Possibilities of using SDMX 3.1, DDI-LS 3.3 and DDI-CDI 1.0 in Estonian data portal based on DCAT-AP 3.0 to describe structural and reference metadata of datasets

Veiko Berendsen
Veiko.Berendsen@stat.ee
leading data governance expert

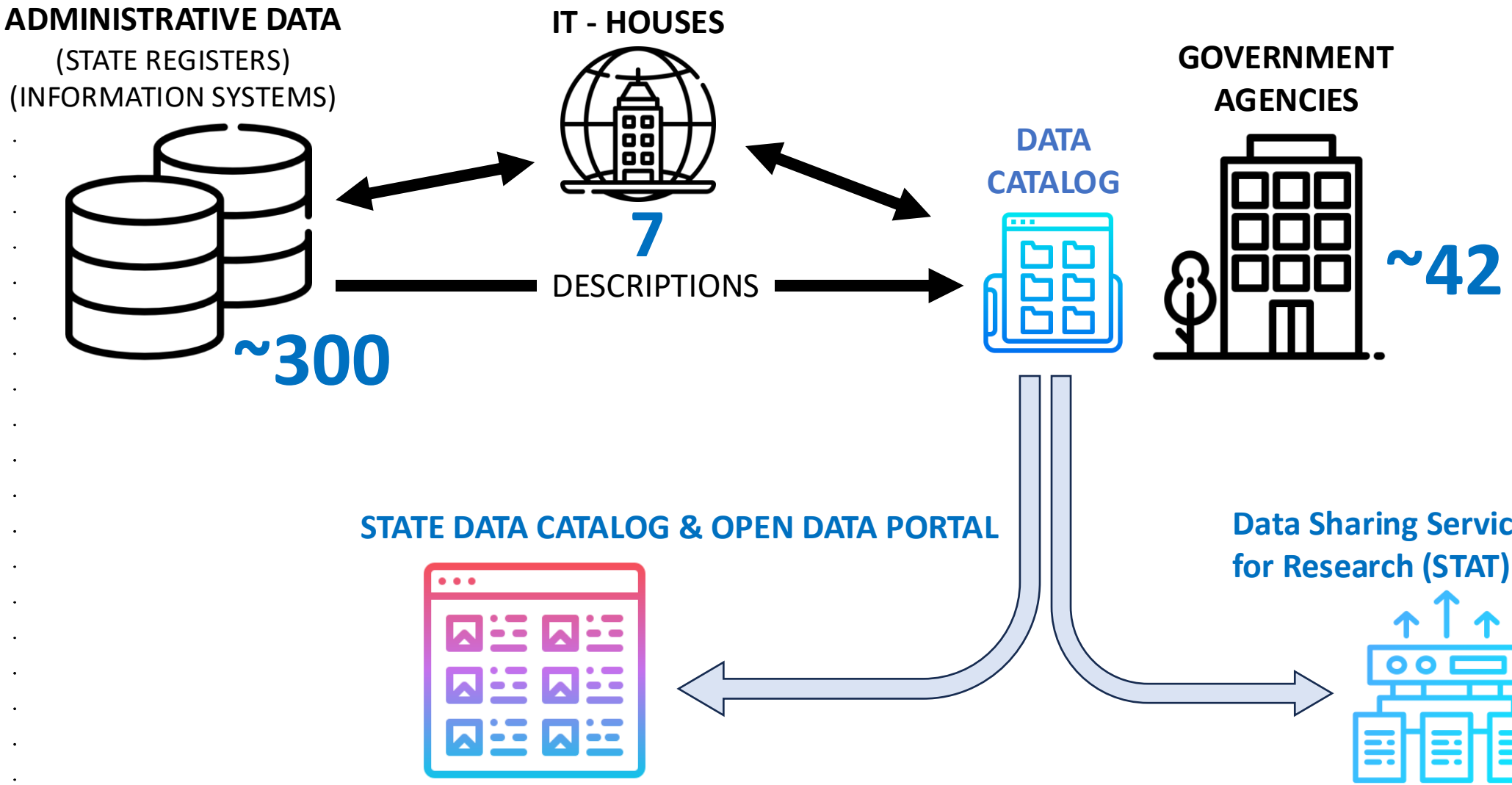
EESTI
STATISTIKA

Possibilities of using SDMX 3.1, DDI-LS 3.3 and DDI-CDI 1.0 in Estonian data portal based on DCAT-AP 3.0 to describe structural and reference metadata of datasets

1. Estonian data portal (ATV) has been in use for more than 15 years to describe administrative data (registers). Semantics have been improved in resent years concerning creation of data dictionaries and business vocabularies. Data service for metadata publishing from organization's data catalogues has been developed via API. Public sector organisations have legal obligation to publish their dataset's metadata and in the case of open data also distribute data in ATV.
2. For secondary use of public sector data, under Data Governance Act (DGA) and national legislation for research, innovation, policy-making, and regulatory activities an infrastructure and maintenance framework has been developed using privacy-enhancing technologies (PETs) for secure and trustworthy reuse of individual data. Data catalogues from both organizational and national level are created to find, understand, and reuse data (FAIR principle).
3. End users have need to understand data in the datasets on variable level. Neither DCAT-AP 3.0 nor most of application profiles - like MobilityDCAT or HealthDCAT - does not have coherent variables and codelists description part in their models. StatDCAT have variable level description possibility - taking semantics from SDMX - but it is not widely used.
4. Consensus about variable level description possible in metadata standards in use by statistical community (SDMX, DDI, DDI-CDI, GSIM) implementation into DCAT-AP and its thematic profiles is needed. Does SDMX can serve as model for it and is it better to have one variable level description model or might many models with semantic mappings be used?

EESTI STATISTIKA

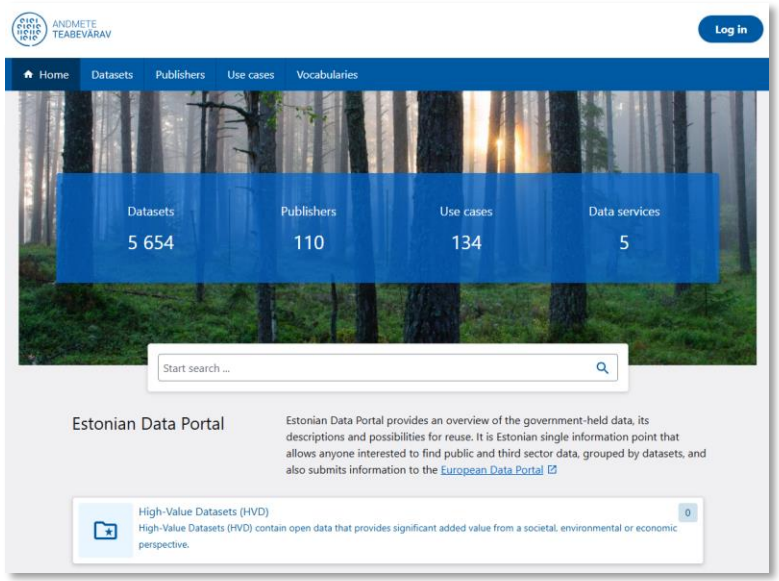
Business probleem. USE OF DATA: (1) secondary use for administrative activities, (2) secondary use for analysis / statistics



EESTI STATISTIKA

Data Portals / Data Catalogues interoperability

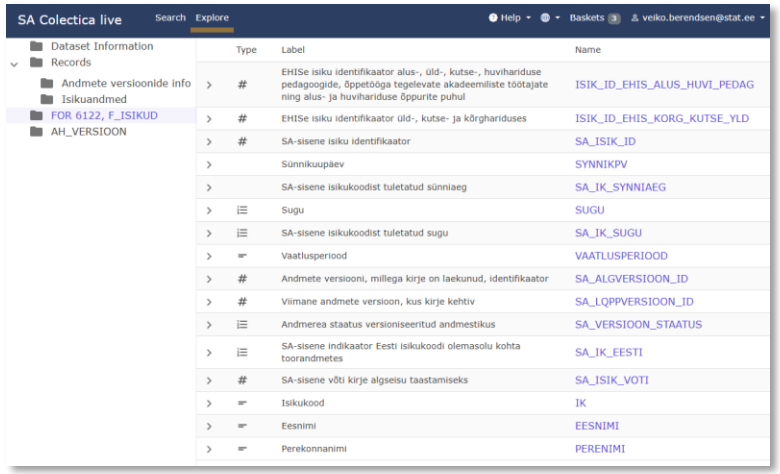
1



Open Data Portal + State Data Catalog (registers description)
andmed.eesti.ee
data set level description
DCAT-AP 3.0 + DCAT-AP-ET
+
Data dictionary & Business Vocabularies (for registers)
+
HealthDCAT, MobilityDCAT etc

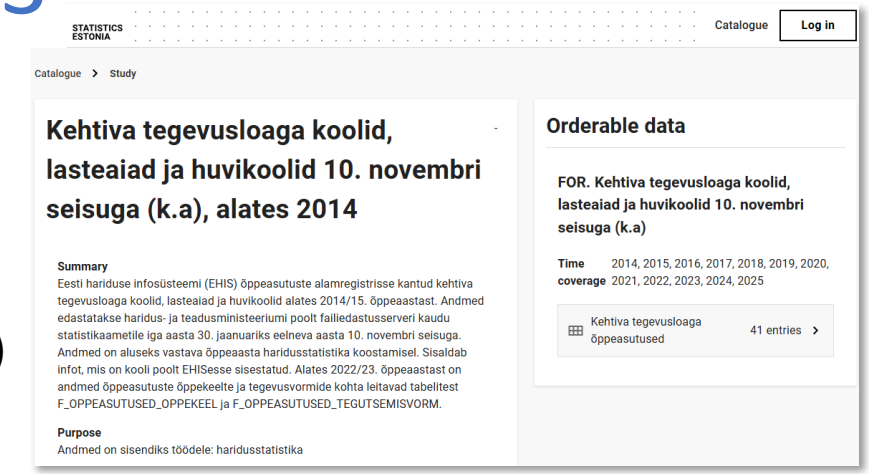
2

Statistics Estonia metadata system
colectica.stat
variable level description + code lists
DDI-LS 3.3



3

Data Sharing Service for Research (Secure Data Analysis Environment)
erika.stat.ee/en/studies
Studies > Data Files > Records (Tables) > Variables
+
Statistics Estonia and others (Health)
variable level description & ordering



Need for „standard“ about variable level descriptions exchange

- StatDCAT-AP – DCAT Application Profile for description of statistical datasets Version 1.0.1 (28.05.2019)
  - 6.2 Extensions and specific usage for description of statistical datasets
    - 6.2.1. **Dimensions and attributes**
      - **Attributes:** components used to qualify and interpret observed values such as units of measure, scaling factors
      - **Dimensions:** components that identify observations such as time, sex, age, regions
    - 6.2.2. **Quality aspects** > Data Quality Vocabulary (DQV)
    - 6.2.3. **Visualisation** - This property is to be used to indicate the type of a Distribution, in particular when the Distribution is a visualisation
    - 6.2.4. **Number of data series** - The actual number of series in the data set as referenced in the Distribution
    - 6.2.5. **Unit of measurement**
    - 6.2.6. **Specifying the length of time series**

- Additional optional classes for StatDCAT-AP

Annotation | **Attribute Property** | **Dimension Property** | **Size or duration**

EESTI STATISTIKA

Standardisation and best practicies

Need for variable level description interoperable profile (standard)

1. GSIM, DDI-LS
   - Series > Study > Data File > Data Layout > Records Layout > NCube > Variable
   - Instance Variable > Represented Variable > Conceptual Variable > Concept
2. DDI-CDI
   - Model-driven
   - Domain-independence
   - Datum-oriented data description
3. SDMX
   - Observation level metadata
     - Attributes – purely descriptive
     - Dimensions – both identify and describe
   - Table, or Dataflow, level metadata

our practice

NEW

not our practice

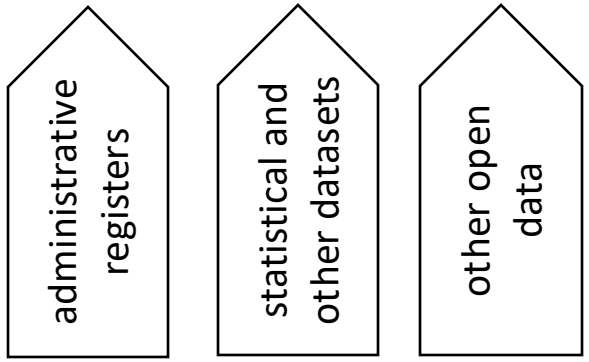
EESTI STATISTIKA

Is it more useful to change in StatDCAT extensions and specific usage for description of statistical datasets form SDMX to DDI-CDI?

European data
data.europa.eu/en

⇐

Estonian
Open Data Portal

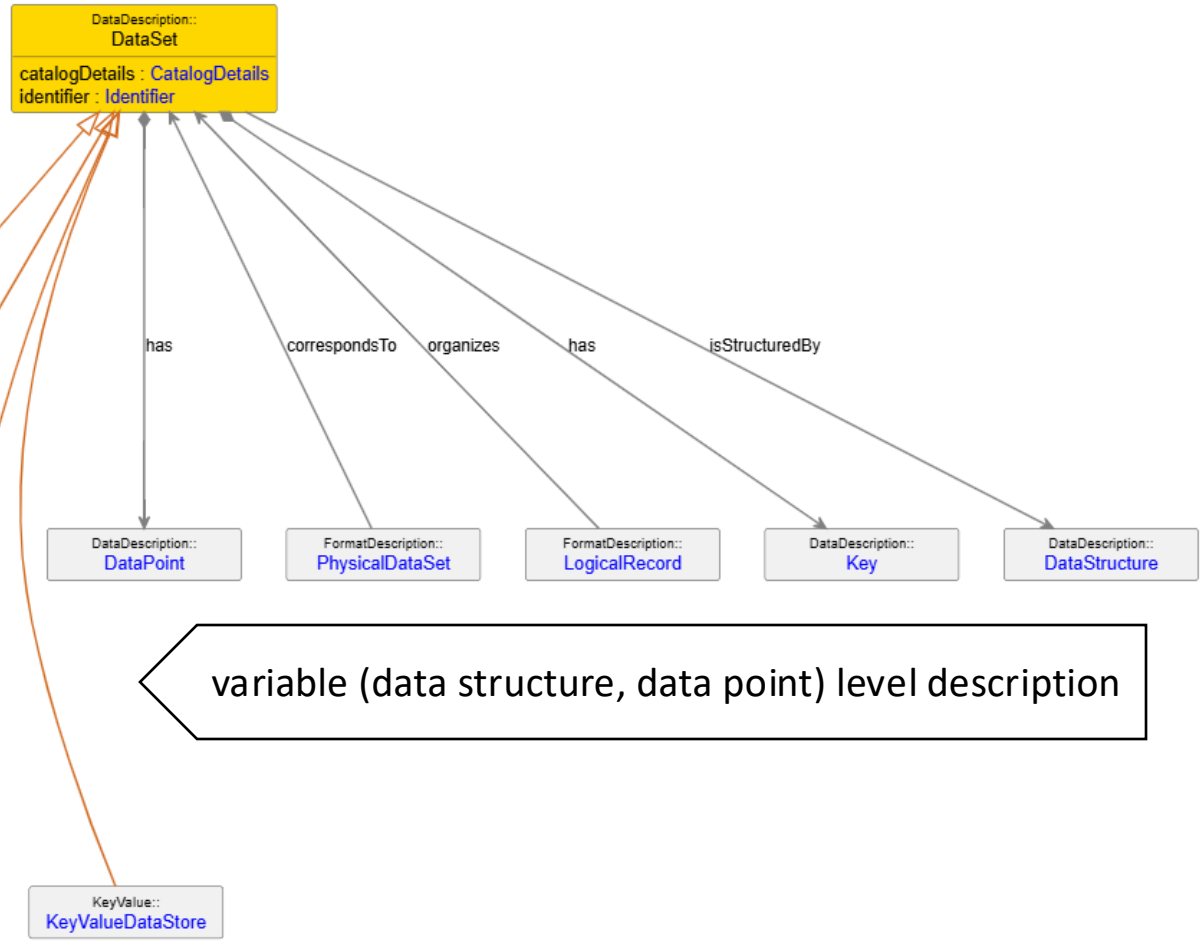


- long
- wide
- key-value

- wide
- dimensional

DDI-CDI 1.0

Data Set



EESTI STATISTIKA

Thank You!

Veiko Berendsen

STATISTICS ESTONIA
www.stat.ee
Tatari 51, 10134 Tallinn, Estonia

EESTI
STATISTIKA