REPUBLIC OF ESTONIA
NATIONAL ARCHIVES

Case Study

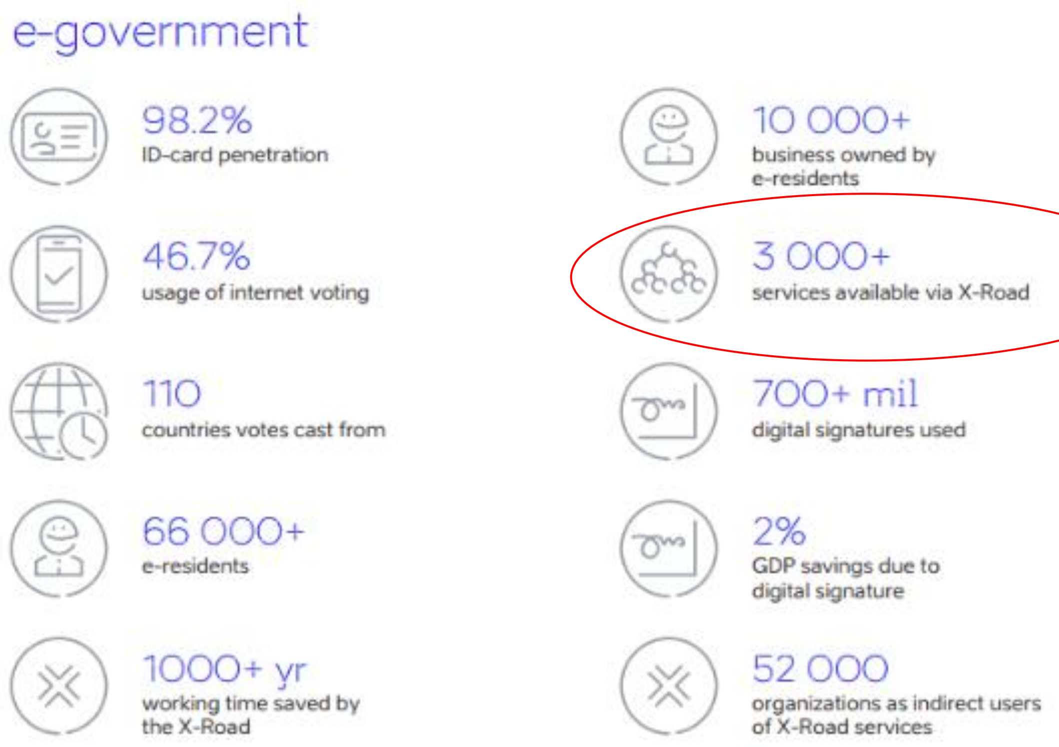
Database archiving at the National Archives of Estonia

Kuldar Aas
29 - 30 July 2020

Database preservation using the Database Preservation Toolkit and SIARD : A Practical Workshop

Republic of Estonia
National Archives

enter e-estonia



e-government

98.2%
ID-card penetration

10 000+
business owned by e-residents

46.7%
usage of internet voting

3 000+
services available via X-Road

110
countries votes cast from

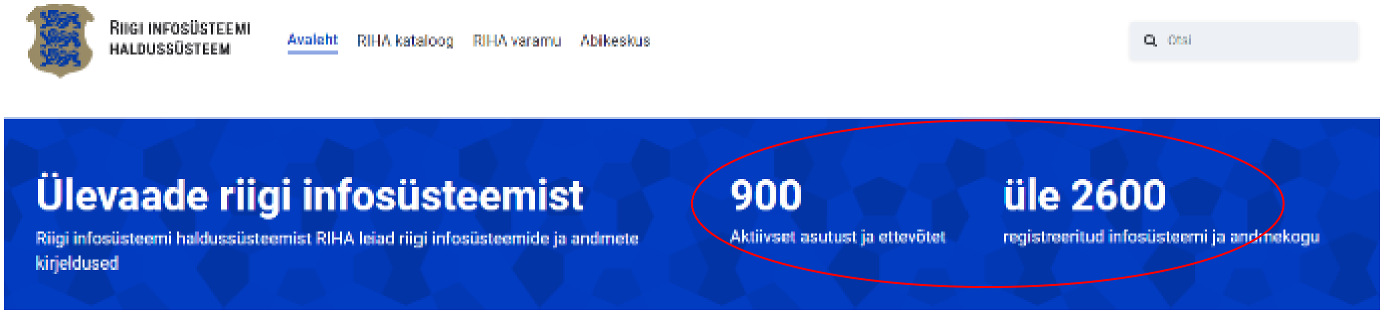
700+ mil
digital signatures used

66 000+
e-residents

2%
GDP savings due to digital signature

1000+ yr
working time saved by the X-Road

52 000
organizations as indirect users of X-Road services



Ülevaade riigi infosüsteemist

Riigi infosüsteemi haldussüsteemist RIHA leiad riigi infosüsteemide ja andmete kirjeldused

900
Aktiivset asutust ja ettevõtet

üle 2600
registreeritud infosüsteemi ja andmekogu

https://e-estonia.com/e-estonia-toolkit/

REPUBLIC OF ESTONIA
NATIONAL ARCHIVES

Questions to address

- Appraisal of 2600+ datasets and 3000+ services

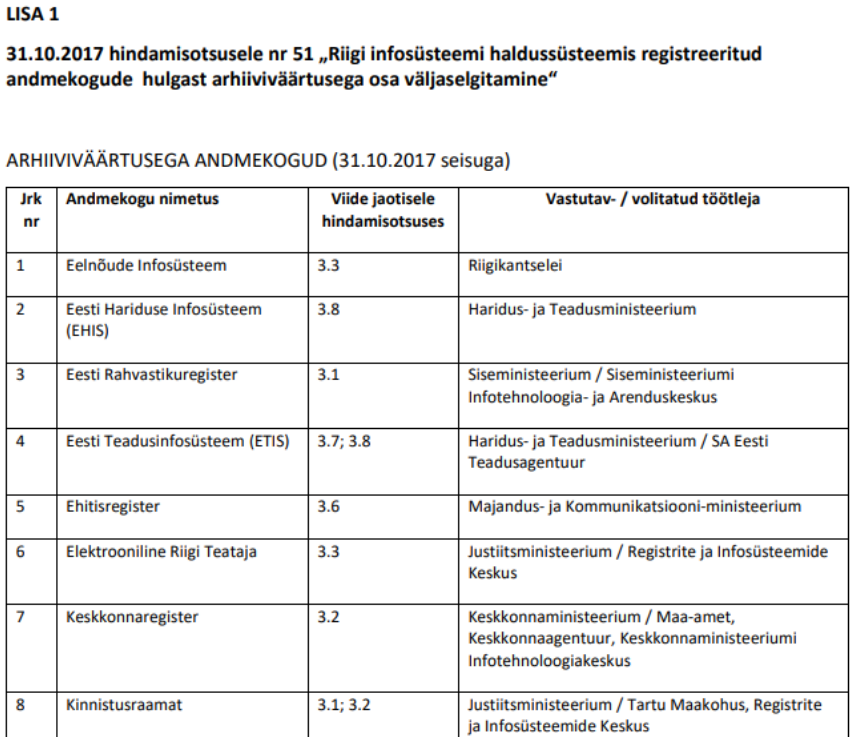
Case study: State Construction Dataset

- Managing the size and complexity of a relational database
- Pre-ingest process
- SIARD creation, archiving and reuse with DBPTK

Very Macro Appraisal

- ✓ Look at the descriptions of 2600 information systems
- ✓ Separate „datasets“ and „processing systems“
- ✓ Classify all datasets according to government functions
  - ✓ Remove the ones supporting non-valuable functions (f.ex. finance, staffing)
- ✓ Within a function analyse primary vs secondary data (i.e. redundancy) and the value of services being offered
  - ✓ Remove the ones which are only using secondary data and/or offer services where the data is not of archival value

Very Macro Appraisal



LISA 1

31.10.2017 hindamisotsusele nr 51 „Riigi infosüsteemi haldussüsteemis registreeritud andmekogude hulgast arhiiviväärtusega osa väljaselgitamine"

ARHIIVIVÄÄRTUSEGA ANDMEKOGUD (31.10.2017 seisuga)

Jrk nr	Andmekogu nimetus	Viide jaotisele hindamisotsuses	Vastutav- / volitatud töötleja
1	Eelnõude Infosüsteem	3.3	Riigikantselei
2	Eesti Hariduse Infosüsteem (EHIS)	3.8	Haridus- ja Teadusministeerium
3	Eesti Rahvastikuregister	3.1	Siseministeerium / Siseministeeriumi Infotehnoloogia- ja Arenduskeskus
4	Eesti Teadusinfosüsteem (ETIS)	3.7; 3.8	Haridus- ja Teadusministeerium / SA Eesti Teadusagentuur
5	Ehitisregister	3.6	Majandus- ja Kommunikatsiooni-ministeerium
6	Elektrooniline Riigi Teataja	3.3	Justiitsministeerium / Registrite ja Infosüsteemide Keskus
7	Keskkonnaregister	3.2	Keskkonnaministeerium / Maa-amet, Keskkonnaagentuur, Keskkonnaministeeriumi Infotehnoloogiakeskus
8	Kinnistusraamat	3.1; 3.2	Justiitsministeerium / Tartu Maakohus, Registrite ja Infosüsteemide Keskus

→ Two years of effort

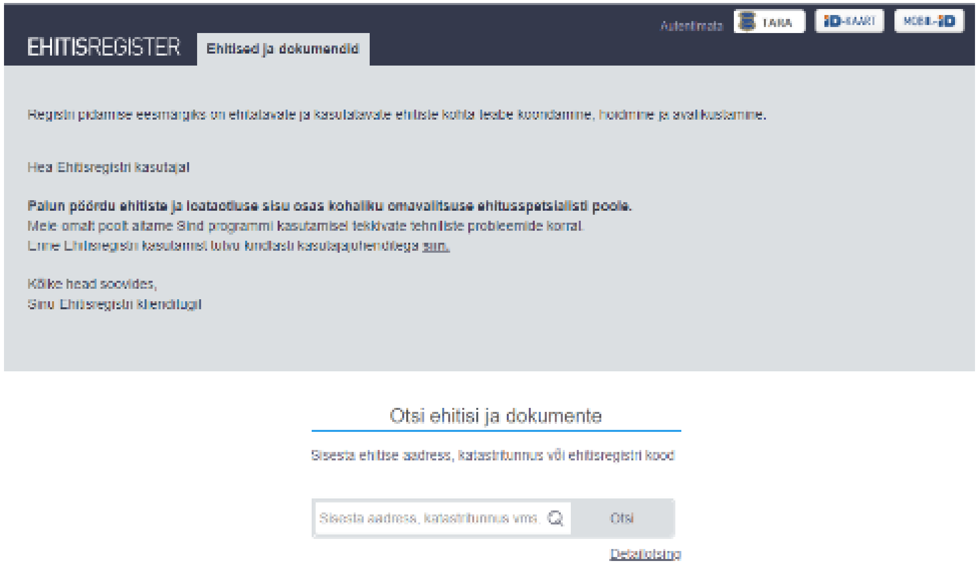
→ Preliminary list of 26 key datasets of „high value"

** List to be regularly updated*

*** Assumed total number of valuable datasets in the Estonian public sector **70 – 90***

**** Does not include scientific, statistical etc datasets*

State Construction Dataset



Registri pidamise eesmärgiks on ehitatavate ja kasutatavate ehitiste kohta teabe koondamine, hoidmine ja avalikustamine.

Hea Ehitisregistri kasutaja!

Palun pöördu ehitiste ja loataotluse sisu osas kohaliku omavalitsuse ehitusspetsialisti poole.
Meie omalt poolt aitame Sind programmi kasutamisel tekkivate tehniliste probleemide korral.
Enne Ehitisregistri kasutamist tutvu kindlasti kasutajajuhenditega siin.

Kõike head soovides,
Sinu Ehitisregistri klienditugi

Otsi ehitisi ja dokumente

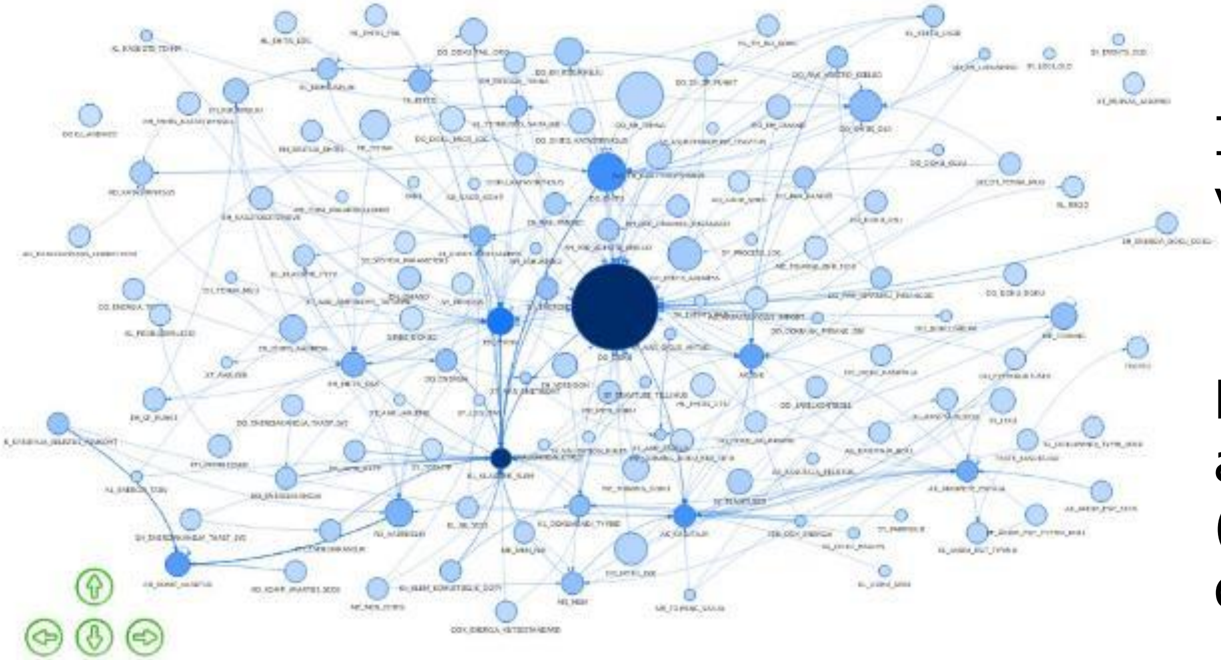
Sisesta ehitise aadress, katastritunnus või ehitisregistri kood

Sisesta aadress, katastritunnus vms.

- ✓ Core data on all buildings in Estonia
- ✓ Process of issuing permits
- ✓ List of companies certified for construction supervision, energy audits, ..
- ✓ Documentation: permits, models, supervision and audit reports

- ✓ Used by all municipalities
- ✓ Public / open interface for accessing core building data

Size and Complexity



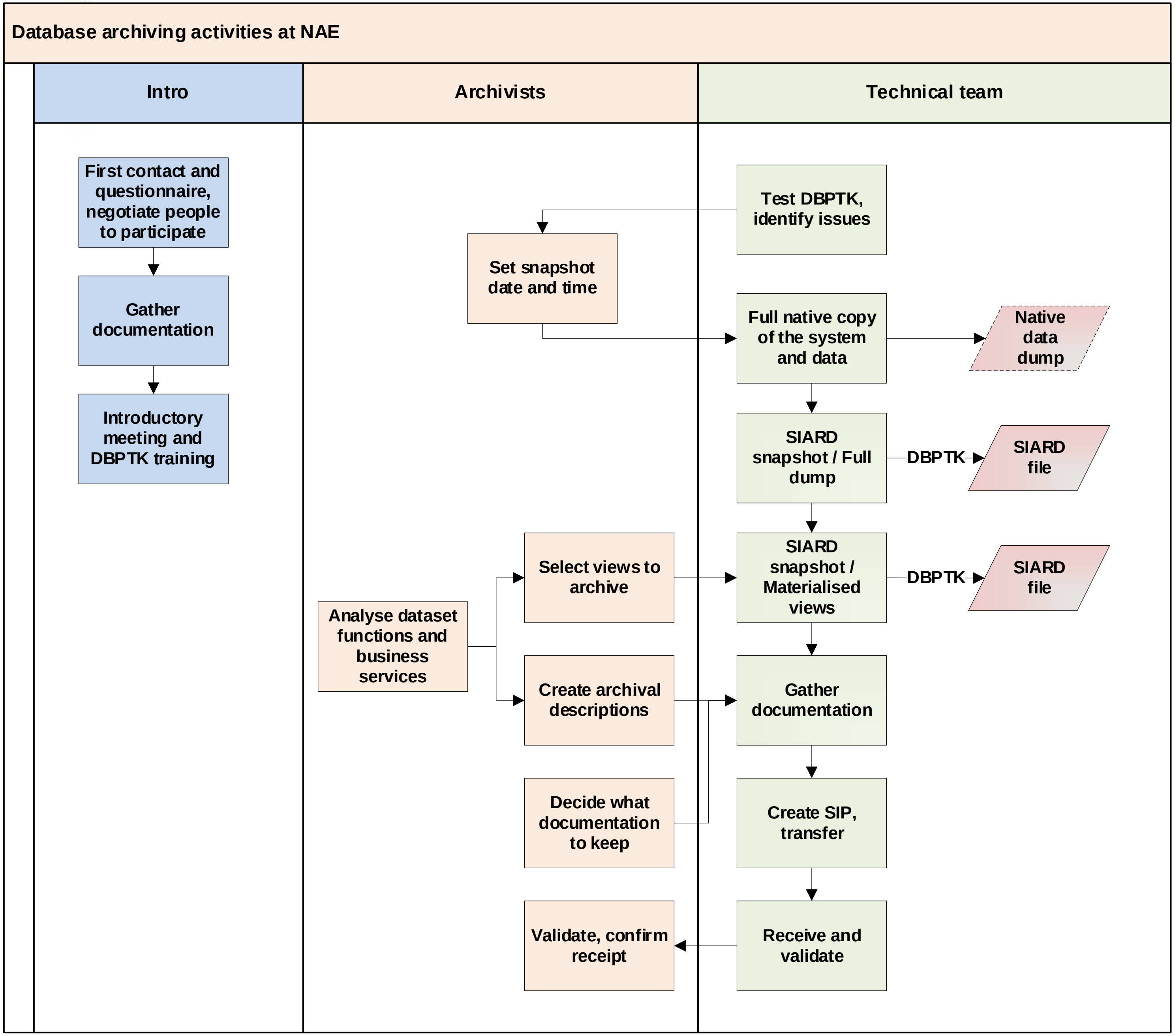
Is all of it really valuable?

How to present to archival users (technical skills, data protection)?

133 tables

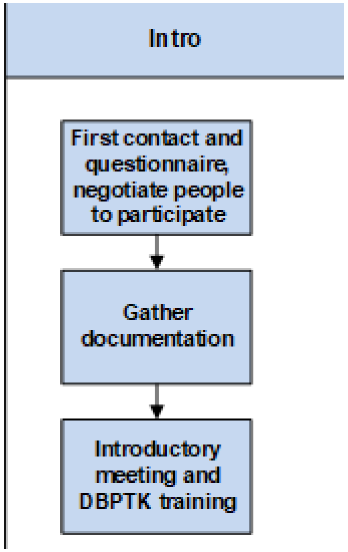
8.9 TB

P.S! Numbers as of „after technical cleaning"


Database archiving activities at NAE
Intro
Archivists
Technical team
First contact and questionnaire, negotiate people to participate
Gather documentation
Introductory meeting and DBPTK training
Test DBPTK, identify issues
Set snapshot date and time
Full native copy of the system and data
Native data dump
SIARD snapshot / Full dump
DBPTK
SIARD file
Analyse dataset functions and business services
Select views to archive
SIARD snapshot / Materialised views
DBPTK
SIARD file
Create archival descriptions
Gather documentation
Decide what documentation to keep
Create SIP, transfer
Validate, confirm receipt
Receive and validate

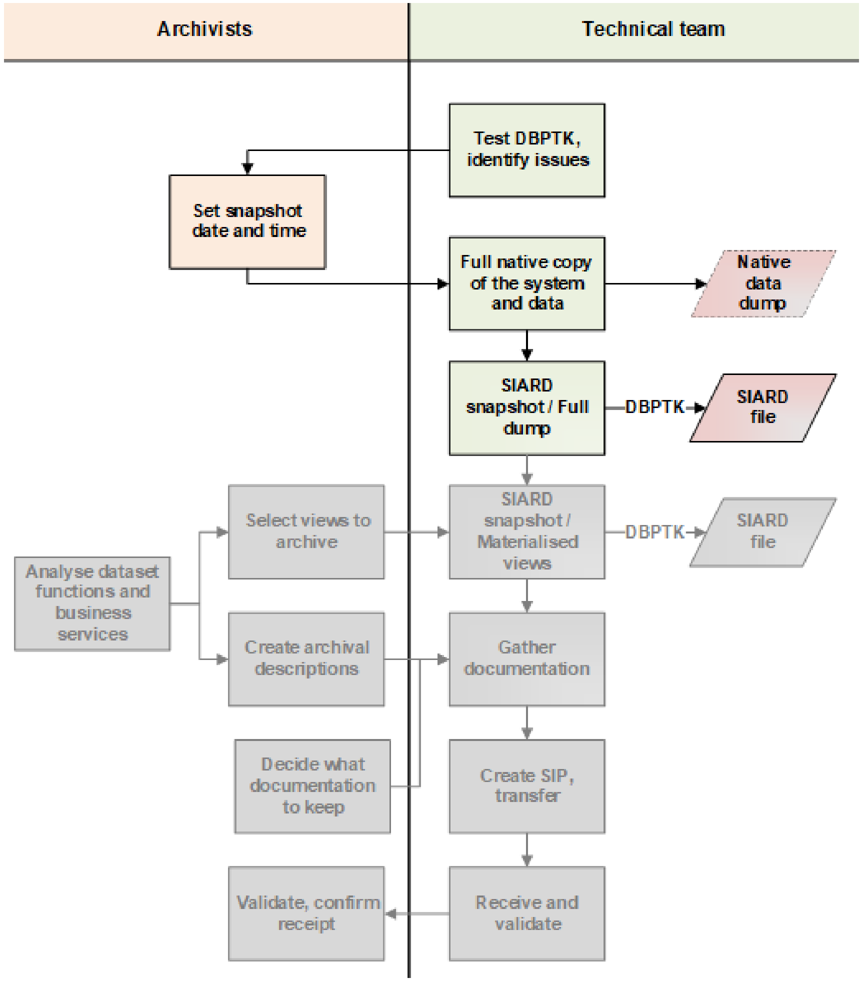
Republic of Estonia
National Archives

Process - intro


Intro
First contact and questionnaire, negotiate people to participate
Gather documentation
Introductory meeting and DBPTK training

- ✓ First contact (phone, e-mail)
- ✓ Questionnaire
  - **Who has to be involved (business owners, archivists, DB admins, developers, hosts)**
  - **Availability and timing**
  - Key technical details (DBMS, size, number of tables)
  - „Known issues“ (LOBs, geodata, queries in SQL vs app layer)
- ✓ Availability of documentation: data model, data descriptions, architecture, service descriptions, user guides, etc.
- ✓ Introductory meeting: discuss all details, determine next steps, explain the process and DBPTK

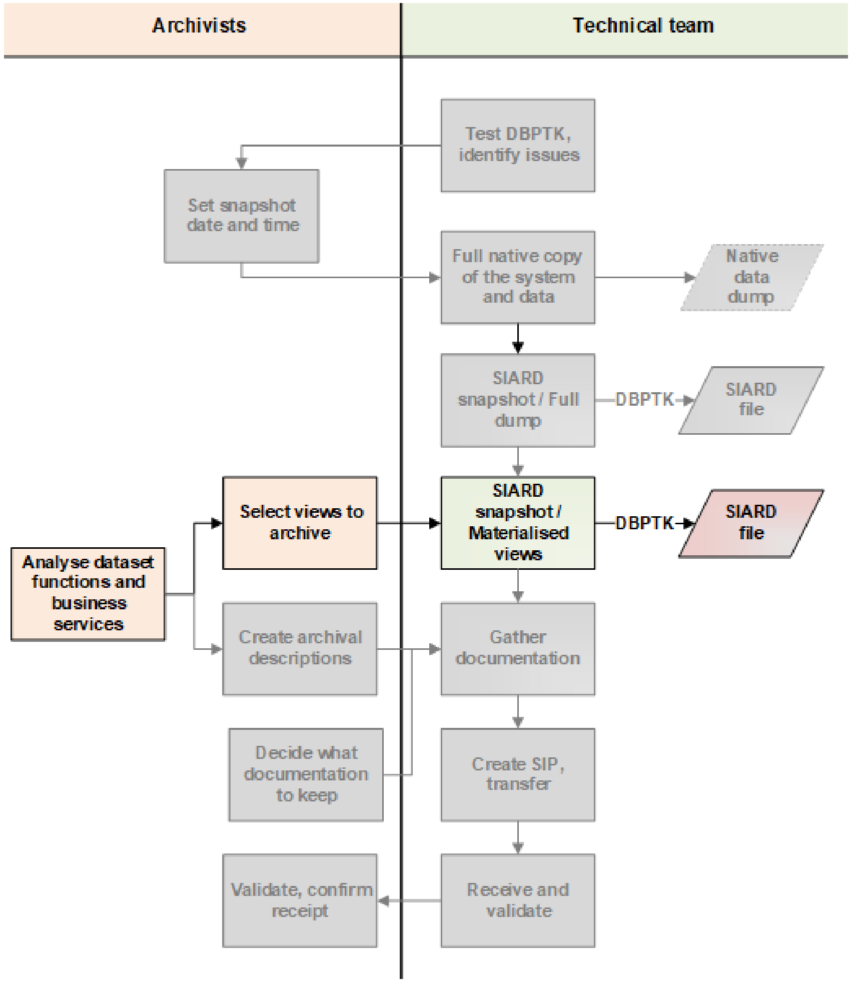
Process – full dump


Archivists
Technical team
Test DBPTK, identify issues
Set snapshot date and time
Full native copy of the system and data
Native data dump
SIARD snapshot / Full dump
DBPTK
SIARD file
Select views to archive
SIARD snapshot / Materialised views
DBPTK
SIARD file
Analyse dataset functions and business services
Create archival descriptions
Gather documentation
Decide what documentation to keep
Create SIP, transfer
Validate, confirm receipt
Receive and validate

Lessons learned

- ✓ Run DBPTK as soon as possible!!!
  - Helps to assess resources (disk space, servers, time)
  - Helps to evaluate errors (connecting to the database, external LOBs, geo-data)
  - Not worth speculating on paper if the tool can be executed in 10 minutes..
- ✓ Create a full copy of the system
  - SIARD creation and validation can take a lot of time (read: weeks)
  - Turn off functions and procedures,
  - Data remains unchanged throughout the rest of the process!

Process – materialised views

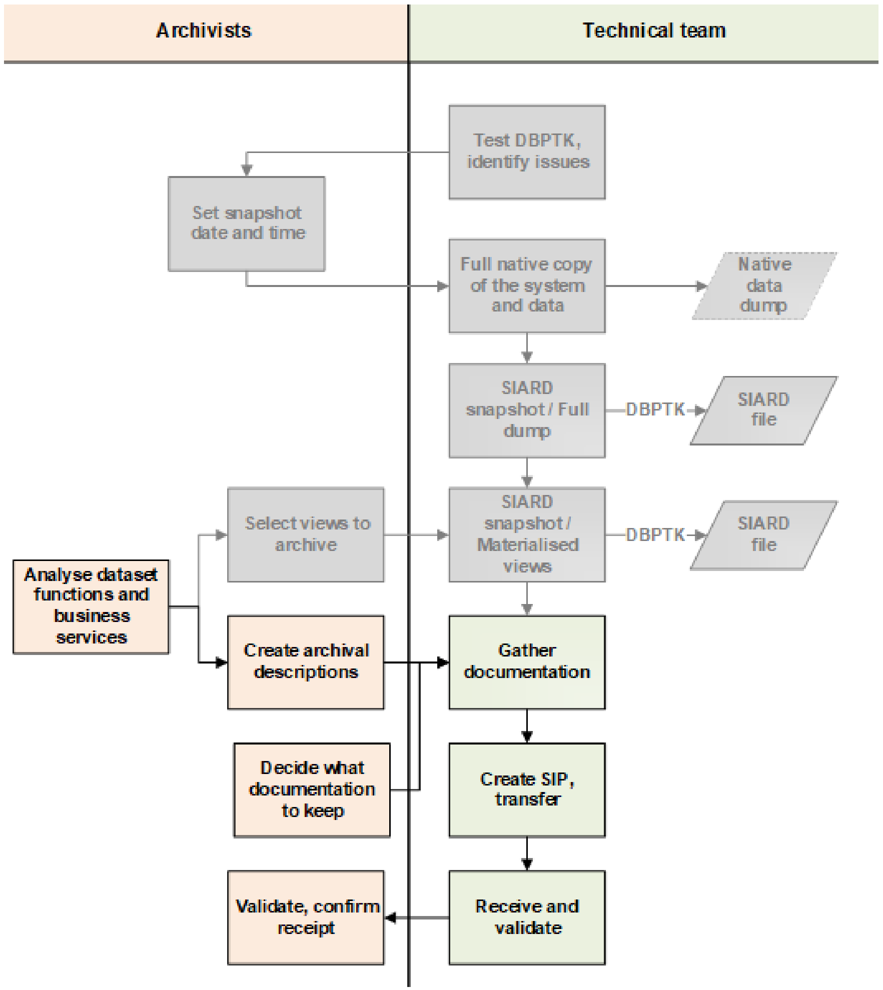

Archivists
Technical team
Test DBPTK, identify issues
Set snapshot date and time
Full native copy of the system and data
Native data dump
SIARD snapshot / Full dump
DBPTK
SIARD file
Select views to archive
SIARD snapshot / Materialised views
DBPTK
SIARD file
Analyse dataset functions and business services
Create archival descriptions
Gather documentation
Decide what documentation to keep
Create SIP, transfer
Validate, confirm receipt
Receive and validate

Motivation: simplified „single table“ representation of data for simple users

- ✓ More than 200 views already available
- ✓ 84 views after removing technical system views
- ✓ Ask owner to describe all remaining views
- ✓ Archivists decide which views to materialise
  - Connection to business function and activity (records series, service)
  - Usage statistics
  - Data protection
- ✓ 13 views selected for materialisation

REPUBLIC OF ESTONIA
NATIONAL ARCHIVES

Process – finalisation


Archivists
Technical team
Test DBPTK, identify issues
Set snapshot date and time
Full native copy of the system and data
Native data dump
SIARD snapshot / Full dump
DBPTK
SIARD file
Select views to archive
SIARD snapshot / Materialised views
DBPTK
SIARD file
Analyse dataset functions and business services
Create archival descriptions
Gather documentation
Decide what documentation to keep
Create SIP, transfer
Validate, confirm receipt
Receive and validate

Lessons learned

- ✓ Transfer of 12 TB data can take a lot of time… consider packaging with tar, zip, ..
- ✓ Technical SIARD validation with DBPTK to be done at agency!!
- ✓ Technical documentation in bespoke formats (.eap)
- ✓ Creation of screencasts / videos of the original GUI (data input, services, queries)

Timing and resources

- Eight people involved
  - 2 NAE archivists
  - **NAE technical expert**
  - NAE project supervisor
  - Agency system owner
  - **Agency database administrator**
  - two technical experts at hosting company

+ technical support from KEEP Solutions

Timing and resources

- Whole process six calendar months
- Technical tasks
  - Technical testing, determining DBPTK configuration: 3 months
  - Setup of dedicated archiving infrastructure: 2 weeks
  - Creation of SIARD snapshots (three tries): 1 month
  - Copying and validation: 1.5 months (during Christmas..)
- View selection: 2 weeks
- Archival description, documentation selection: 2 weeks

What’s next

- Happy with the performance of DBPTK
- SIP for SIARD – work going on in E-ARK
- Checksum / manifest creation before transfer
- Set up DBPTK Enterprise as public access portal
    - Pre-load unrestricted materialised views
- Maintain and grow a list of prerequisites for database archiving
    - Try to influence national IT guidelines to implement relevant ones
- Archiving only materialised views where relevant
    - In 5 – 10 years
    - Requires good data governance to be in place!