WELCOME.
IPS0313
Metadata-driven Data Gate
at Statistics Estonia

Mrs Kaia Kulla
Statistics Estonia
Thursday 20 July, 10:00AM – 12:00AM
Outline/Content

1. Current data capture methods
2. Goal - minimizing manual effort in receiving data
3. Metadata for Data Gate, but not only
4. DDI comes in
5. Benefits
6. Use case
Current data capture methods
Current Data Capture Methods

Automated
- xml
- Push
- x-Road over public internet
- Pull
- X-Gate
- AHM

Manual
- csv
- txt
- xls(x)
- Push
- public internet
- email
- ftp
- SharePoint
- NextCloud
Goal - minimizing manual effort in receiving data
New solution – Data Gate

1 – request metadata
2 – request metadata (translated)
3 – send metadata
4 – send metadata (translated)
5 – upload data-file
6 – store data-file in repository
7 – trigger workflow
8 – trigger workflow (translated)
9 – download data-file from repository
10 – report file validation progress
11 – store data in data warehouse
Metadata for Data Gate, but not only
9 Data capture steps to consider

Data Files

Agreement

Raw Data

Initial Observation Register

Processing

Data Storage Area

Cleaned and processed data

Final Observation Register

Data Gate

COLECTICA

HULFT
Metadata for Data Gate – what we need?

- Obligations -> originate from agreement
  - Information about data provider, incl. information about administrative register
  - Datasets
    - Variables, types, descriptions
    - Code Lists and Classifications
    - Description of dataset (files: type, separator, )
  - Data submitting date(s), frequency
  - Data collection dates
  - Periods covered by the data
- But also, information about
  - Processing
  - Storage
  - Data Lineage
  - ???
Users to consider

• Several applications
  – Denodo
  – Data editing application
  – Data collection applications
  – Environment for scientists (in development)
  – Etc.

• Programs (R-language, Oracle)

• Physical users
  – Analysts (inside)
  – Scientists (outside)
  – Students (outside)
  – Etc.
DDI comes in
Where to start?

- Collect existing information from everywhere (existing data, Confluence, Excel files, emails, Word documents) in one place -> Colectica
- What is missing, what would simplify everyday work?
- Use of metadata as much as possible in the data capture, data storage and data processing processes
- Discover DDI model and features of Colectica
  - Models behind Colectica and our current metadata system are different
  - DDI is much more detailed compared to our current model (Statistical activities, Variables, Code Lists, for processing: Data Sets, Tables, Variables)
  - Lack of attributes or nests for information that we already have
  - DDI Controlled Vocabularies don’t match always with our need
Result of the analysis - per DDI item

- DDI Items involved:
  - Series -> Data Registry or Information System
    - Subseries -> Sub-registry, if that exists
  - Study -> kind of umbrella for data capture
  - Data Files (Data Layout) -> Dataset
    - Records -> Tables within the Dataset
  - Organization -> Owner of the Data Registry
  - Other Material -> Agreement

- Meaning remains the same as a meaning of DDI item
  - Data Collection
  - Variables
  - Code Lists and Categories
  - Classifications and Classification Items
Result of the analysis - Data model, draft version
Benefits
Simplified process and counterparts

Data Provider
Data Sets
Frequency and
observation period, etc.

Data Provider

Obligation(s)

Nextcloud

Data Gate

WSO2

DDI, Colectica

JSON

Hulft

IOR

Check(s)

Error?

CANCELLED

FOR

OK

FOR

OK?
Benefits, what we (will) get

For data capture purposes, in Data Gateway
- Automatic processes for data capture (at least from SE side)
- All information about agreements, metadata based on these agreements and obligations are in one place
- Calendar (schedule) for data capture – until now this was lacking
- Monitoring panel – which data in which state, status etc.
- Validation of captured data according to the agreement

Also, ...
- All metadata, that we have and located everywhere or that we missed, are stored now or will be stored in one place
- Reuse of metadata
- Overview of existing administrative registry data
- Implementing DDI stimulates the thinking of what else could be possible to document

- In the future:
  - Automatization of pseudonymization – document in metadata, what data should pseudonymize
Use case
All obligations are listed with submission dates and statuses.
Data Provider Portal

- Overview on one obligation
Data Provider Portal

- Submitting agreed data file
Data Provider Portal

Submitted data – data has arrived

<table>
<thead>
<tr>
<th>Date</th>
<th>Institution</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 01.2024</td>
<td>AITA</td>
<td>Data of international automatic exchange of TAX information application (AITA) (sub reger of Register of Taxable Persons) For your information, table FOR_50369.AITA_20210224 contains unversioned AITA data, years 2014-2020.</td>
</tr>
</tbody>
</table>

**MESSAGES**
Data File is submitted, but it is not final yet.

Institution: Estonian Tax and Customs Board
# SE portal – data arrival monitoring

## Monitorning

### Today's load

- **Tähtaeg**: 4 t
- **50%** 2 tk tekemud
- **50%** 2 tk tana saabumis
- **0%** 0 tk tekemata
- **0%** 0 tk tagasi liikatud

### Overview of received files

- **Ouskord**: 32 tk
- **19%** 6 tk tois
- **81%** 25 tk saabunud
- **0%** 0 tk tagasi liikatud
- **0%** 0 tk korras
- **0%** 0 tk vigased

### Last month's receipts statistics

- **Laskumised**: 14%
- **100%** 14 tk otseregistreeritud
- **0%** 0 tk hõlmemud

---

### Kohustuse ID  | Kohustuse/sisendfail | Andmedlik | Estiga  | Tähtaeg  | Olek
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>240</td>
<td>AITA</td>
<td>Data of international automatic exchange of TAXI information application (AITA)</td>
<td>ONE TESTNUMBER</td>
<td>10.01.2024 00:00</td>
<td>SASKUMUS</td>
</tr>
<tr>
<td>240</td>
<td>AITA</td>
<td>Data of international automatic exchange of TAXI information application (AITA)</td>
<td>ONE TESTNUMBER</td>
<td>10.01.2024 00:00</td>
<td>SASKUMUS</td>
</tr>
<tr>
<td>240</td>
<td>AITA</td>
<td>Data of international automatic exchange of TAXI information application (AITA)</td>
<td>ONE TESTNUMBER</td>
<td>10.07.2024 00:00</td>
<td>SASKUMUS</td>
</tr>
<tr>
<td>240</td>
<td>AITA</td>
<td>Data of international automatic exchange of TAXI information application (AITA)</td>
<td>ONE TESTNUMBER</td>
<td>10.07.2024 00:00</td>
<td>SASKUMUS</td>
</tr>
</tbody>
</table>
SE portal – detailed overview of arrivals
## New obligation

### ORGANIZATION AND DATASET

**Organization**
- Vali või sisesta uus

**Dataset**
- Vali või sisesta uus

### SUBMISSION

**Time series**
- Vali aegline kohtus

**From**
- Palun vali koostööpea

**To**
- Palun vali koostööpea

### PERSONS

**Client manager**
- Mõõdja liendiassistent

**Content manager**
- Mõõdja arstiahelk

### FILE STRUCTURE

**File format**
- txt, csv, xls(x), xml, json

**Column separator**
- Punkt, komma või

**String attribute**
- Punkt, komma või

**Decimal separator**
- Punkt, komma või

<table>
<thead>
<tr>
<th>Nr</th>
<th>Variable name</th>
<th>Description</th>
<th>Data type</th>
<th>Data subtype</th>
<th>Pattern</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Kõrduja nimetus</td>
<td>Kõrduja nimetus</td>
<td>Formaat</td>
<td>Alamformaat</td>
<td>Formaad kõrduja</td>
<td>Pikkus</td>
</tr>
</tbody>
</table>
THANK YOU.