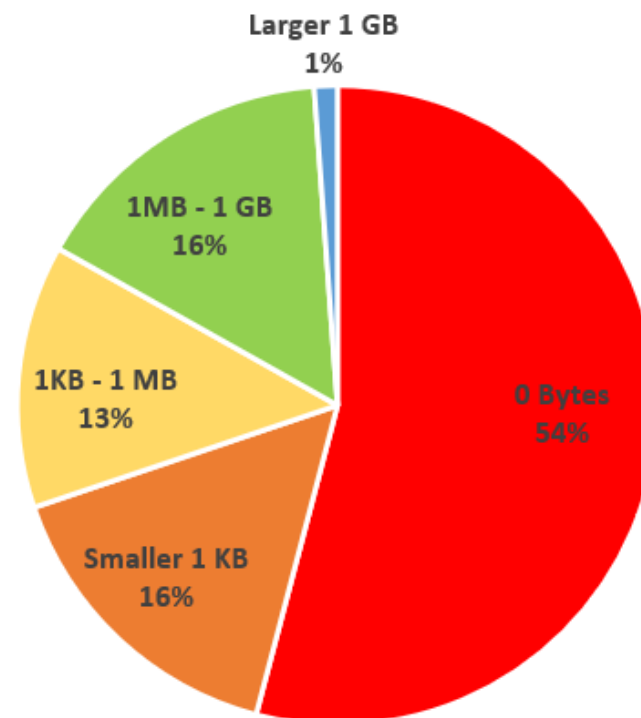
	Start time	End time	Duration
.frl harvest (2022)	14-3-2022 09:48	19-4-2022 10:51	5 weeks 1 day, 1 hour, 3 minutes

# Comparing NAS Metadata with Log File Data

BYTES HARVESTED  
.FRL DOMEIN



Logs - Bytes



# Crawling Top Level Domain

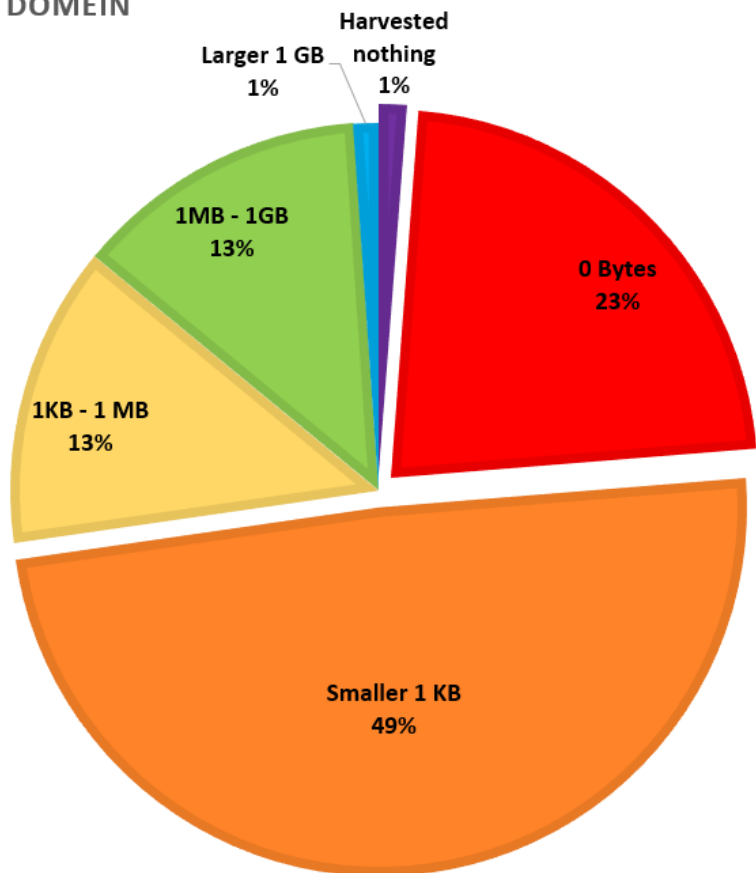
*Harvest results according to NAS metadata*

	Bytes	Documents	Domains
.frl harvest (2022)	296.115.975.995	3.035.059	9592
Expert list	180.520.734.507	2.098.732	828

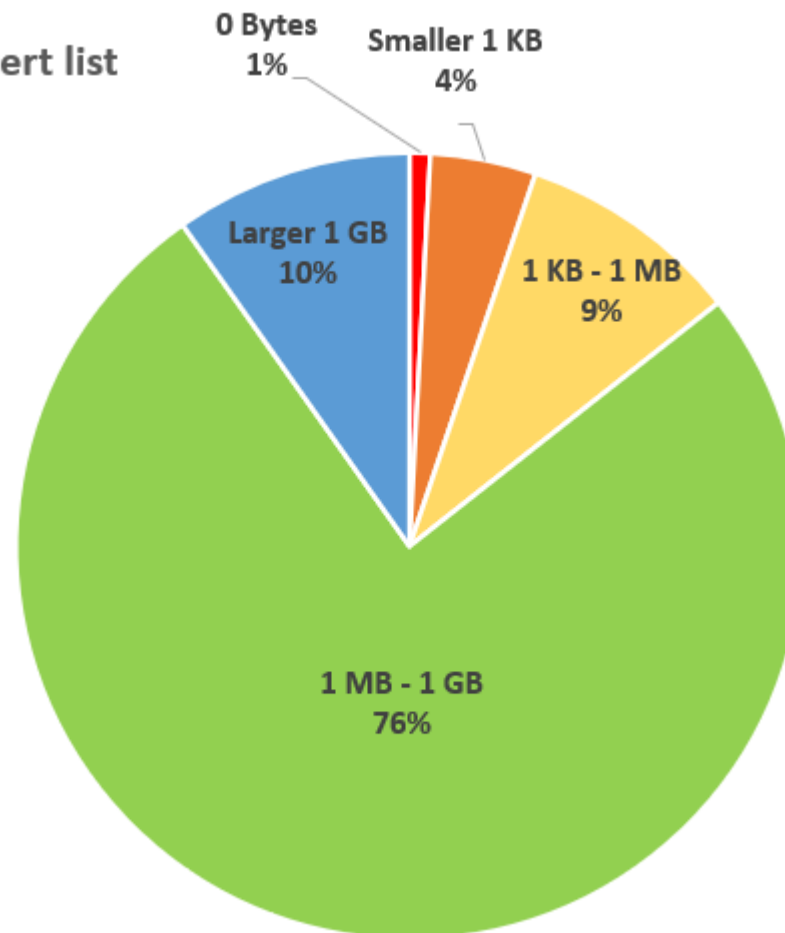
	Start time	End time	Duration
.frl harvest (2022)	14-3-2022 09:48	19-4-2022 10:51	5 weeks 1 day, 1 hour, 3 minutes
Expert list	20-5-2022 11:40	30-6-2022 10:37	5 weeks 5 days, 22 hours, 57 minutes

# Seedlist Quality Matters

BYTES HARVESTED  
.FRL DOMEIN



Expert list





# Crawling Top Level Domain

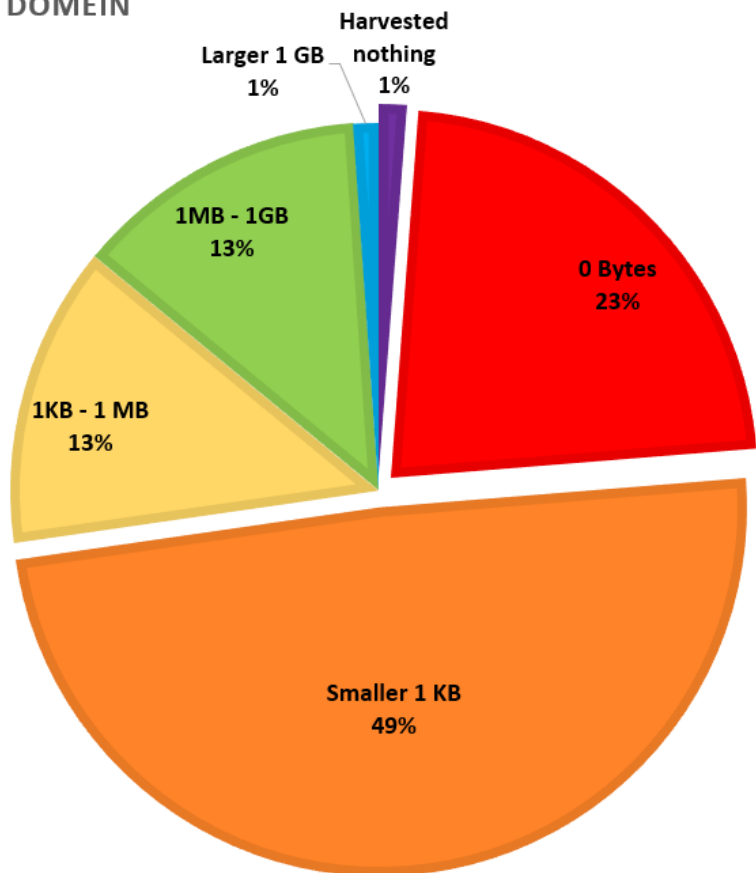
*Harvest results according to NAS metadata*

	Bytes	Documents	Domains
.frl harvest (2022)	296.115.975.995	3.035.059	9592
Expert list	180.520.734.507	2.098.732	828
.frl harvest (2023)	476.344.883.541	5.950.322	6898

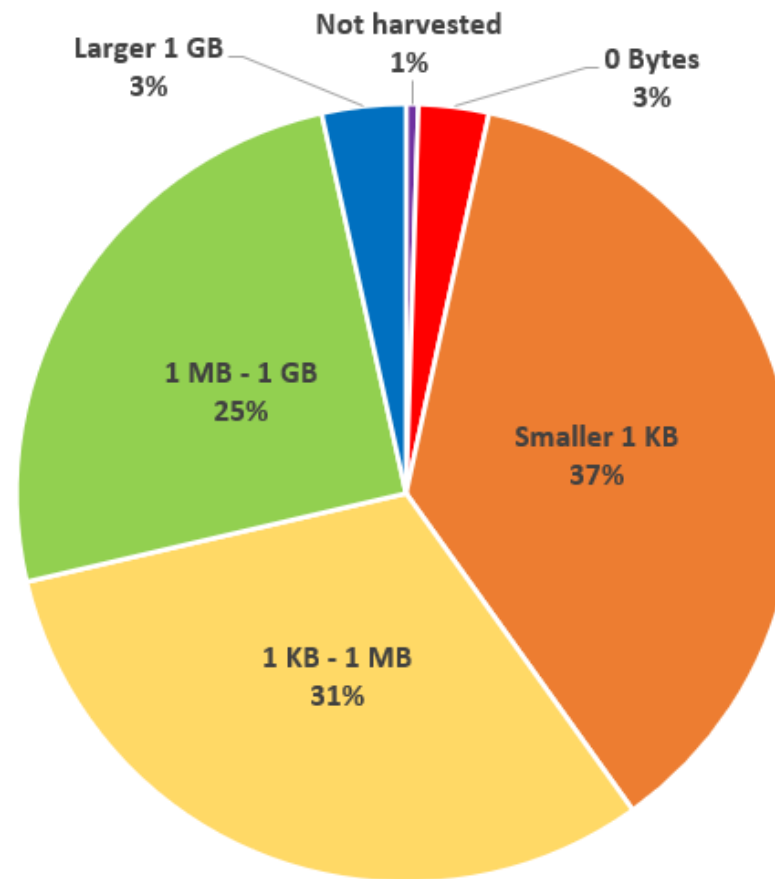
	Start time	End time	Duration
.frl harvest (2022)	14-3-2022 09:48	19-4-2022 10:51	5 weeks 1 day, 1 hour, 3 minutes
Expert list	20-5-2022 11:40	30-6-2022 10:37	5 weeks 5 days, 22 hours, 57 minutes
.frl harvest (2023)	30-11-2022 10:55	13-2-2023 15:30	Almost 3 months

# Seedlist Quality Matters

BYTES HARVESTED  
.FRL DOMEIN



FRL 2023



# Profile Settings Wikipedia

[https://fy.wikipedia.org/wiki/23\\_novimber](https://fy.wikipedia.org/wiki/23_novimber)  
[https://fy.wikipedia.org/wiki/Wiki:Koartlyn\\_feroare](https://fy.wikipedia.org/wiki/Wiki:Koartlyn_feroare)  
[https://fy.wikipedia.org/wiki/Wiki:Samar\\_in\\_side](https://fy.wikipedia.org/wiki/Wiki:Samar_in_side)  
<https://www.mediawiki.org/wiki/Special:MyLanguage/Help:Contents>  
[https://donate.wikimedia.org/wiki/Special:FundraiserRedirector?utm\\_source=donate&utm\\_medium=sidebar&utm\\_campaign=C13\\_fy.wikipedia.org&uselang=fy](https://donate.wikimedia.org/wiki/Special:FundraiserRedirector?utm_source=donate&utm_medium=sidebar&utm_campaign=C13_fy.wikipedia.org&uselang=fy)

[https://fy.wikipedia.org/wiki/Obe\\_Sikkes\\_Bangma](https://fy.wikipedia.org/wiki/Obe_Sikkes_Bangma)  
[https://fy.wikipedia.org/w/index.php?title=Oerlis:Obe\\_Sikkes\\_Bangma&action=edit&redlink=1](https://fy.wikipedia.org/w/index.php?title=Oerlis:Obe_Sikkes_Bangma&action=edit&redlink=1)

[https://fy.wikipedia.org/wiki/Obe\\_Skkes\\_Bangma](https://fy.wikipedia.org/wiki/Obe_Skkes_Bangma)  
[https://fy.wikipedia.org/w/index.php?title=Obe\\_Sikkes\\_Bangma&veaction=edit](https://fy.wikipedia.org/w/index.php?title=Obe_Sikkes_Bangma&veaction=edit)  
[https://fy.wikipedia.org/w/index.php?title=Obe\\_Sikkes\\_Bangma&action=edit](https://fy.wikipedia.org/w/index.php?title=Obe_Sikkes_Bangma&action=edit)  
[https://fy.wikipedia.org/w/index.php?title=Obe\\_Sikkes\\_Bangma&action=history](https://fy.wikipedia.org/w/index.php?title=Obe_Sikkes_Bangma&action=history)

[https://fy.wikipedia.org/wiki/Wiki:Myn\\_oerlis](https://fy.wikipedia.org/wiki/Wiki:Myn_oerlis)  
[https://fy.wikipedia.org/wiki/Wiki:Myn\\_bydragen](https://fy.wikipedia.org/wiki/Wiki:Myn_bydragen)  
[https://fy.wikipedia.org/w/index.php?title=Wiki:Nije\\_ynstellings\\_oanmeitsje&returnto=Obe+Sikkes+Bangma](https://fy.wikipedia.org/w/index.php?title=Wiki:Nije_ynstellings_oanmeitsje&returnto=Obe+Sikkes+Bangma)  
<https://fy.wikipedia.org/w/index.php?title=Wiki:Oanmelde&returnto=Obe+Sikkes+Bangma>

[rmei\\_keppele/](#)  
[keppelings/](#)  
[Bndere\\_siden](#)  
[id=904916](#)  
[ction=info](#)  
[angma&id=904916&](#)  
[Page/Q2689267](#)

The screenshot shows the Wikipedia article for "Obe Sikkes Bangma" in the Frisian language. The browser address bar is [fy.wikipedia.org/wiki/Obe\\_Sikkes\\_Bangma](https://fy.wikipedia.org/wiki/Obe_Sikkes_Bangma). The article title is "Obe Sikkes Bangma". The text describes him as a mathematician and seafarer from Penjum, born on 30 March 1768 and died on 23 November 1829. It mentions his work as a secretary and editor of the "Tydskriften fan it Wiskundich Genoatskip" and his role as a seafarer's school teacher in Amsterdam. A table of contents is visible, listing sections like "Wurk", "Sekundêre literatuer", and "Keppeling om utens". The article content is partially obscured by a red box. The navigation bar at the top right includes "Lêze", "Bewurkje", "Boarne bewurkje", and "Skiednis besjen". The sidebar on the left contains links for "Haadside", "Wikipedy-mienskip", "Hjoed", "Koartlyn feroare", "Samar in side", "Help", "Donaasjes", "Hjirmei keppele", "Keppelings folgje", "Bysûndere siden", "Fêste keppeling", "Sidegegevens", "Dizze side siterje", and "Wikidata-ïtem".

[https://fy.wikipedia.org/wiki/Wiki:Myn\\_oerlis](https://fy.wikipedia.org/wiki/Wiki:Myn_oerlis)  
[https://fy.wikipedia.org/wiki/Wiki:Myn\\_bydragen](https://fy.wikipedia.org/wiki/Wiki:Myn_bydragen)  
[https://fy.wikipedia.org/w/index.php?title=Wiki:Nije\\_ynstellings\\_oanmeitsje&returnto=Obe+Sikkes+Bangma](https://fy.wikipedia.org/w/index.php?title=Wiki:Nije_ynstellings_oanmeitsje&returnto=Obe+Sikkes+Bangma)  
<https://fy.wikipedia.org/w/index.php?title=Wiki:Oanmelde&returnto=Obe+Sikkes+Bangma>

# Profile Settings Wikipedia

*Excludes:*

## **Edit pages**

. \*veaction\=edit.\*  
. \*action\=edit.\*  
. \*action\=formedit.\*

## **Login pages**

. \*oanmeitsje.returnto\=.\*  
. \*Oanmelde.returnto\=.\*  
. \*Wiki\:Myn\\_oerlis.\*  
. \*Wiki\:Myn\\_bydragen.\*

## **Formats**

. \*bookcmd\=book\\_creator.referer\=.\*  
. \*action\=show\-download\-screen.\*  
. \*action\=smartbook.\*  
. \*mobileaction\=toggle\\_view\\_mobile.\*  
. \*\printable\=yes.\*

# Crawling Top Level Domain

*Harvest results according to NAS metadata*

	Bytes	Documents	Domains
.frl harvest (2022)	296.115.975.995	3.035.059	9592
Expertlijst	180.520.734.507	2.098.732	828
.frl harvest (2023)	476.344.883.541	5.950.322	6898
fy.wikipedia.org	90.272.584.820	947.553	1

	Start time	End time	Duration
.frl harvest (2022)	14-3-2022 09:48	19-4-2022 10:51	5 weeks 1 day, 1 hour, 3 minutes
Expertlijst	20-5-2022 11:40	30-6-2022 10:37	5 weeks 5 days, 22 hours, 57 minutes
.frl harvest (2023)	30-11-2022 10:55	13-2-2023 15:30	Almost 3 months
fy.wikipedia.org	7-7-2022 09:57	12-7-2022 18:03	5 days, 8 hours, 6 minutes

# Lessons learned: Ad Hoc Issues During Pilot

- Issues with installing and using tool
  - Importing URLs
  - Available metadata
  - + Importing settings and crawler traps
- Understanding the settings before testing
- Process for quality control
- Preservation requirements


# Lessons learned: Preservation

- How would we preserve a national domain?
- Some core principles for future policies on a national domain:
  - Integrity:
    - Checksums on file level.
    - Preserve context information (settings, log files and collection descriptions) with the IP.
  - Authenticity:
    - Determine default settings for a domain crawl.
    - Reconstruct preservation events (creation) for the life cycle of digital objects.

# Lessons learned: Selection

- Keep seedlist up-to-date
- Reserved Names

- Punycode websites

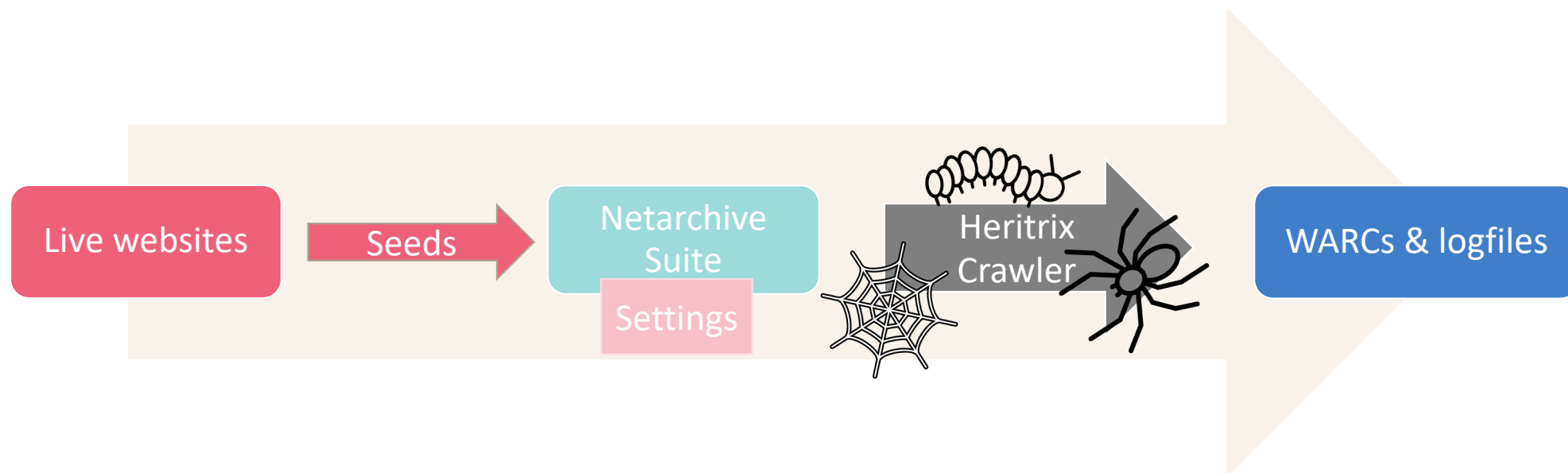
xn--112frysln-i2a.frl =  https://112fryslâ.n.frl

- Check response
  - Websites without IP-adres
  - Websites that redirect



# Conclusion

- Lots of trial and error
- There is no handbook how to do it 'right'
- Understanding settings and tool is very important.





KB } nationale  
bibliotheek