Making Web Collections for Research Sustainable & Reusable

Possibilities and Challenges Experienced

ELD ZIERAU
Digital Preservation Specialist, PhD

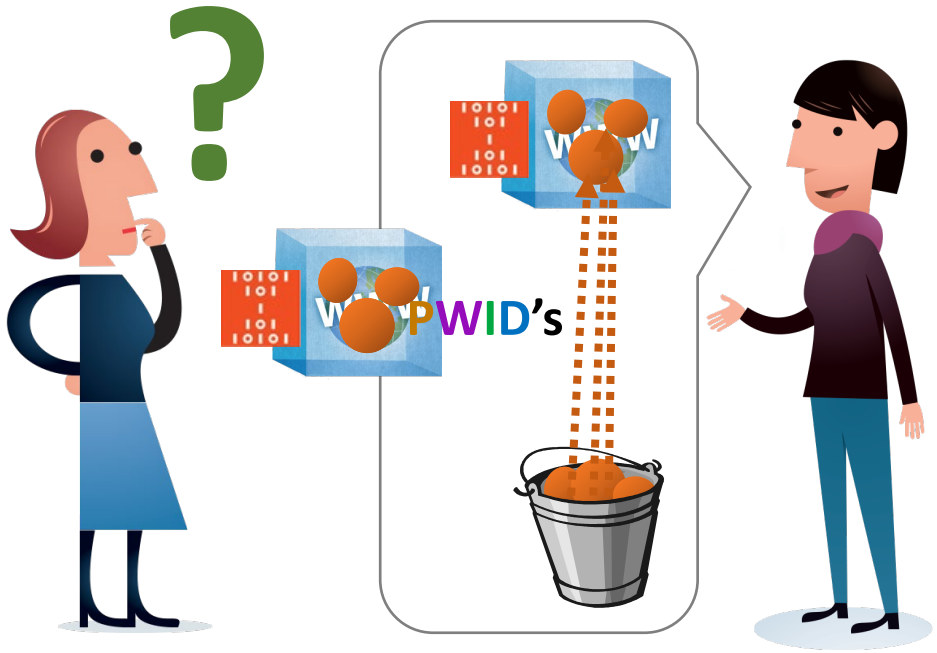
PER MØLDRUP-DALUM
IT Consultant

IIPC 2021, Virtual
June 2021

DET KGL. BIBLIOTEK
Royal Danish Library

Making Web Collections for Research Sustainable & Reusable

Possibilities and Challenges Experienced



A Case Story

"Probing a Nation's Web Domain"

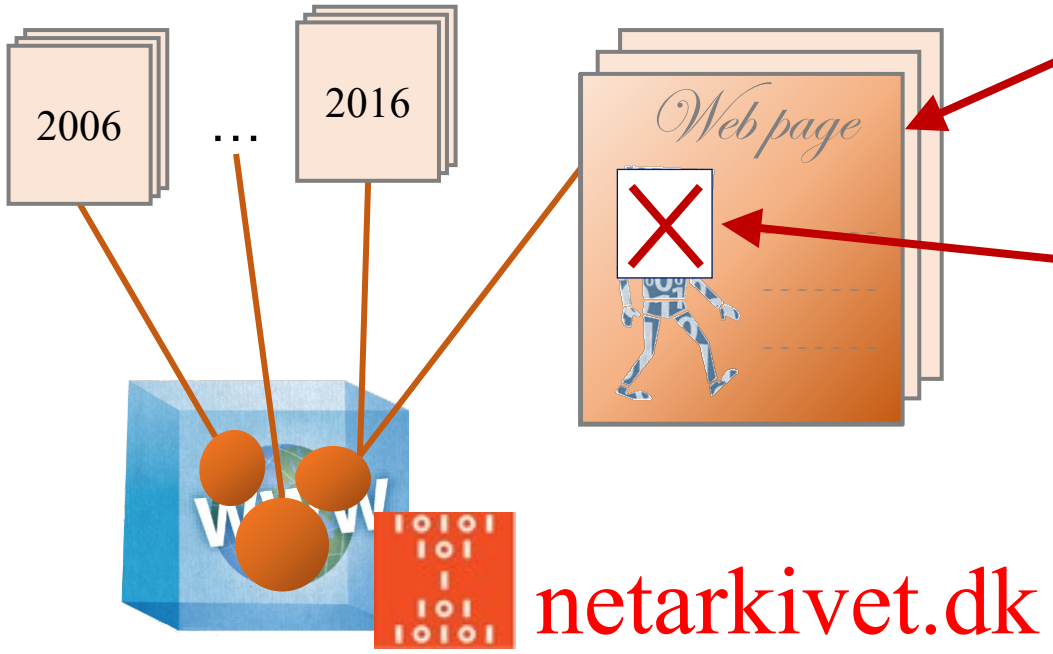
Motivation

Ensuring persistency of web collections for later reuse and result verification:

- assess reliability and provenance
- retrace and reproduce research steps
- enable continued work

The Case Story

Project “Probing a Nation’s Web Domain”
- Investigating changes of the Danish web preserved in Netarkivet
- Approach: compare annually collection - only one occurrence of elements



html for http://digitalbevaring.dk/hjemmesider/
harvested: 2016-12-03 05:04:37 UTC

Image http://digitalbevaring.dk/uploads/image.png
harvested: 2011-08-21 06:52:06 UTC

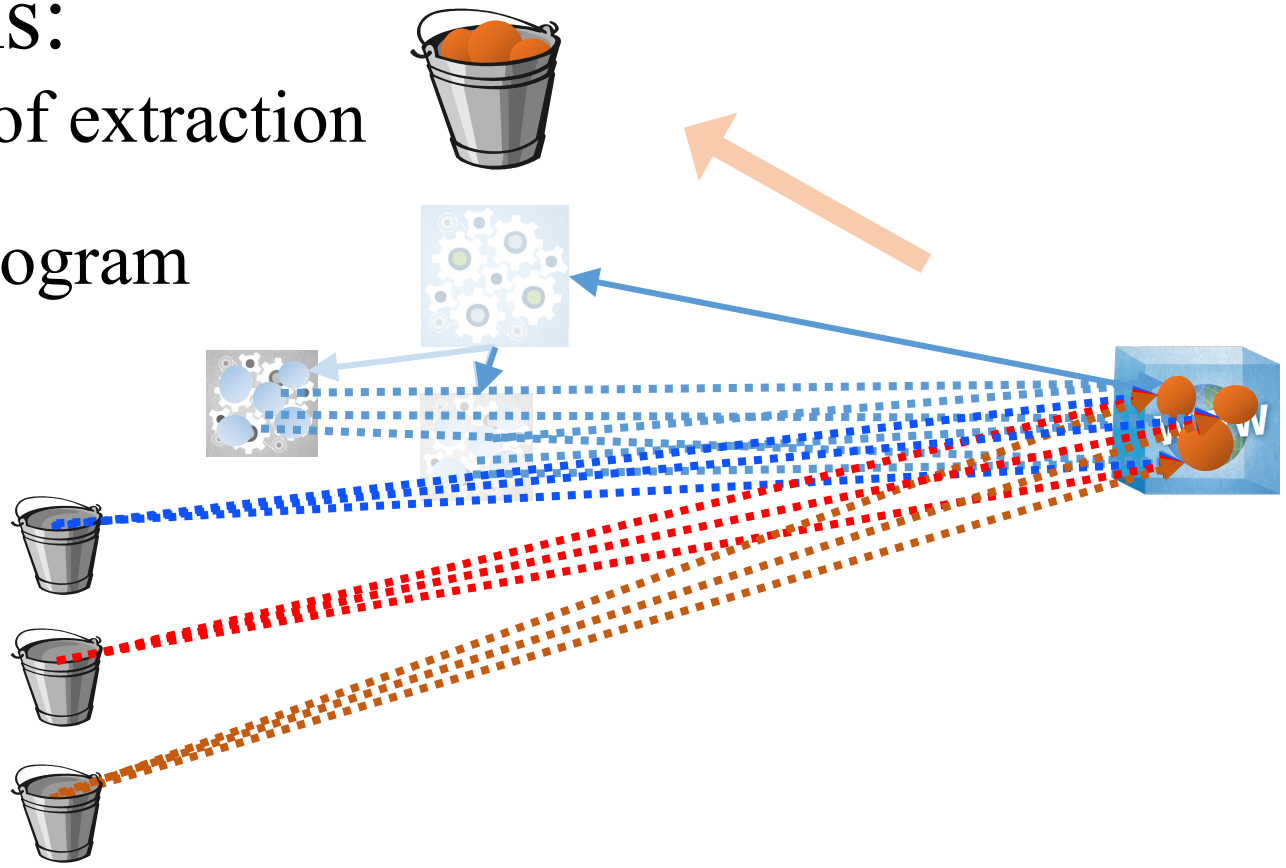
Only metadata and text from elements with text

Persistency of the web collections is essential for future use of the project results.

Challenges in making it reusable

Investigated options:

- Separate preservation of extraction
- Collection selection program
- Results from program
- Archive URL list
- CDX list
- PWID list



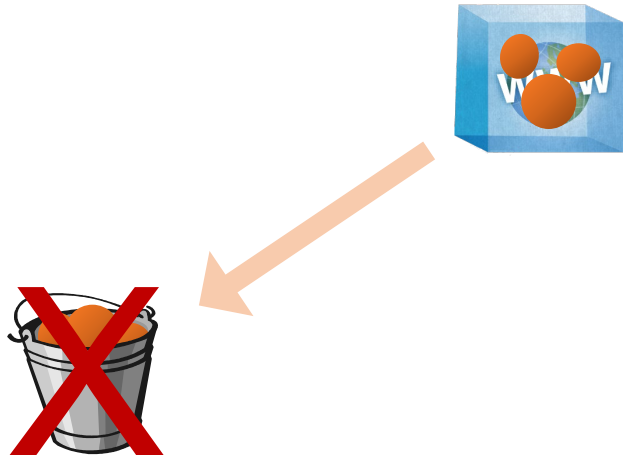
Separate preservation of extraction

Challenges

- Size

11 corpora – one per year
2006 -2016

In total app. <u>24 TB</u>



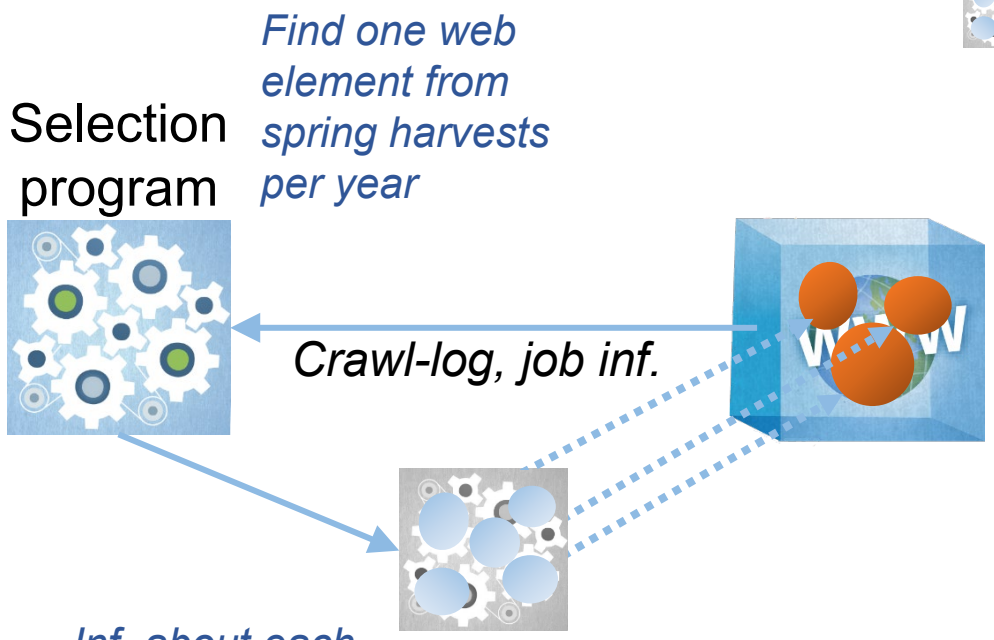
Extract is not feasible

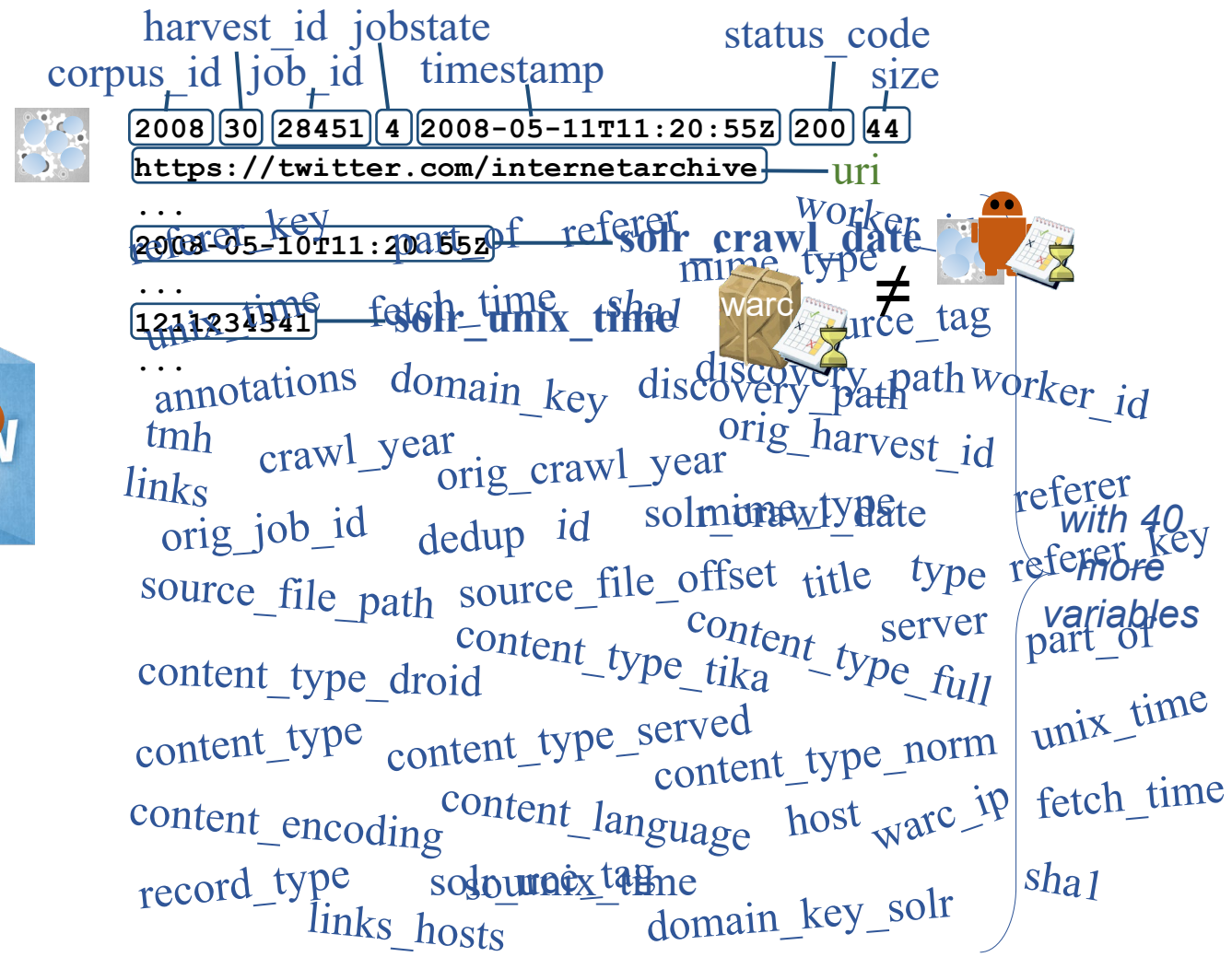
- Jurisdictions

Legal issues

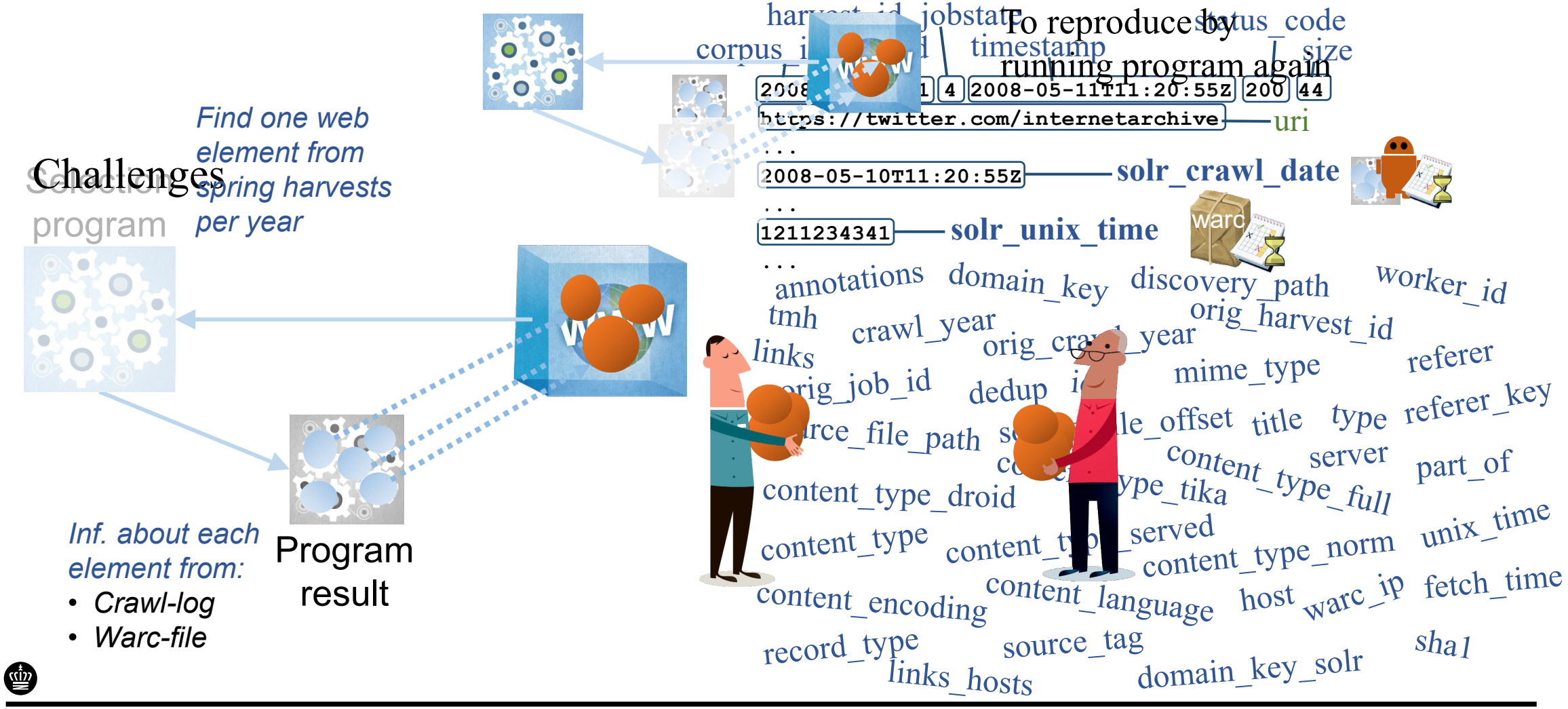
Handout, but
deletion after 5 years

Collection selection program

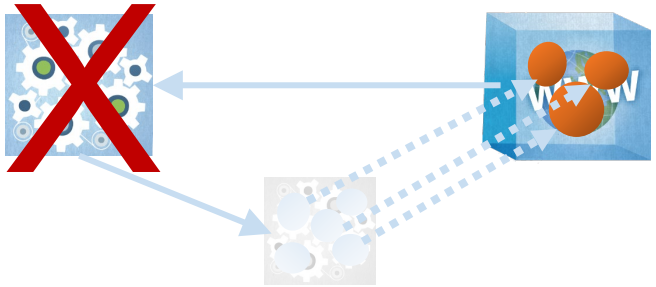




Collection selection program



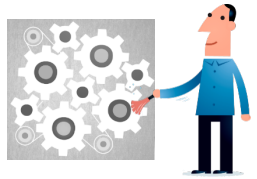
Collection selection program



To reproduce by
running program again

Challenges

- Hard to preserve programs long term



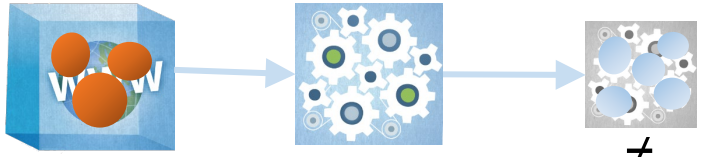
emulation

documentation

changed archive APIs

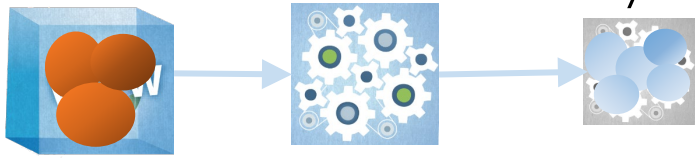
- Enriched web archive data in data range

2020

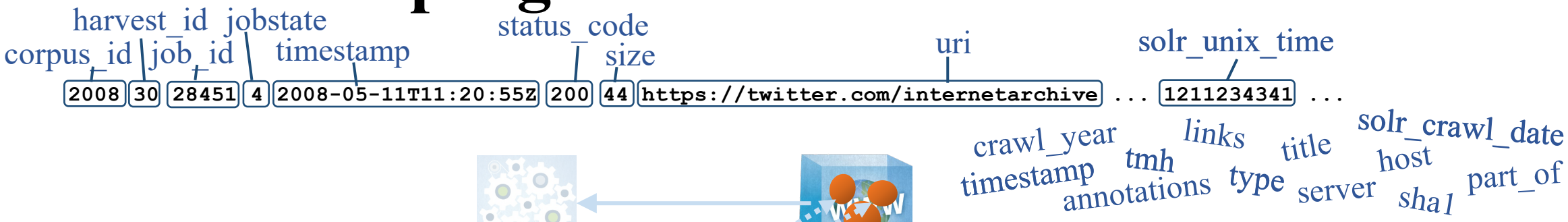


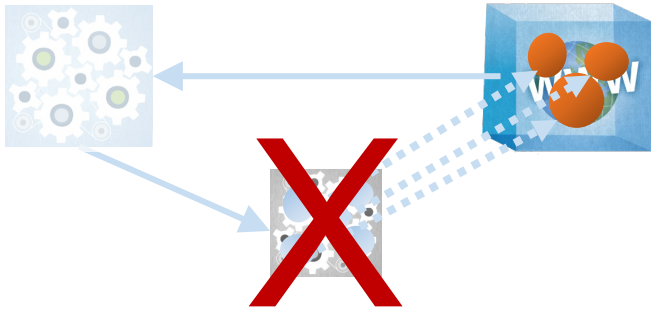
≠

2030



Results from program





Challenges

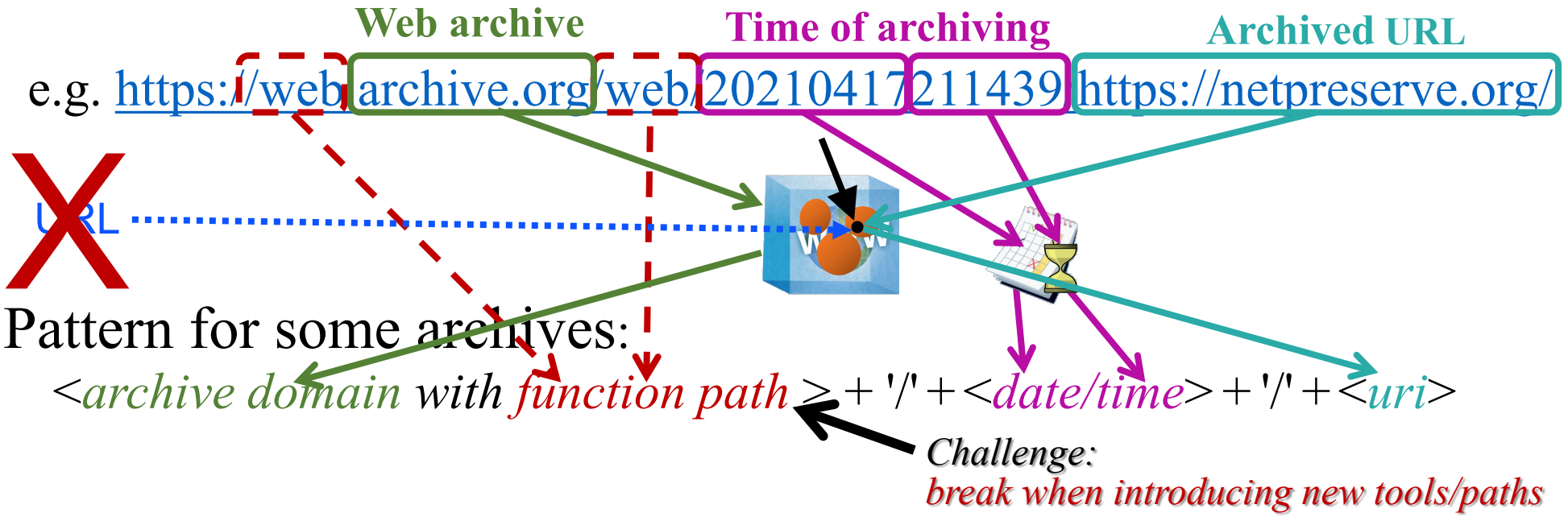
- Too much information

  Do not need harvest_id, jobstate, corpus_id, crawl_year, job_id, timestamp, status_code, referer, tmh, size, annotations, . . .

- Understandable in the future?

- Not a standard – but case based

Archive URL list



Other patterns (Danish Web Archive)

archive file & off-set ← *Challenge:* *break when migrating to .gz*

23171-53-20080102165640-00082-sb-prod-har-001.statsbiblioteket.dk.arc/3323344

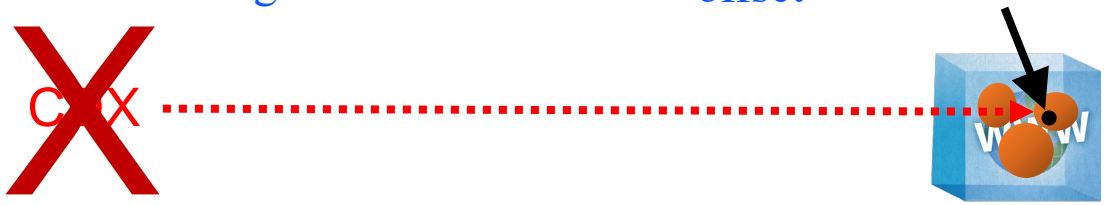
Not persistent / Not technology agnostic

CDX list

url-key timestamp original-url mime-type status-code

```
id.kb.dk/pwid/pwid.ppsm 20210518072011 http://id.kb.dk/pwid/pwid.ppsm text/html 200
5MLNDFDYNYWROURB6XMUNZITGBIFGNCO - 262066812 364550-333-20210518063849309-00002-kb-prod-har-010.kb.dk.warc.gz
```

digest offset file-name



Challenges

- Too much information

  Do not need url-key mime-type file-name status-code digest offset

- Too little information for special cases

  If parts in CDX-list is from different web archives

- Not stabile standard

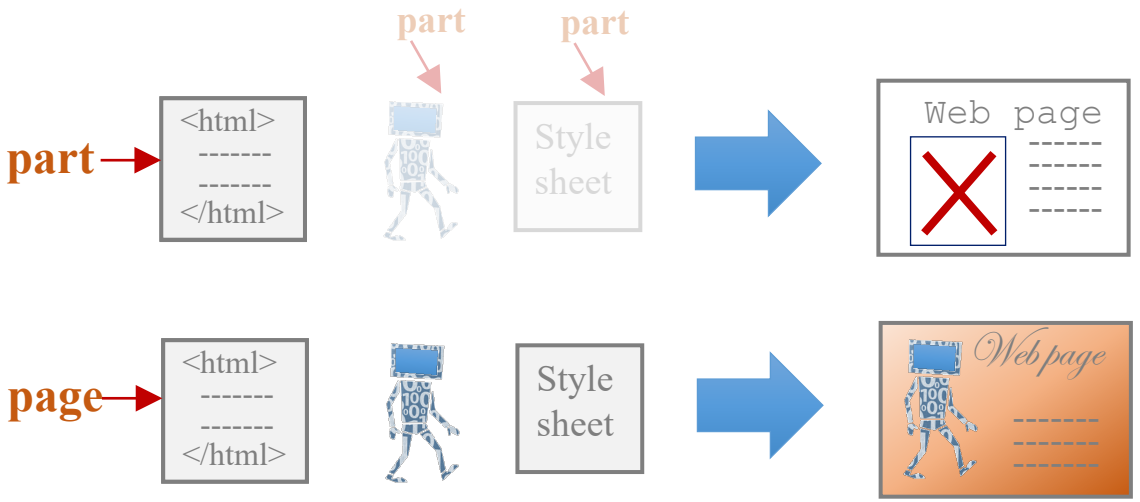
Persistent Web Identifier (PWID) list

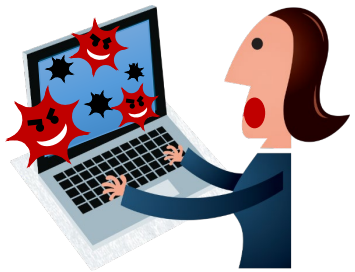
Web archive | Time of archiving | Precision | Archived URL

e.g. pwid:archive.org:2021-04-17T21:14:39Z:part:https://netpreserve.org/

Precision

Recommendation:
Use Part for collections



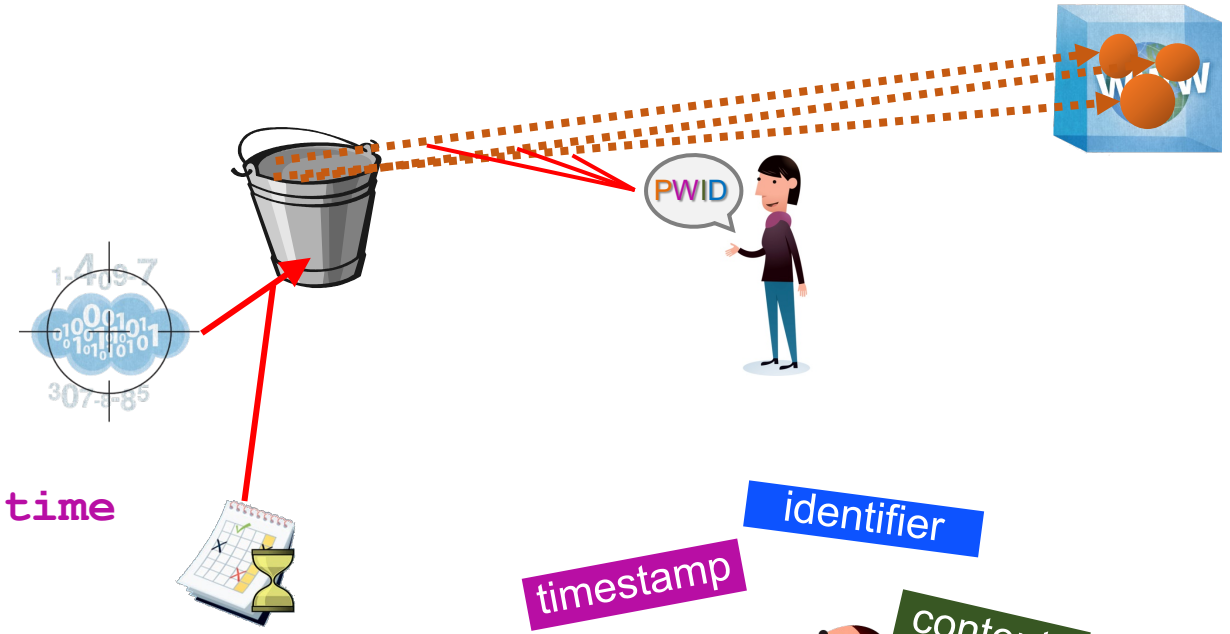


PWID collection

Collection definition: `collection`

- **Identifier** `collection-identifier`
  findable and re-usable corpus
- **Timestamp** `collection-archival-time`
  distinguish different versions
  registered at different times (UTC)
- **Contents** `collection-contents`
  persistent global references to corpus parts
  use PWIDs

In structure with standard names

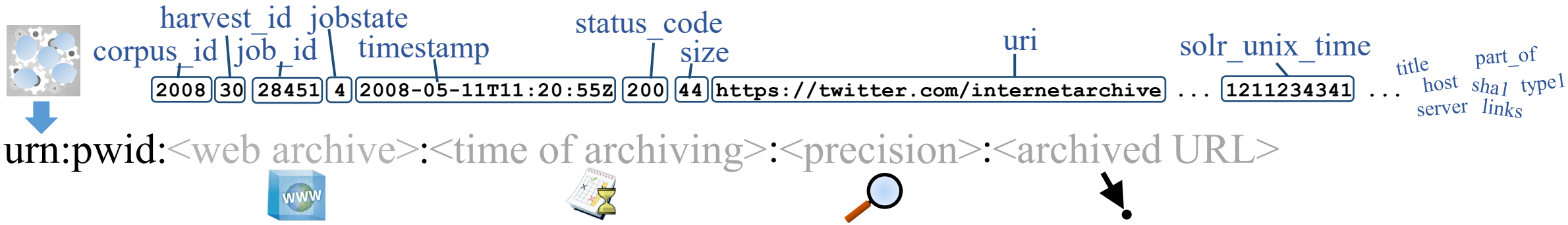


Other metadata elsewhere

Example in XML

```
<collection>
 <collection-identifier>
    urn:uuid:631578c2-a8cf-11eb-bcbc-0242ac130002
 </collection-identifier>
 <collection-archival-time>2021-05-01T12:04:40Z</collection-archival-time>
 <collection-contents>
  <part>urn:pwid:netarkivet.dk:2008-05-20T12:32:01Z:part:http://dr.dk/Nyheder/Temaer/
Politik+temaer/2007/Valg/2007/11/14/150307.htm
  </part>
  <part>urn:pwid:netarkivet.dk:2008-05-19T21:45:27Z:part:http://www.dr.dk/DR2/Temaaften/
Udsendelser/tirsdag/2005/20060224111822.htm
  </part>
  <part>urn:pwid:netarkivet.dk:2008-05-19T04:38:09Z:part:http://www.dr.dk/Regioner/Kbh/
Nyheder/Furesoe/2008/04/23/073010.htm
  </part>
  ...
 </collection-contents>
</collection>
```

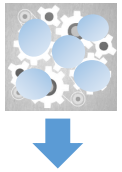
Creation of PWID collection



Challenges:

- Timestamp

Creation of PWID collection



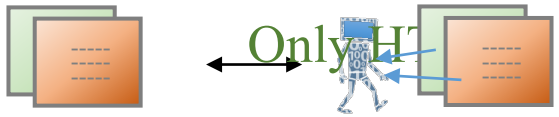
harvest_id jobstate
corpus_id job_id timestamp status_code size uri solr_unix_time title part_of host sha1 type1 server links

```
2008 30 28451 4 2008-05-11T11:20:55Z 200 44 https://twitter.com/internetarchive ... 1211234341 ...
```

urn:pwid:<web archive>:<time of archiving>:<precision>:<archived URL>

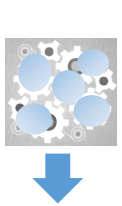
Challenges:

- Timestamp
- Deduplication



Only HTML
Only what researchers have

Creation of PWID collection



harvest_id jobstate
corpus_id job_id timestamp status_code size uri solr_unix_time title part_of host sha1 type1 server links

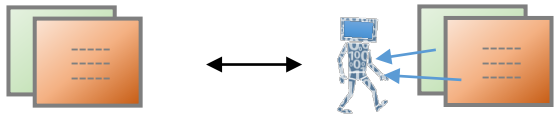
```
2008 30 28451 4 2008-05-11T11:20:55Z 200 44 https://twitter.com/internetarchive ... 1211234341 ...
```

urn:pwid:<web archive>:<time of archiving>:<precision>:<archived URL>

Challenges:

- Timestamp

- Deduplication



Learnings:

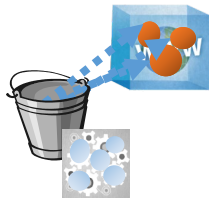
Make PWID collection from the start

Creation of PWID collection


```
log("Create a PWID data frame in Spark memory")
archive ← "netarkivet.dk"
sdf_pwid ← solr_corpus %>%
mutate(pwid_datetime = from_unixtime(solr_unix_time,"yyyy-MM-dd'T'HH:mm:ss'Z'")) %>%
mutate(pwid = paste("urn","pwid", archive, pwid_datetime, "part", uri, sep = ":")) %>%
select(pwid) %>%
compute(name = "sfd_pwid")
```

Statistics

Accumulated research data: 24TB (text and metadata)



Statistics

Accumulated research data: 24TB

year	Metadata size (GB)	PWID size (GB)
2006	67.249	23.005
2007	138.599	43.889
2008	152.322	51.752
2009	208.988	72.632
2010	220.594	70.008
2011	187.803	51.455
2012	194.682	54.698
2013	196.395	56.132
2014	188.742	48.378
2015	191.345	46.100
2016	182.085	38.621
	1,928.804	556.671

Statistics

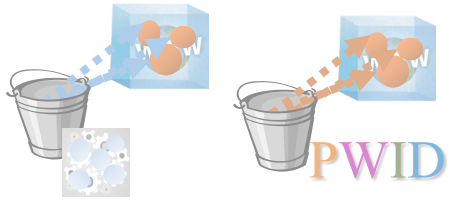
Accumulated research data: 24TB (text and metadata)

year	Metadata size (GB)	PWID size (GB)	No. of PWIDs
2006	67.249	23.005	167,892,081
2007	138.599	43.889	300,828,679
2008	152.322	51.752	332,899,148
2009	208.988	72.632	469,981,110
2010	220.594	70.008	473,144,026
2011	187.803	51.455	356,766,440
2012	194.682	54.698	366,267,978
2013	196.395	56.132	368,262,556
2014	188.742	48.378	316,052,138
2015	191.345	46.100	326,196,901
2016	182.085	38.621	255,325,163
	1,928.804	556.671	3.733.616,220

Statistics

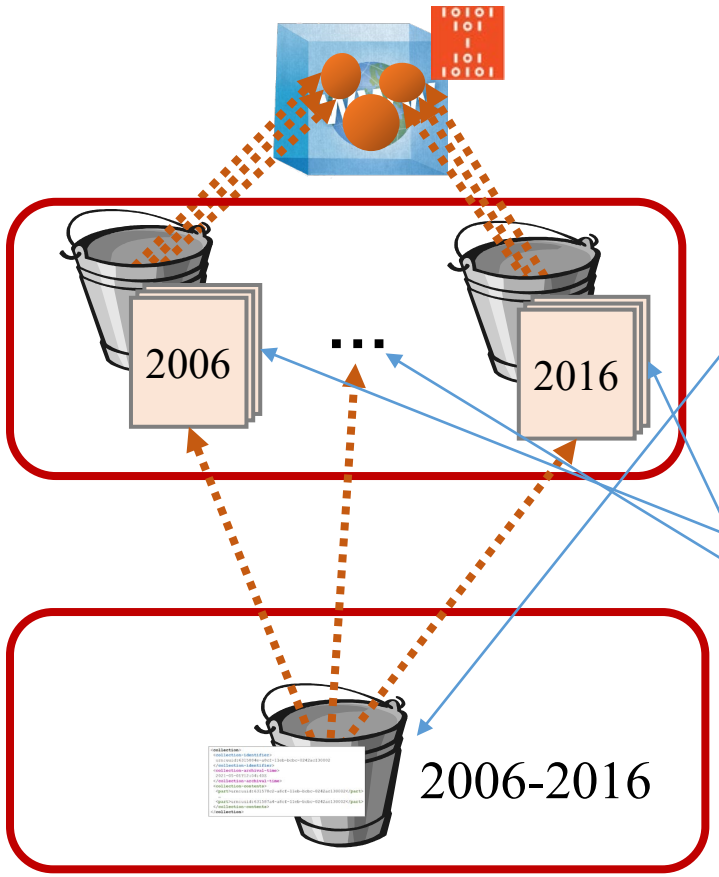
Accumulated research data: 24TB (text and metadata)

year	Metadata size (GB)	PWID size (GB)	No. of PWIDs	Generation of PWID		
				start	end	Δt
2006	67.249	23.005	167,892,081	15:08:20	15:23:08	14m48s
2007	138.599	43.889	300,828,679	15:23:13	16:00:13	37m00s
2008	152.322	51.752	332,899,148	16:00:19	16:48:55	48m36s
2009	208.988	72.632	469,981,110	16:49:00	17:48:56	59m56s
2010	220.594	70.008	473,144,026	17:49:02	18:44:00	54m58s
2011	187.803	51.455	356,766,440	18:44:05	19:22:18	38m13s
2012	194.682	54.698	366,267,978	19:22:23	20:04:50	42m27s
2013	196.395	56.132	368,262,556	20:04:56	20:50:01	45m05s
2014	188.742	48.378	316,052,138	20:50:07	21:26:04	35m57s
2015	191.345	46.100	326,196,901	21:26:09	21:58:48	32m39s
2016	182.085	38.621	255,325,163	21:58:53	22:24:42	25m49s
	1,928.804	556.671	3.733.616,220	15:08:20	22:24:42	7h16m22s



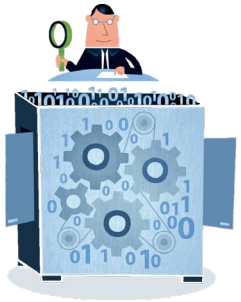
- National DeiC Cultural Heritage Cluster
- Hadoop Spark cluster:
- 9 Dell PowerEdge R730 servers
  - 36 hyper threaded cores
  - 256 GB RAM
  - 4 x 10Gb Ethernet
- HDFS with 288TB storage
- Use R as interface

Collection of collection



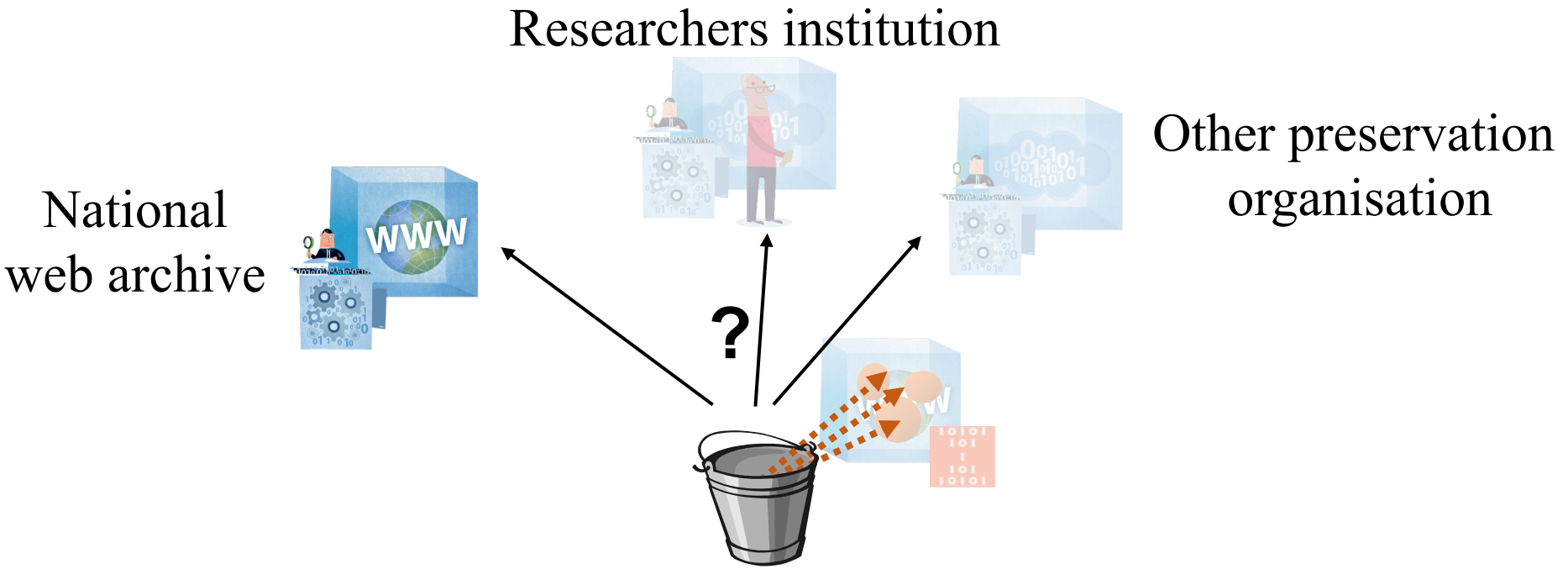

```
<collection>
 <collection-identifier>
  urn:uuid:6315884e-a8cf-11eb-bcbc-0242ac130002
 </collection-identifier>
 <collection-archival-time>
  2021-05-01T12:04:40Z
 </collection-archival-time>
 <collection-contents>
  <part>urn:uuid:631578c2-a8cf-11eb-bcbc-0242ac130002</part>
   ...
  <part>urn:uuid:631587a4-a8cf-11eb-bcbc-0242ac130002</part>
 </collection-contents>
</collection>
```

Preservation Challenges



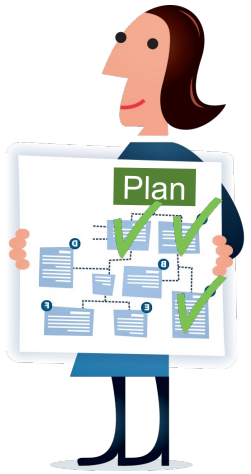
Persistent web corpora:

- As long as Netarkivet exist
- As long as PWID-collection is preserved



Status

PWID as an URN: Both for collection or single reference



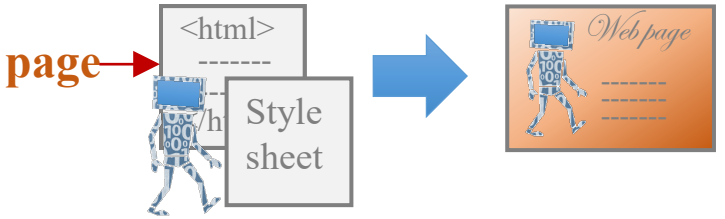
- PWID as an URN

In process since 2017
2 out 4 experts have approved

- Establish web archive domain registry

Subject for IIPC?

- PWID support and resolution possibilities



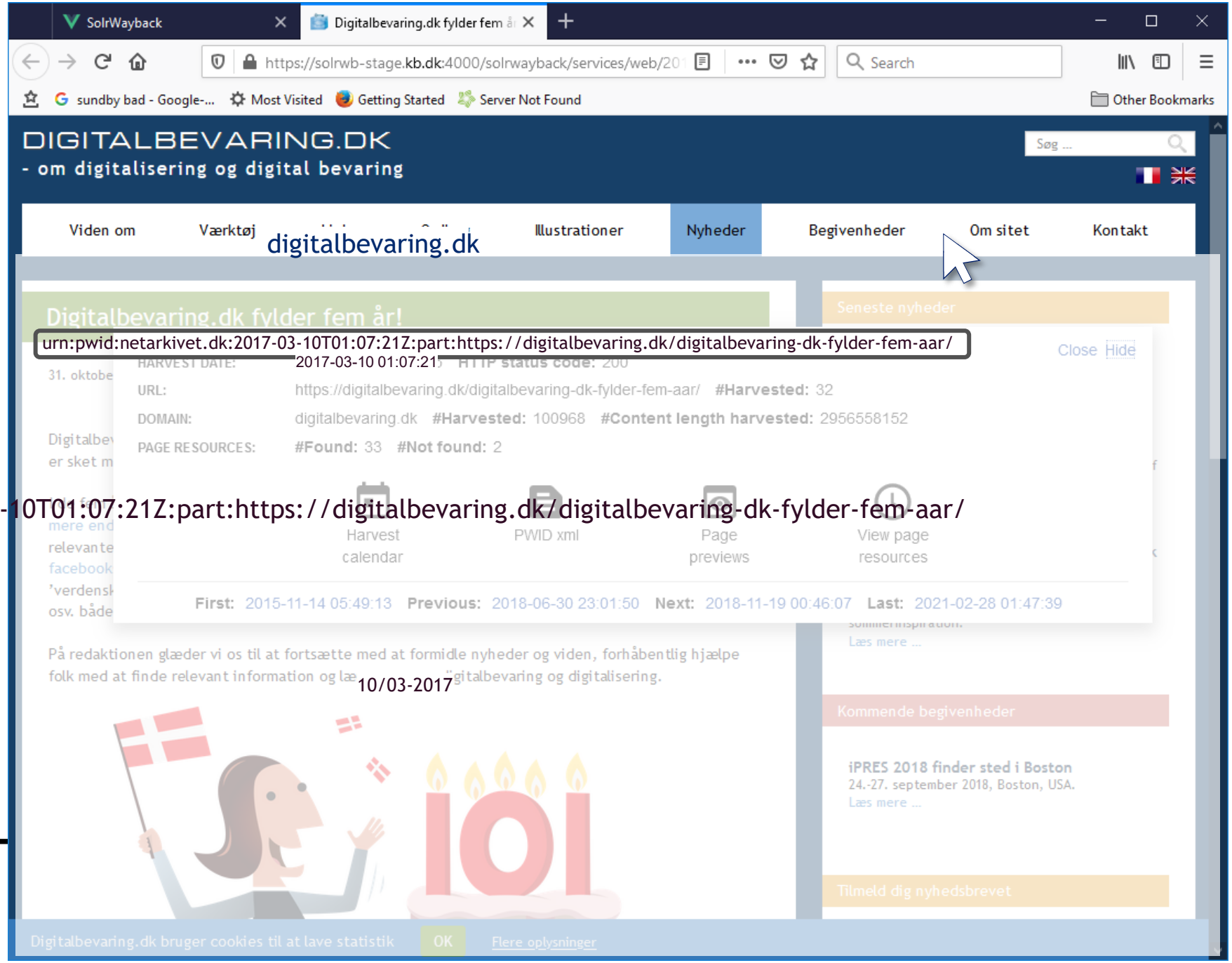
Prototype exists
IIPC? OPF?

- SOLRWayback support

DET KGL. BIBLIOTEK SOLRWAYBACK

PWID support

urn:pwid:netarkivet.dk:2017-03-10T01:07:21Z:part:https://digitalbevaring.dk/digitalbevaring-dk-fylder-fem-aar/



DET KGL. BIBLIOTEK
Royal Danish Library

PWID support

Resolve PWID

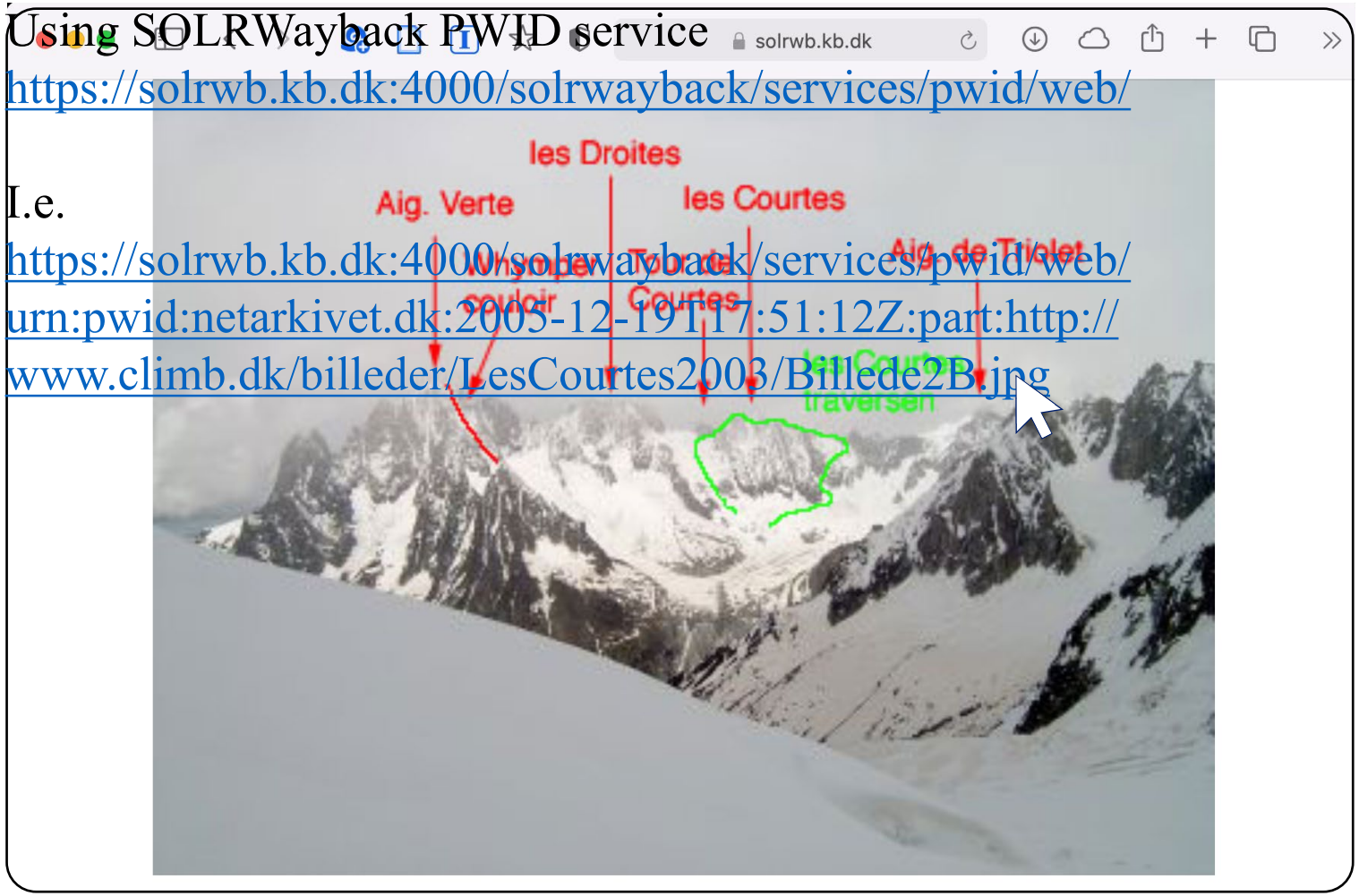
urn:pwid:netarkivet.dk:2005-12-19T17:51:12Z:part:http://www.climb.dk/billeder/LesCourtes2003/Billede2B.jpg

Using SOLRWayback PWID service

https://solrwb.kb.dk:4000/solrwayback/services/pwid/web/

I.e.

https://solrwb.kb.dk:4000/solrwayback/services/pwid/web/urn:pwid:netarkivet.dk:2005-12-19T17:51:12Z:part:http://www.climb.dk/billeder/LesCourtes2003/Billede2B.jpg



Status & Further work

PWID as an URN: Both for collection or single reference

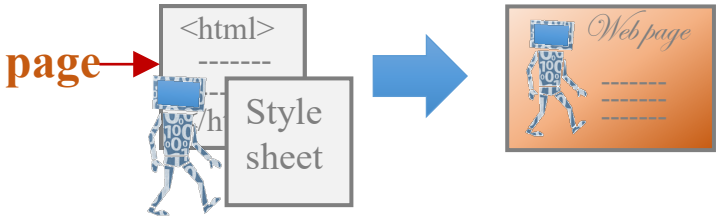
- PWID as an URN

In process since 2017
2 out 4 experts have approved

- Establish web archive domain registry

Subject for IIPC?

- PWID support and resolution possibilities



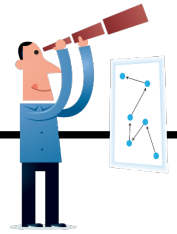
Prototype exists
IIPC? OPF?

- SOLRWayback support

DET KGL. BIBLIOTEK SOLRWAYBACK

- Memento delivering PWID collections for search results

- Policy for making PWIDs up front in research projects



Thank you for your attention!

Images in this style are from digitalbevaring.dk

More information on PWID available from http://id.kb.dk/pwid/PWID.ppsm
urn:pwid:2021-05-28T14:03:02:part:http://id.kb.dk/pwid/PWID.ppsm