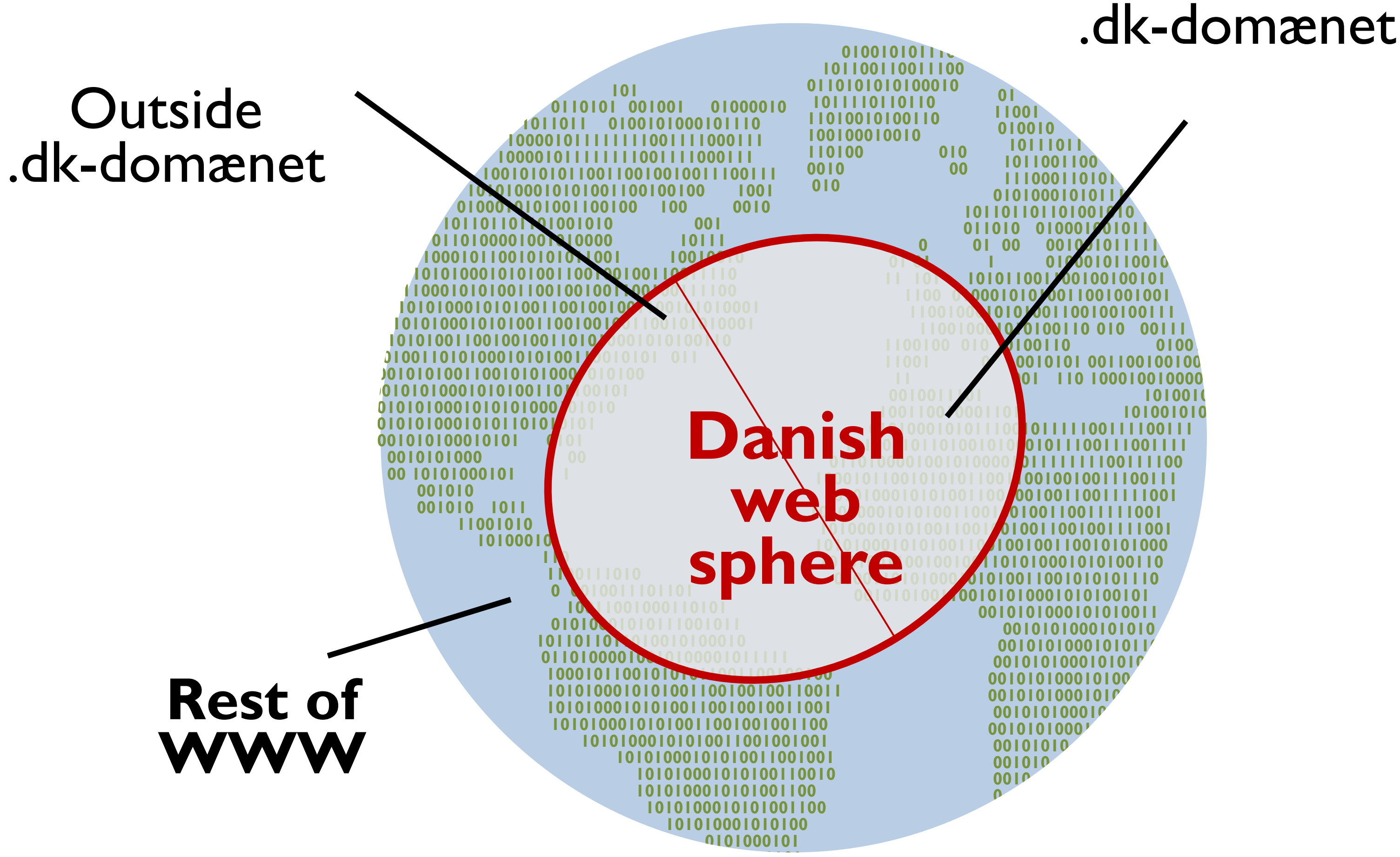
Looking Back, Looking Forward: New Strategies for Coverage of a National Web Sphere

By Eld Zierau
The Royal Library of Denmark
Ditte Laursen
The State and University Library of Denmark

The Web Sphere



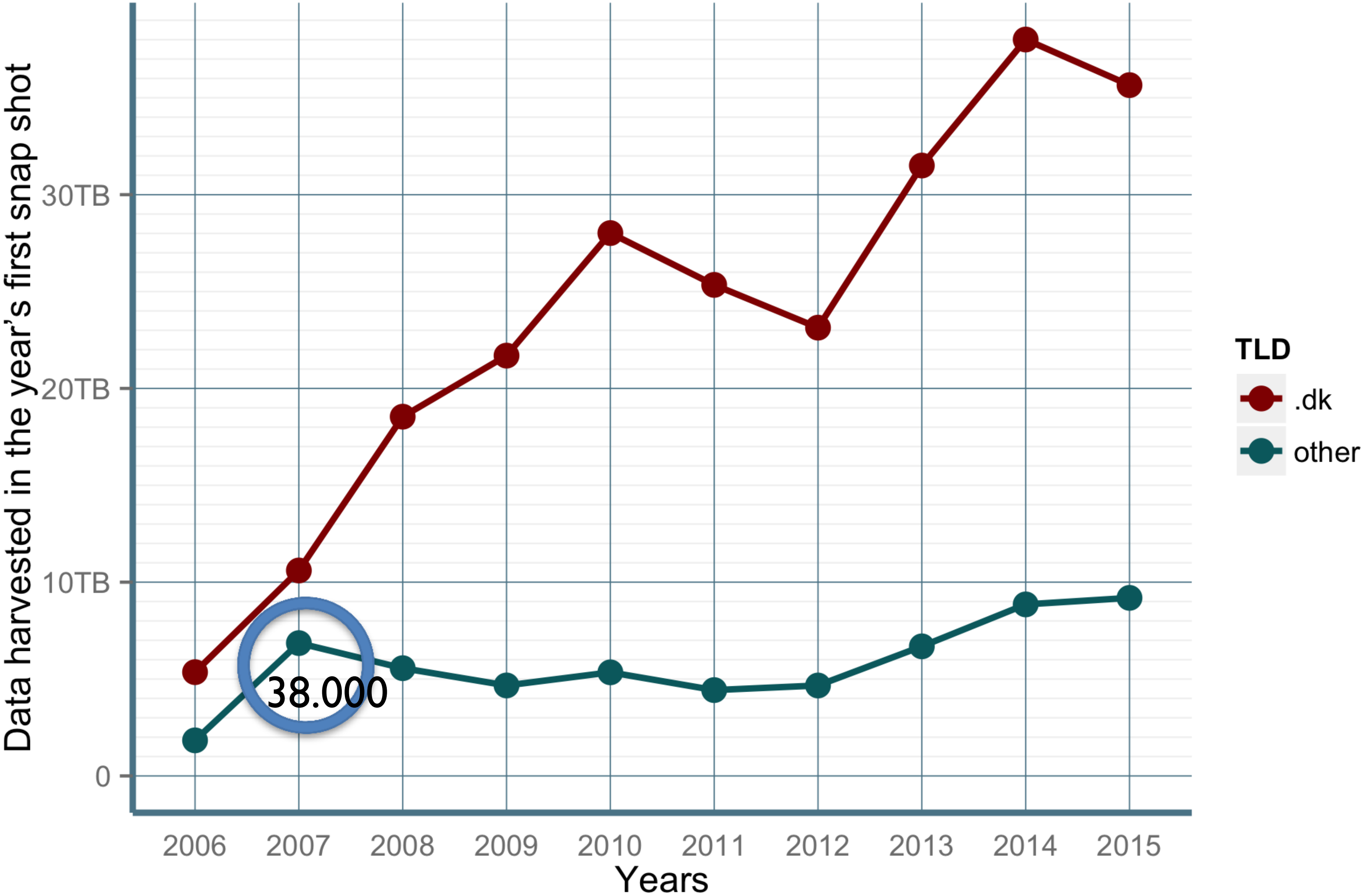
increasing amount of national webpages moves to generic Top Level Domains like .com or .org

Contents

- A bit of history

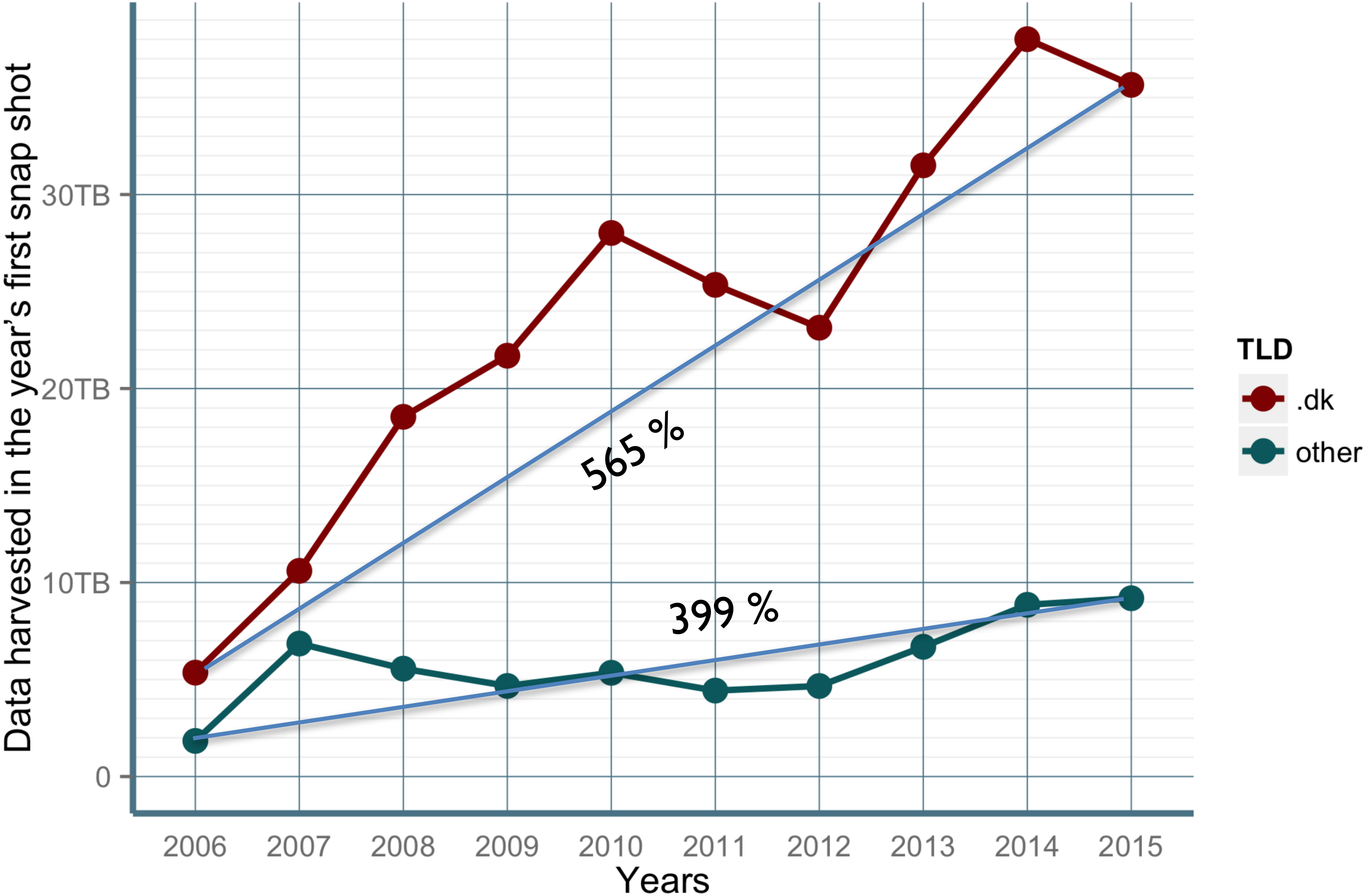
- The implementation

.dk and non .dk



38.000

.dk and non .dk


Data harvested in the year's first snap shot
30TB
20TB
10TB
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Years
565 %
399 %
TLD
.dk
other

Top 10 biggest domains

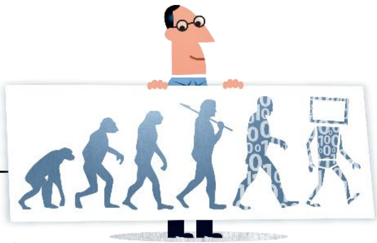
2006

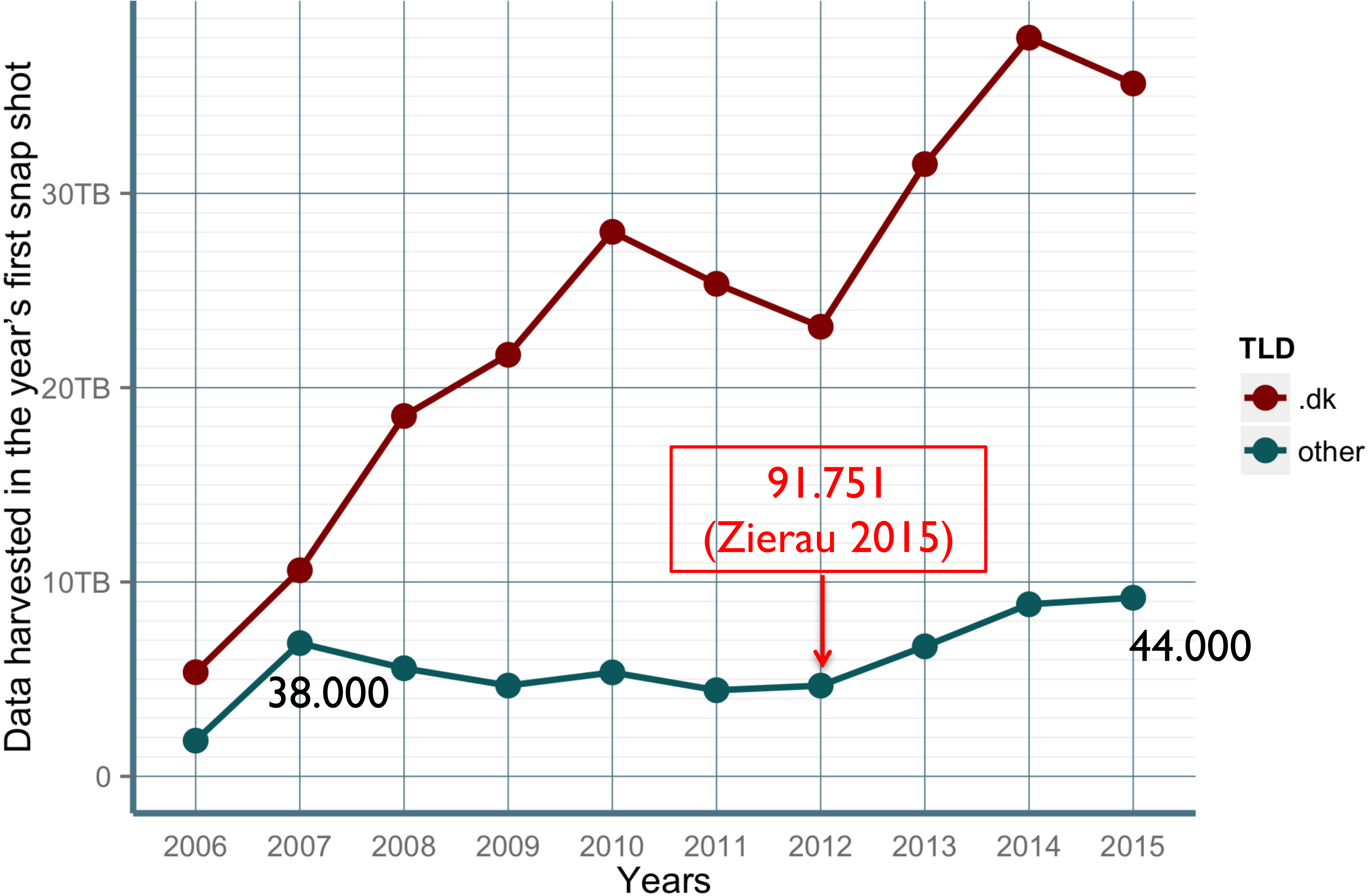
- *tv2.dk*
- *dbu.dk*
- *miniclip.com*
- *microsoft.com*
- *inforce.dk*
- *radioorbit.dk*
- *omegn.dk*
- *karstenskj.dk*
- *um.dk*
- *pornaccess.com*

2015

- *snatch.dk*
- *cloudfront.net*
- *pinterest.com*
- *twitter.com*
- *google.com*
- *vimeocdn.com*
- *googleusercontent.com*
- *blogspot.com*
- *amazonaws.com*
- *dmi.dk*

.dk and non .dk



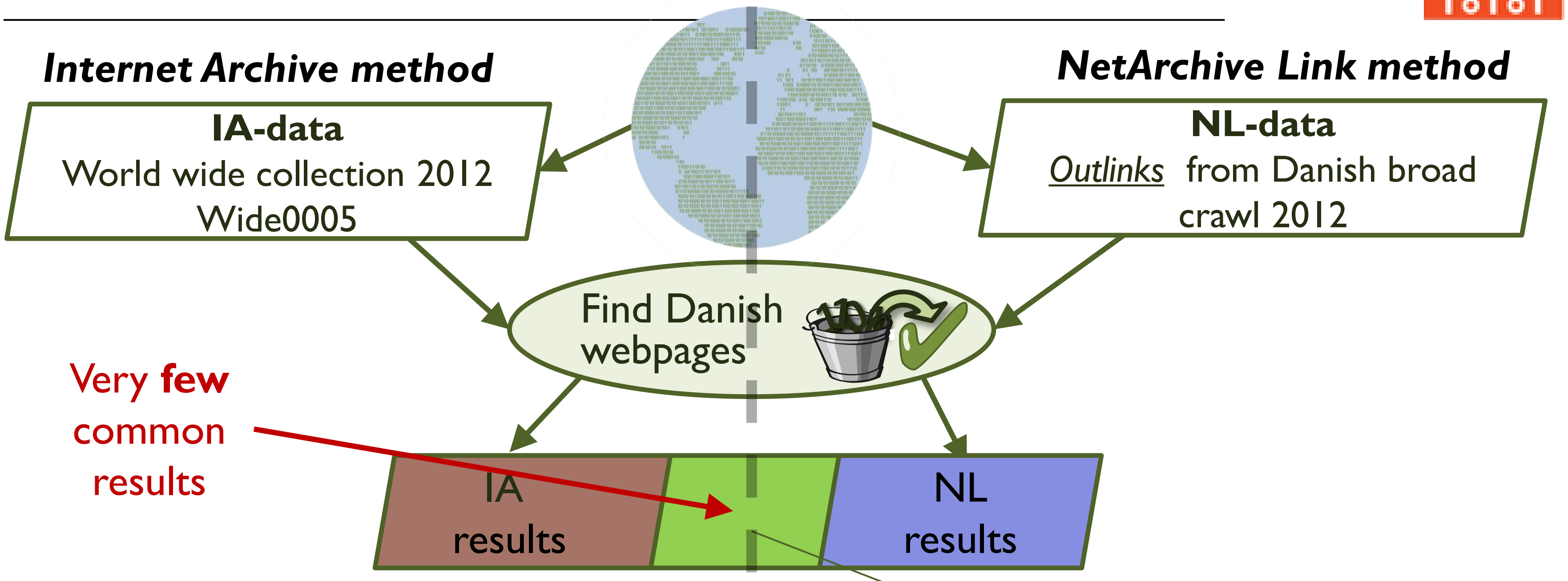


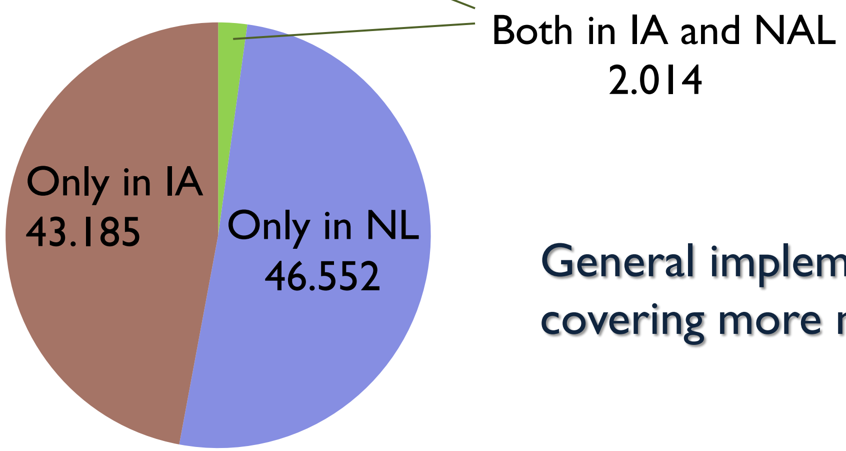
91.751
(Zierau 2015)

38.000

44.000

WebDanica project Tested Different Methods





Host: 1. part of URL
http://abc.xx/def/ghi/...

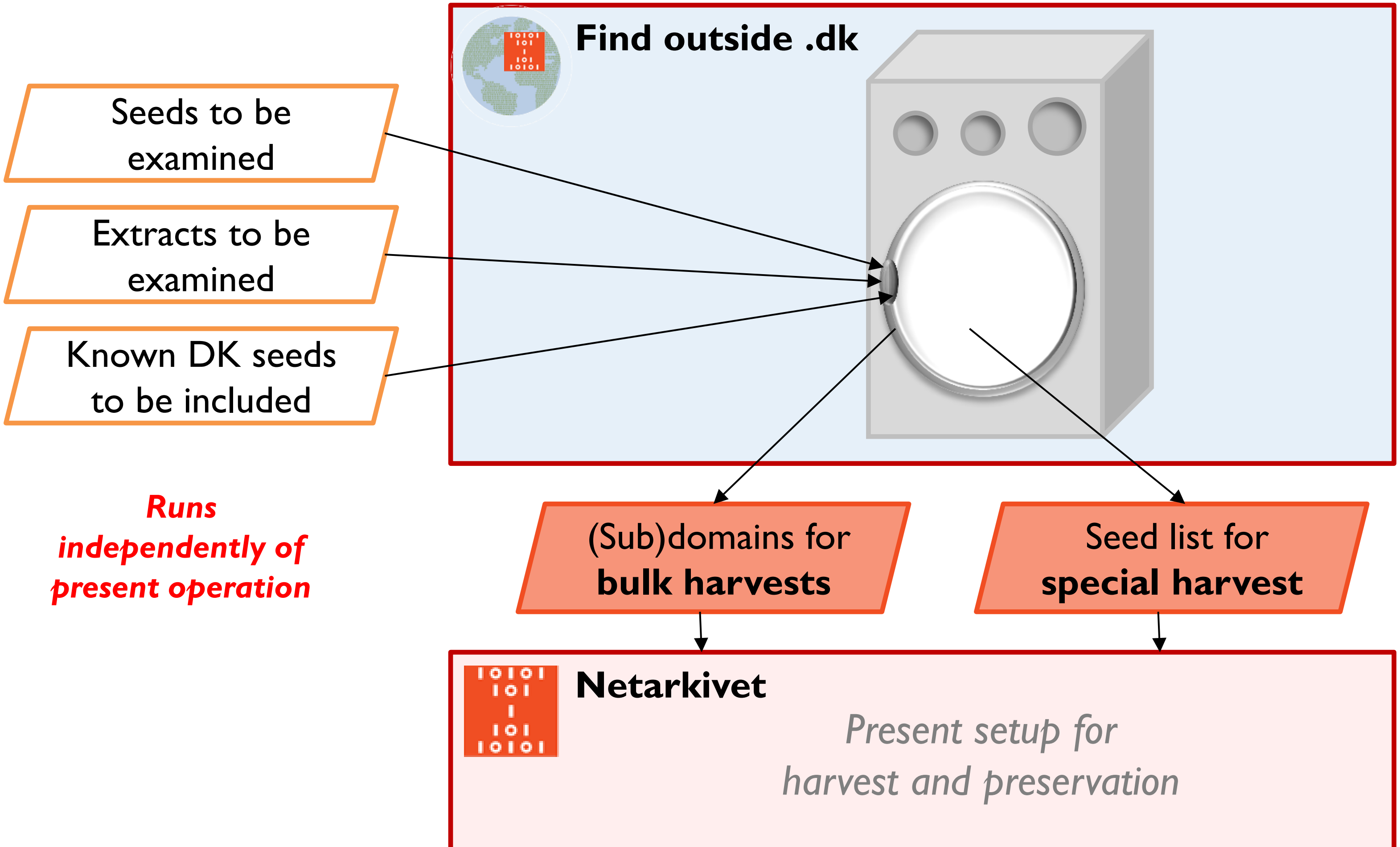
General implementation covering more methods

Implemention – what to do

How to implement?

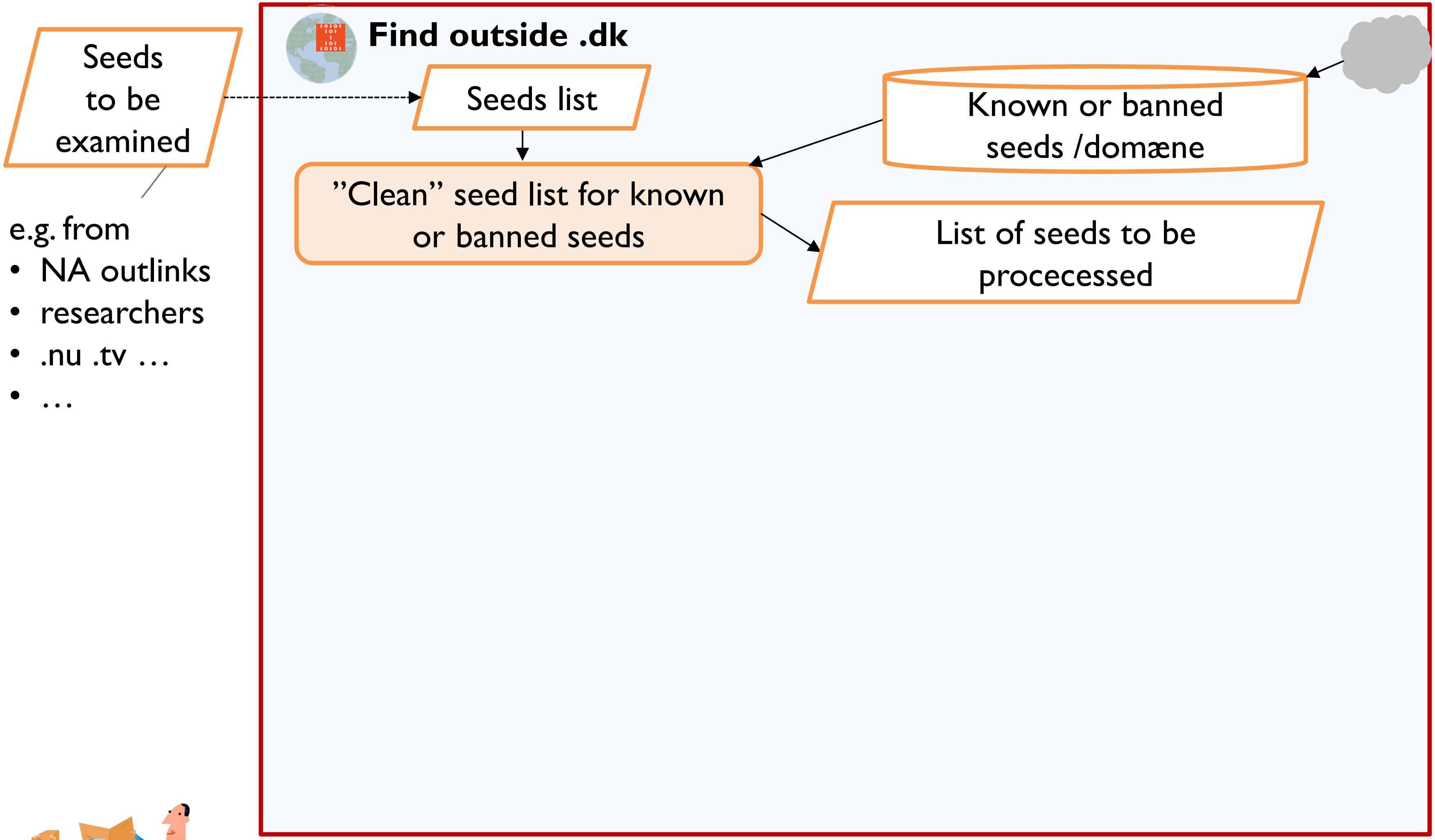
- Automate whenever possible
- Support web curators work
- Support several methods
  - Similar to NL (based on seed)
  - Similar to IA (based on existing extracts)
  - Known Danish outside .dk seeds

Implementation – the seed washer


Find outside .dk
Seeds to be examined
Extracts to be examined
Known DK seeds to be included
(Sub)domains for **bulk harvests**
Seed list for **special harvest**
Netarkivet
Present setup for harvest and preservation

Runs independently of present operation

Seed washer – Seeds method



Seeds to be examined

Find outside .dk

Seeds list

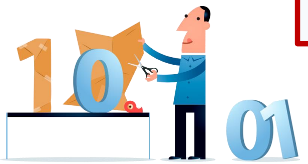
Known or banned seeds /domæne

”Clean” seed list for known or banned seeds

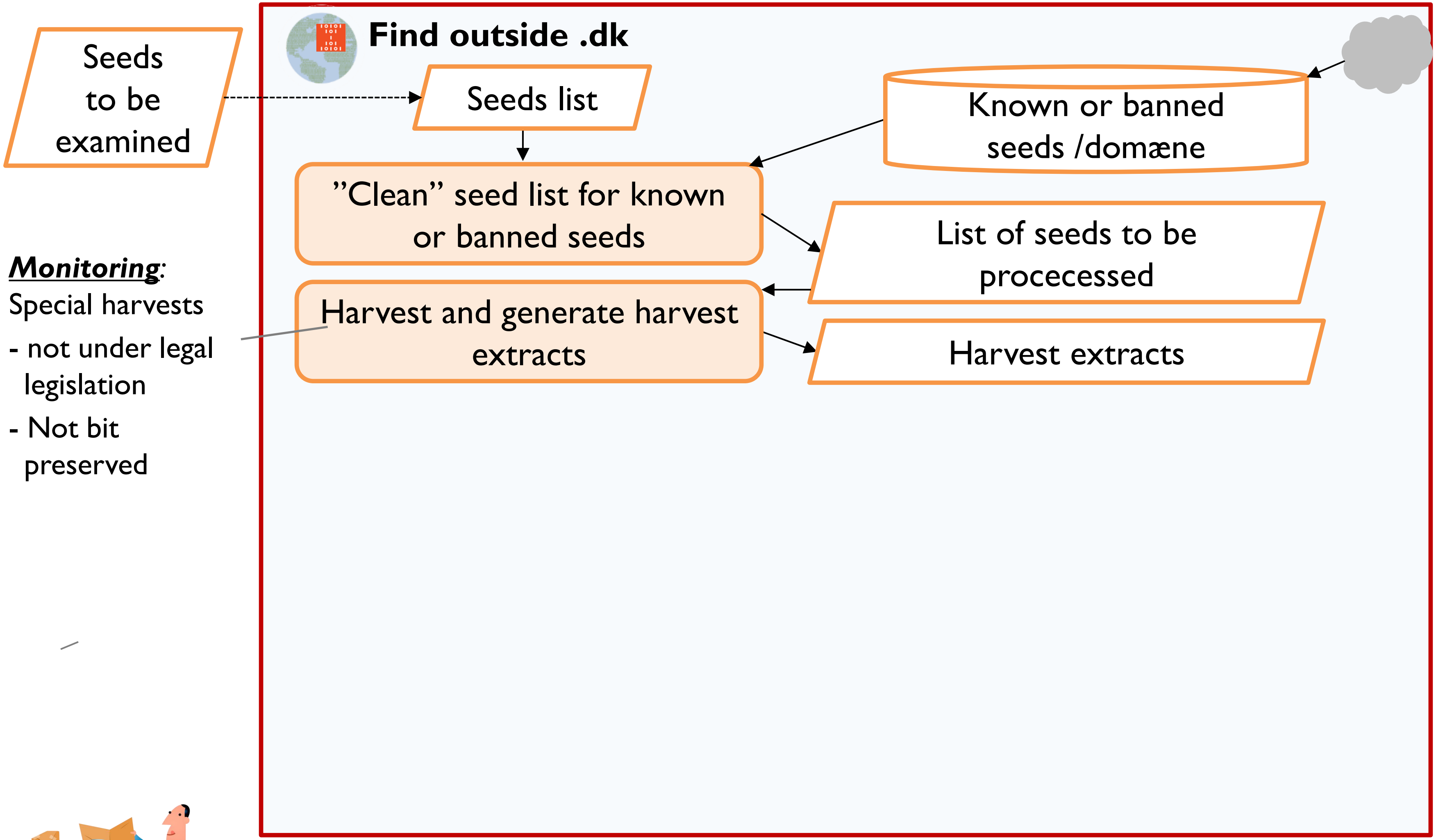
List of seeds to be procecessed

e.g. from

- NA outlinks
- researchers
- .nu .tv …
- …



Seed washer – Seeds method



Seeds to be examined

Find outside .dk

Seeds list

Known or banned seeds /domæne

”Clean” seed list for known or banned seeds

List of seeds to be procecessed

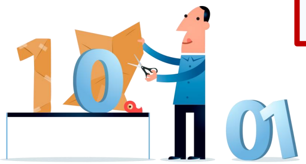
Harvest and generate harvest extracts

Harvest extracts

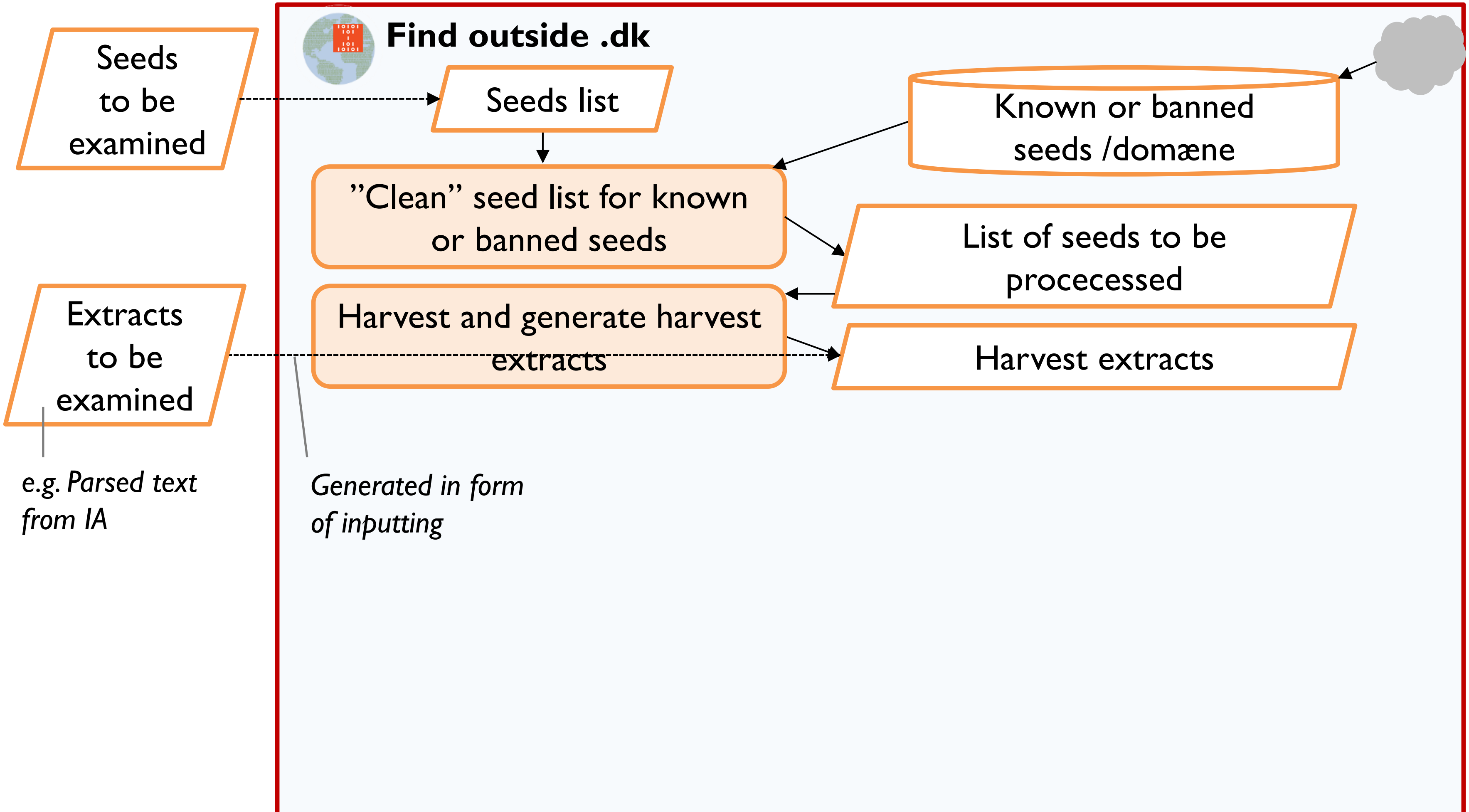
Monitoring:

Special harvests

- not under legal legislation
- Not bit preserved



Seed washer – Seeds method



Seeds to be examined

Find outside .dk

Seeds list

Known or banned seeds /domæne

”Clean” seed list for known or banned seeds

List of seeds to be procecessed

Harvest and generate harvest extracts

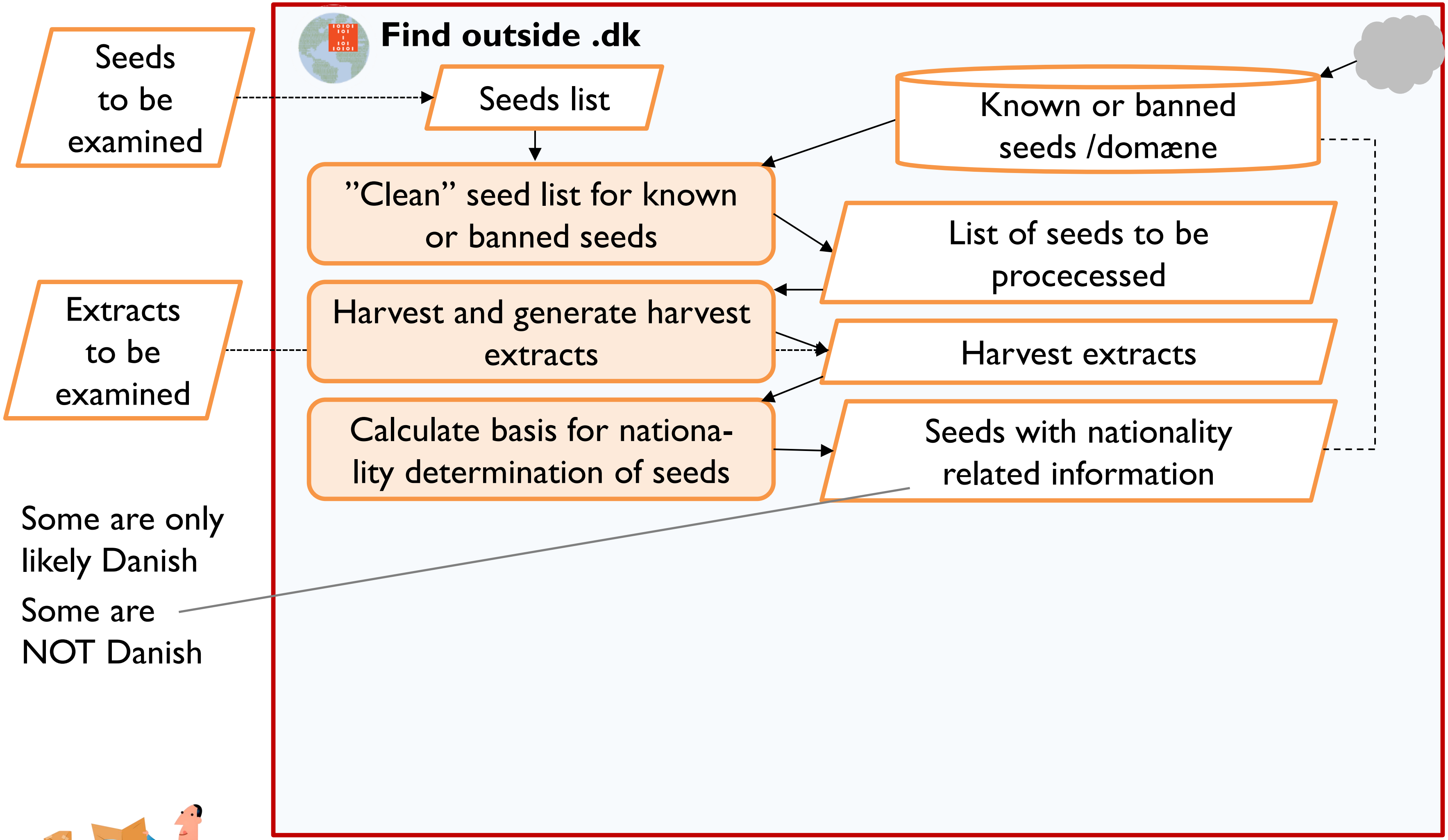
Harvest extracts

Extracts to be examined

e.g. Parsed text from IA

Generated in form of inputting

Seed washer – Seeds method



Find outside .dk

Seeds to be examined

Seeds list

Known or banned seeds /domæne

”Clean” seed list for known or banned seeds

List of seeds to be procecessed

Extracts to be examined

Harvest and generate harvest extracts

Harvest extracts

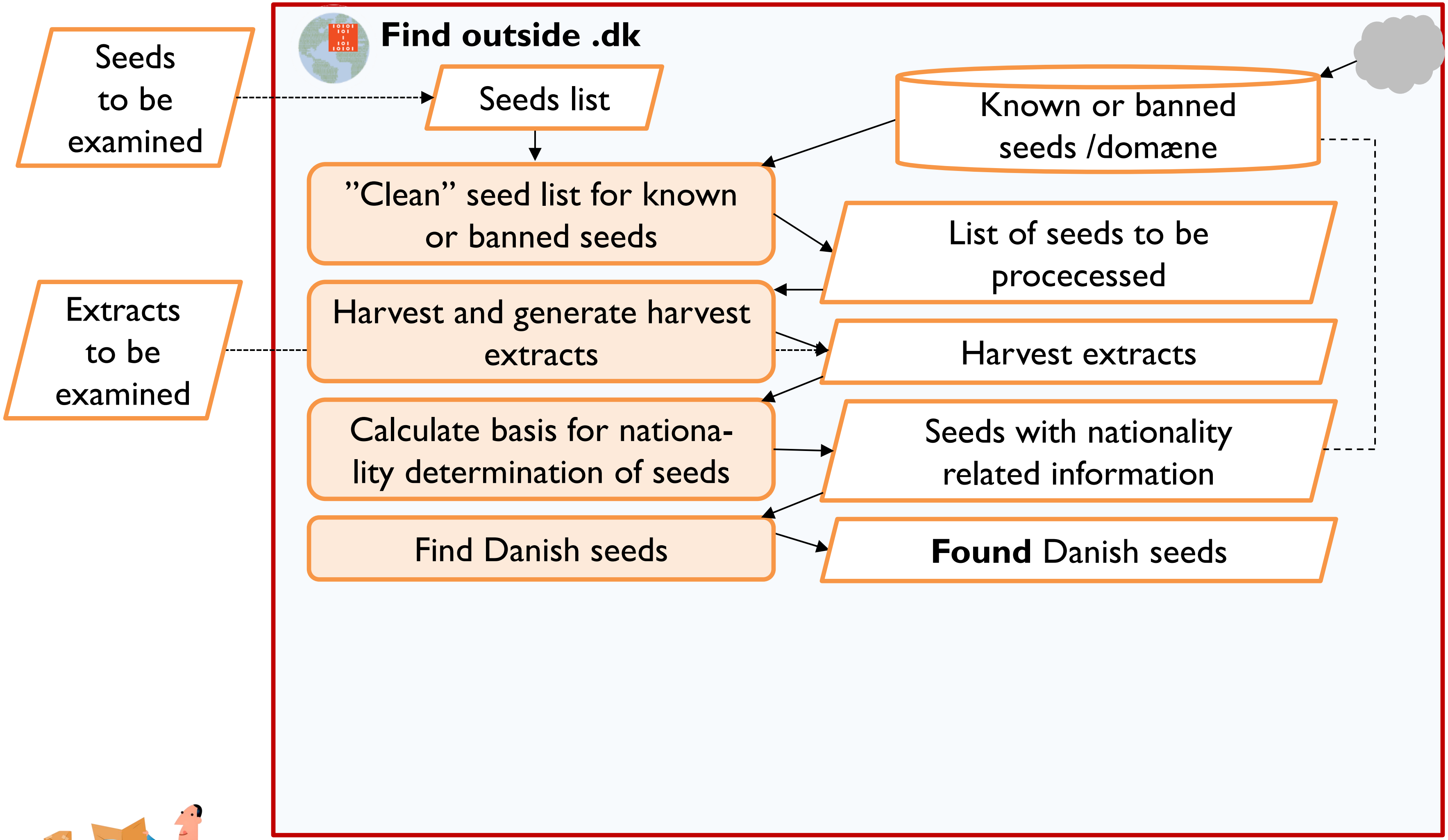
Calculate basis for nationa-lity determination of seeds

Seeds with nationality related information

Some are only likely Danish

Some are NOT Danish

Seed washer – Seeds method



Seeds to be examined

Extracts to be examined

Find outside .dk

Seeds list

Known or banned seeds /domæne

”Clean” seed list for known or banned seeds

List of seeds to be procecessed

Harvest and generate harvest extracts

Harvest extracts

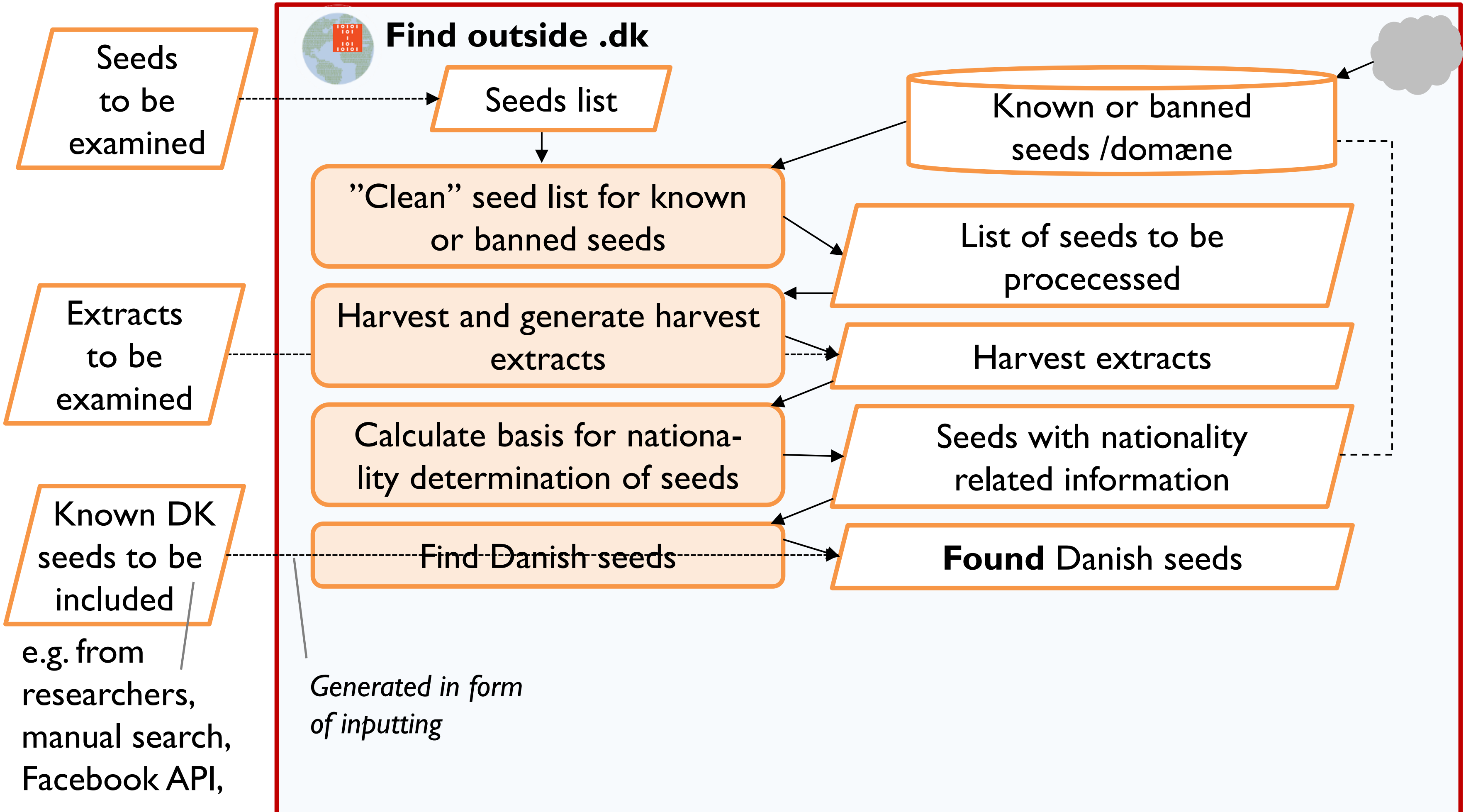
Calculate basis for nationa-lity determination of seeds

Seeds with nationality related information

Find Danish seeds

Found Danish seeds

Seed washer – Seeds method



Find outside .dk

Seeds to be examined

Seeds list

Known or banned seeds /domæne

”Clean” seed list for known or banned seeds

List of seeds to be procecessed

Harvest and generate harvest extracts

Extracts to be examined

Harvest extracts

Calculate basis for nationa-lity determination of seeds

Seeds with nationality related information

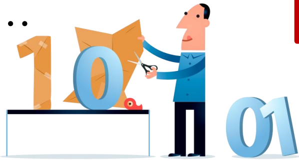
Find Danish seeds

Known DK seeds to be included

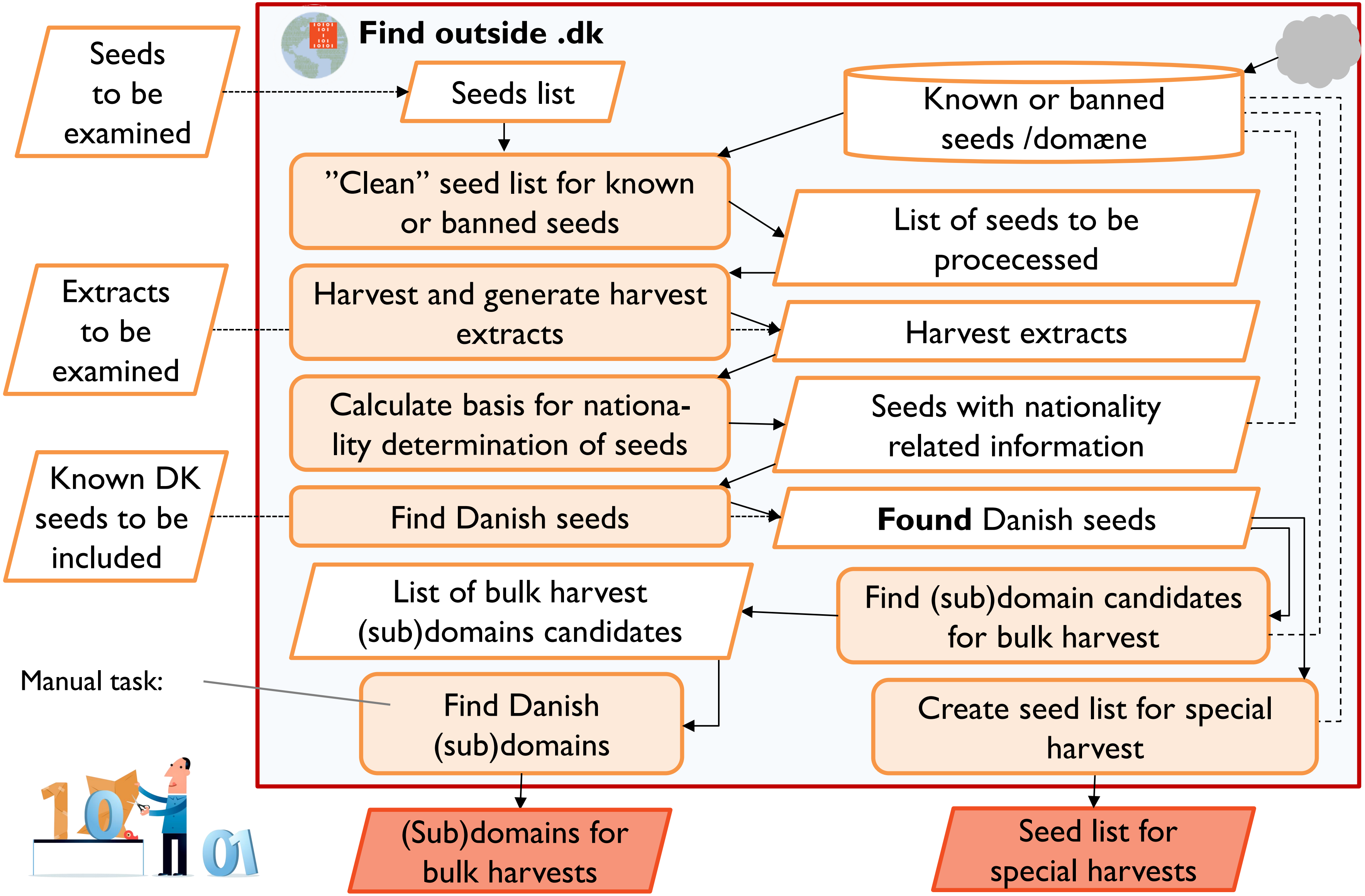
Found Danish seeds

e.g. from researchers, manual search, Facebook API, …

Generated in form of inputting



Seed washer – Seeds method



Seeds to be examined

Extracts to be examined

Known DK seeds to be included

Manual task:

Find outside .dk

Seeds list

Known or banned seeds /domæne

”Clean” seed list for known or banned seeds

List of seeds to be procecessed

Harvest and generate harvest extracts

Harvest extracts

Calculate basis for nationality determination of seeds

Seeds with nationality related information

Find Danish seeds

Found Danish seeds

List of bulk harvest (sub)domains candidates

Find (sub)domain candidates for bulk harvest

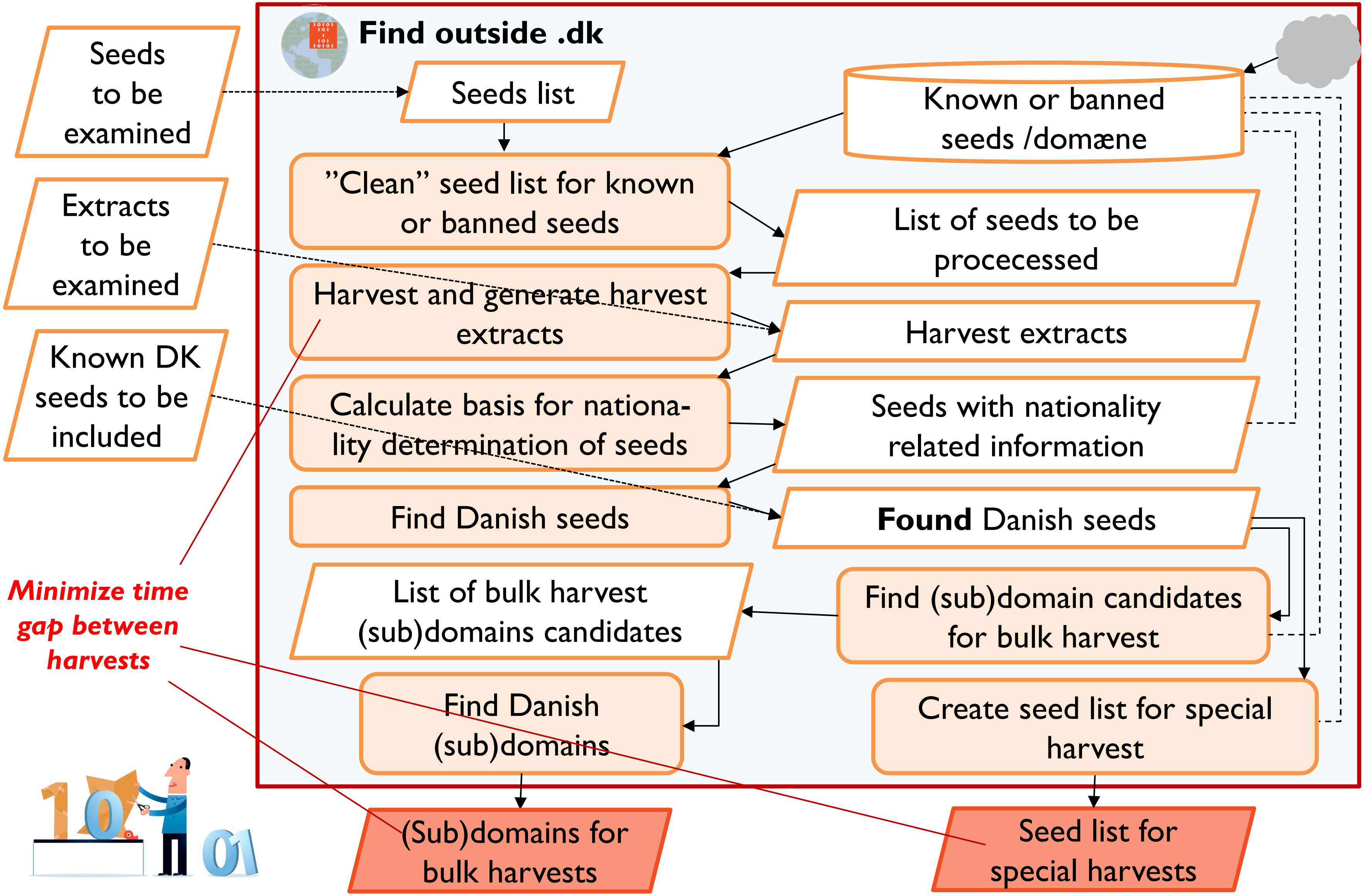
Find Danish (sub)domains

Create seed list for special harvest

(Sub)domains for bulk harvests

Seed list for special harvests

Seed washer – Seeds method



Seeds to be examined

Extracts to be examined

Known DK seeds to be included

Find outside .dk

Seeds list

Known or banned seeds /domæne

”Clean” seed list for known or banned seeds

List of seeds to be procecessed

Harvest and generate harvest extracts

Harvest extracts

Calculate basis for nationa-lity determination of seeds

Seeds with nationality related information

Find Danish seeds

Found Danish seeds

List of bulk harvest (sub)domains candidates

Find (sub)domain candidates for bulk harvest

Find Danish (sub)domains

Create seed list for special harvest

(Sub)domains for bulk harvests

Seed list for special harvests

Minimize time gap between harvests

Assumptions

- It is possible to minimize the harvest gap sufficiently
- There are no legal issues in harvesting outside the .dk top level domain
- It is acceptable
    - That we lose material
      The only included seeds, are the ones where there are about 90% probability of being relevant for Danish heritage
    - That we include some noise:
      About 10% of the included seeds are ***not*** relevant for Danish heritage

The way forward …

- Implementations in 2016 – just started

- Expects to evaluate and adjust after 2-3 years

Questions

Images of this style from digitalbevaring.dk