Using linked metadata and quality standards for the documentation of Insee's statistical operations

EDDI 2020
Franck Cotton
Insee

1 METADATA REPOSITORY

2 STATISTICAL OPERATIONS

3 DOCUMENTATION MODEL

4 IMPLEMENTATION

5 NEXT STEPS AND CONCLUSION

01 METADATA REPOSITORY

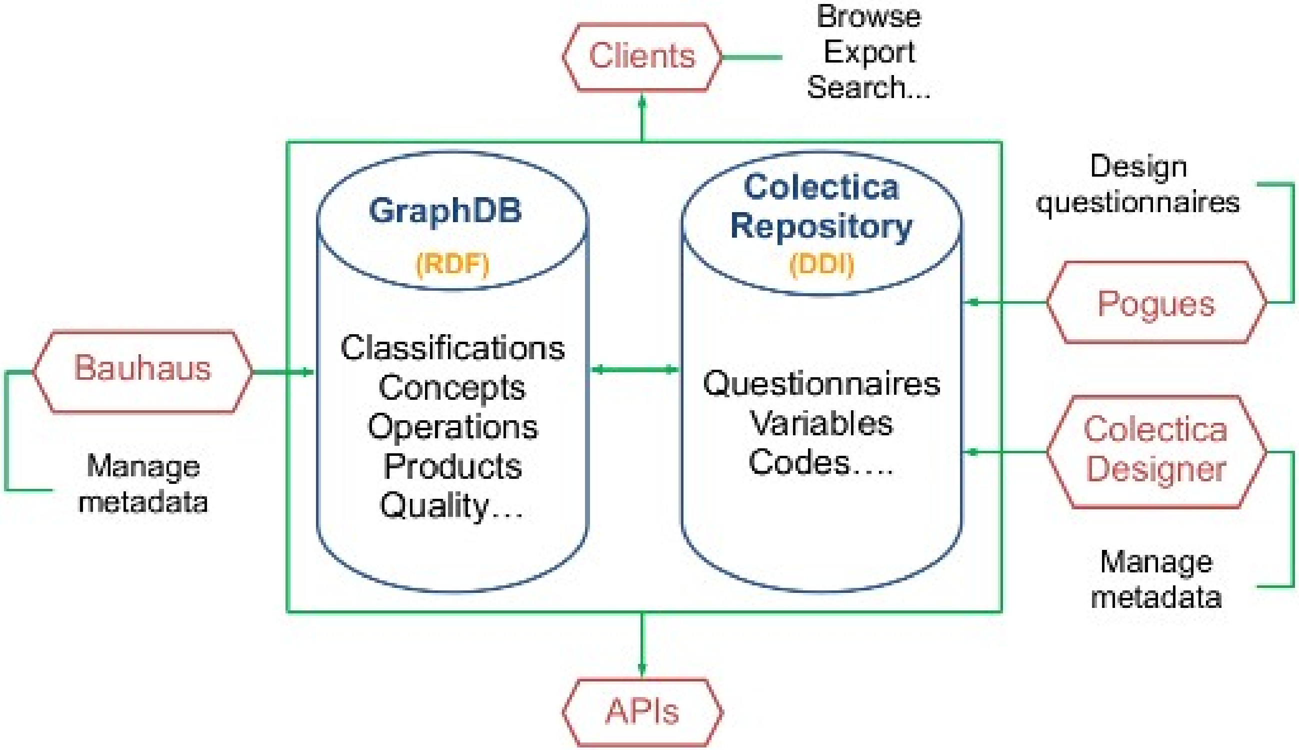
RMÉS PRINCIPLES

- A reference repository
  - Global naming
  - No duplication
- Rely on standards
  - Unece model (GSBPM, GSIM…)
  - DDI for data collection, variables, etc.
  - RDF vocabularies for concepts, classifications, etc.

RMÉS PRINCIPLES

- Active Metadata
  - Lifecycle approach
  - Machine actionability
    - Data collection for business surveys
    - Extension to household surveys
    - Work on data dissemination, administrative data ingestion…
- More on https://www.insee.fr/en/information/4195079

ARCHITECTURE OVERVIEW

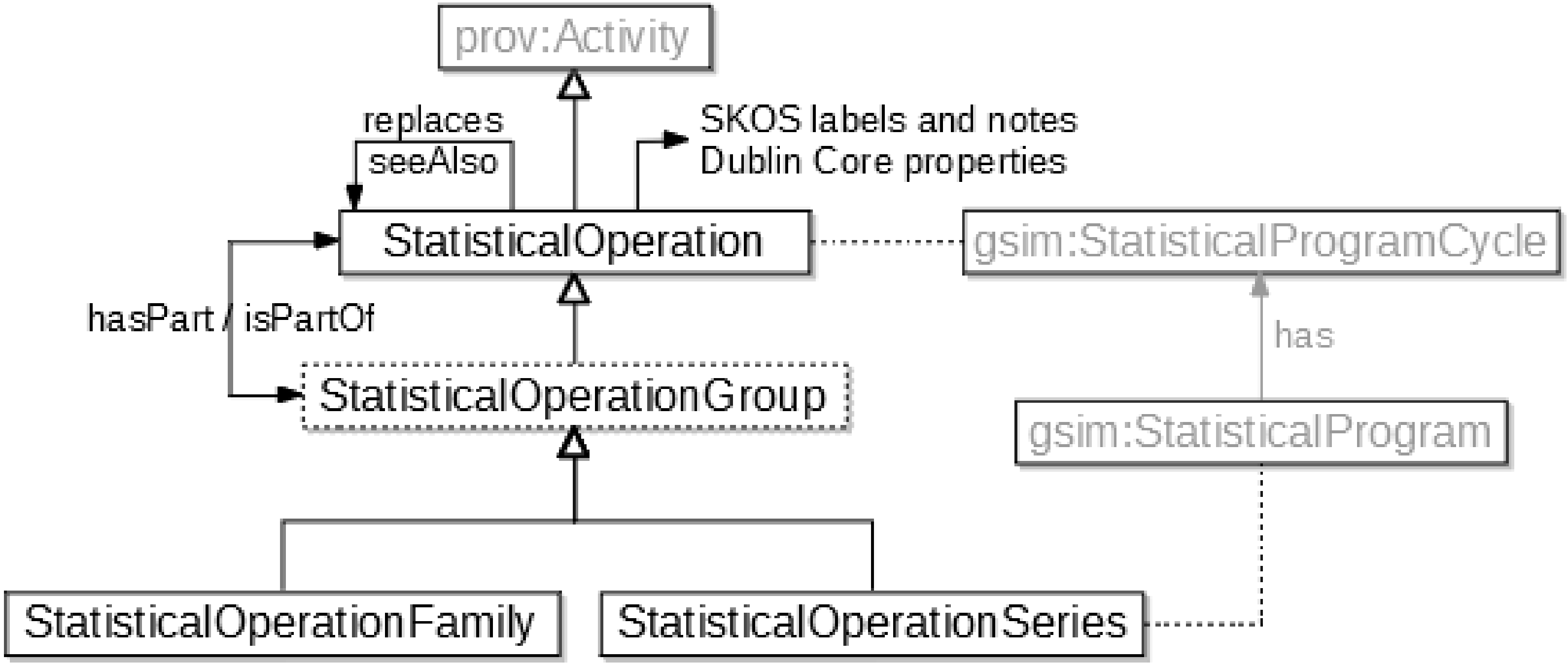


02 STATISTICAL OPERATIONS

DEFINITION

- Previously called « Sources »
  - No formal definition
  - Documented by descriptive texts
    - Published on Insee’s web site
    - No specified structure, length, update policy...
- Now « Statistical Operations »
  - Formal model, linked to GSIM
  - Structured documentation

MODEL OVERVIEW



EXAMPLE

- Family of operations on Information and Communication Technologies (ICT) contains 5 series
  - Organizational Change and ICT use surveys, replaced by:
  - ICT Enterprises, ICT Very Small Enterprises (VSE), ICT Housholds
- Series of ICT surveys on VSE has two operations
  - Survey on ICT usage - 2012
  - Survey on ICT usage and e-Commerce – 2016
- ICT Enterprises groups 13 operations (2006-2020)

03 DOCUMENTATION MODEL

OPTIONS

- Documentations on operations cover a number of subjects
  - Descriptions, actors, methodology, data products, etc.
- A possibility could be to define specialized SKOS notes
  - Example of classifications (inclusions, exclusions, case law…)
  - But more themes here, and distinction not always clear
- Documentations often close to quality reports used in ESS
  - Idea to use then emerging SIMS standard

SINGLE INTEGRATED METADATA STRUCTURE

- Version 2 adopted by the ESS in 2015
- Convergence model
  - ESMS : quality information for users
  - ESQRS : quality reporting for Eurostat
- Includes quality and performance indicators (QPI)
  - Non-response rate, imputation rate, etc.
- Formalized as an SDMX Metadata Structure Definition

SINGLE INTEGRATED METADATA STRUCTURE

- ~80 items in 19 sections: Contact, Metadata update, Statistical presentation, Unit of measure, Reference period, Institutional mandate, Confidentiality, Release policy, Frequency of dissemination, Accessibility and clarity, Quality management, Relevance, Accuracy and reliability, Timeliness and punctuality, Comparability and Coherence, Cost and burden, Data revision, Statistical processing, Comment
- Hierarchical structure (unbalanced)
    - Accuracy and reliability → Non-sampling error → Model assumption error

SIMS ENRICHMENTS

- Added a few items
  - Data collection sub-items (mode, unit, sample method and size)
  - French visa number, survey status, regional extensions
- Used more specific types
  - Dates, geographic features, persons / organisations
- Defined richer text types
  - Formalizing links to external web pages or documents

04 IMPLEMENTATION

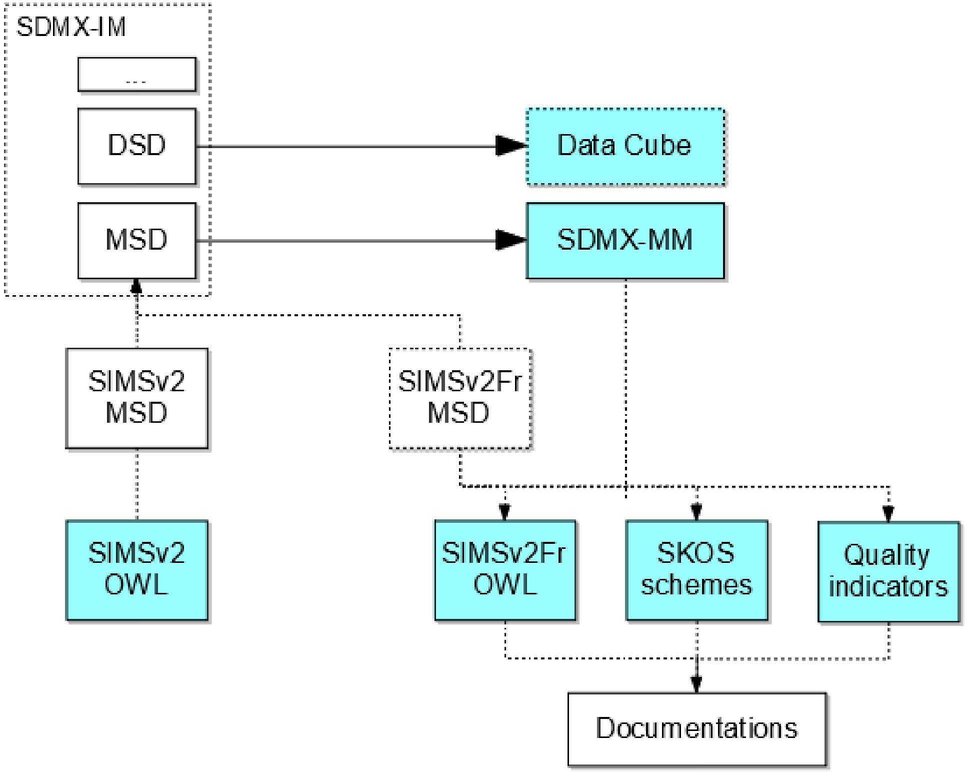
WORK ON THE CONTENT

- Work (still ongoing) with subject-matter experts
- Agree on list of operations with external stakeholders
- Update and structure documentations
- Attach it to the right level
- Intermediate tools developed (Calc files)
- Important effort

CONVERTING SIMS TO RDF

- RDF provide high-quality open metadata
- Link to other RDF models
- Converting SDMX Metadata Model to OWL
  - Metadata Structure Definition (MSD)
    - Metadata attribute to an RDF property
  - Metadata Set
    - Metadata report and reported attributes
- This allows conversion of SIMS (or any other SDMX MSD)

CONVERTING SIMS TO RDF



IMPLEMENTATION

METADATA EDITOR

BAUHAUS

SIMS METADATA MANAGEMENT

- Integrated as a module of Bauhaus
  - Insee’s metadata management tool
  - Other modules are for concepts, classifications, structures…
  - Adapts to any metadata structure
- Bauhaus
  - Is open source at https://github.com/InseeFr/Bauhaus
  - Is bi-lingual (and you can contribute other langages)
  - Can be used for browsing

05 NEXT STEPS AND CONCLUSION

NEXT STEPS

- Publication (April 2021)
  - RDF data for rich queries
  - API on Insee’s API portal
- Exports
  - Improve HTML export for web publication
  - SIMS SDMX for transmission to Eurostat
  - Documents

NEXT STEPS

- Extend to more data producers
  - Work with French ministerial statistical services to include their operations in the system
- Engage with more users
  - In particular researchers and data archives
- Extend to more metadata
  - In particular documentation of microdata

CONCLUSION

- Insee has remade the documentation of its statistical processes using quality standards
- That required a lot of work by metadata and subjet-matter experts, as well as IT people
- The resulting information system is ready to go live, and represents a huge improvement in quality
- A continuous quality improvement process is in place
- More metadata will be published in the future

Join us on

insee.fr

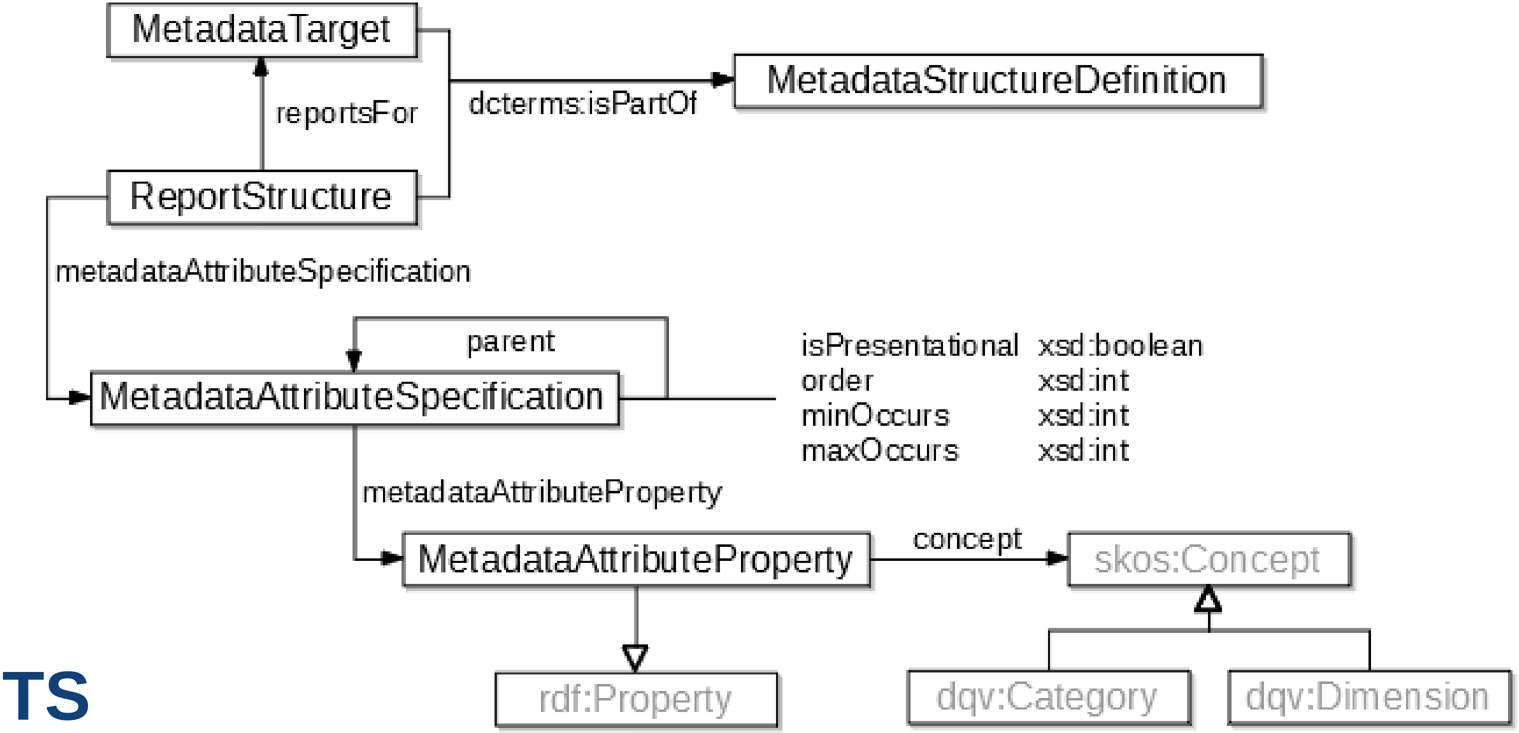
Franck Cotton

Scientific Advisor

Directorate of Information System

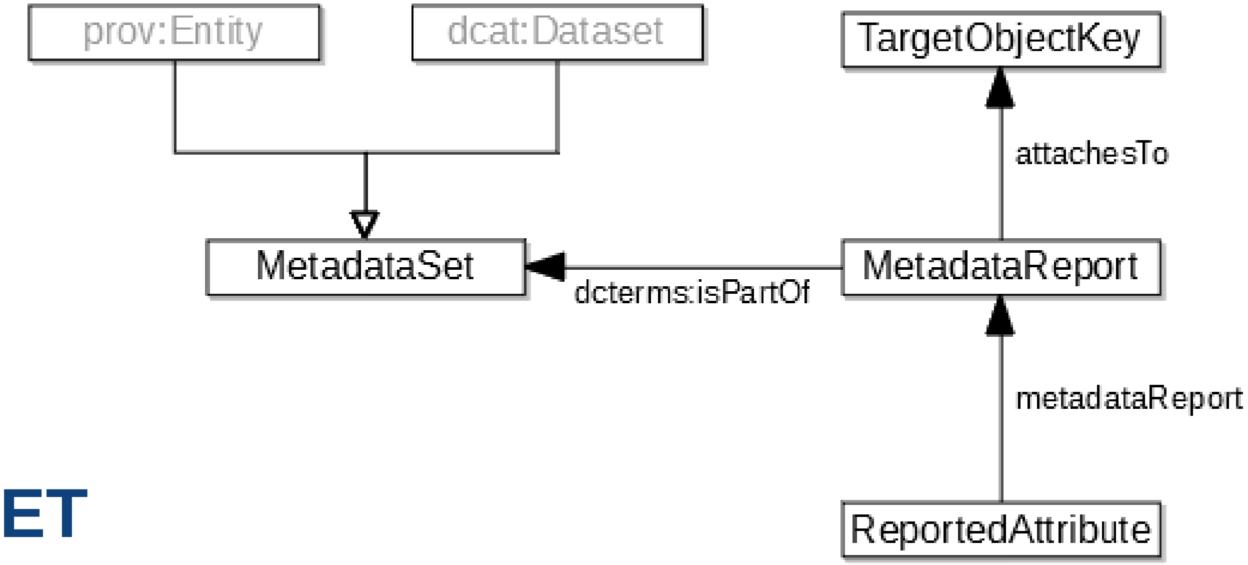
franck.cotton@insee.fr

ANNEX - SDMX METADATA MODEL AS OWL


MetadataTarget
MetadataStructureDefinition
reportsFor
dcterms:isPartOf
ReportStructure
metadataAttributeSpecification
parent
MetadataAttributeSpecification
isPresentational xsd:boolean
order xsd:int
minOccurs xsd:int
maxOccurs xsd:int
metadataAttributeProperty
MetadataAttributeProperty
concept
skos:Concept
rdf:Property
dqv:Category
dqv:Dimension

MSD ARTIFACTS

ANNEX - SDMX METADATA MODEL AS OWL



METADATA SET ARTIFACTS