Core Ontology for Official Statistics – COOS

Franck Cotton, Insee

ModernStats World Workshop, 26-28 June 2019

modernstats
BY HLG - MOS

Insee
Institut national de la statistique et des études économiques
Mesurer pour comprendre

Outline

- What is an ontology? Do we need one?
- Objectives
- Where does it come from?
- How do we work?
- Examples of questions
- What is in there?
- Next steps

What is an ontology? Do we need one?

- An ontology provides formal naming, definitions and descriptions of the concepts of a given domain, and the relations between these concepts.
- In the field of Official Statistics, a number of models, vocabularies or other semantic assets already exist, but they are not often formally expressed or coherent with one another.
- There are also very few connections between these semantic assets and models or vocabularies from other domains.
- The work on the COOS (Core Ontology for Official Statistics) aims at tackling these problems.

Objectives

- The objectives of COOS are to:
  - produce a formal representation (including global identification: OWL is used) of the *higher-level* concepts present in the main models, especially those which are not formally modeled yet (GSPBM, GAMSO)
  - specify the relations between these concepts, in particular those who belong to different models
  - define relations between statistical concepts and objets defined in other ontologies or vocabularies (actually the main objective)
- Additional goal is to document use cases showing how COOS can be put in practice
- An indirect but important objective is to constitute a community of statisticians interested by ontology design

Where does it come from?

- Early works by Istat and Insee
  - Formal representation of GSBPM, GSIM, CSPA services…
  - Scientific papers, pair-reviewed, presented at SemStats workshop, published in CEUR
- Failed in 2017 to qualify as a full-fledged HLG-MOS project (came 3rd)
- Activity was launched by the HLG-MOS after its November 2018 meeting, and placed under the umbrella of the Supporting Standards group

How do we work?

- Specific task team (13 members, Canada, France, Eurostat, Italy, Poland, Mexico, US)
- Started to work in February 2019 via virtual meetings on the basis of a dedicated GitHub repository
- Discussions are made via GitHub issues
- OWL file updated as we go along
- HTML document not updated yet
- All is open (CC-BY-4)
- You are invited

Examples of questions

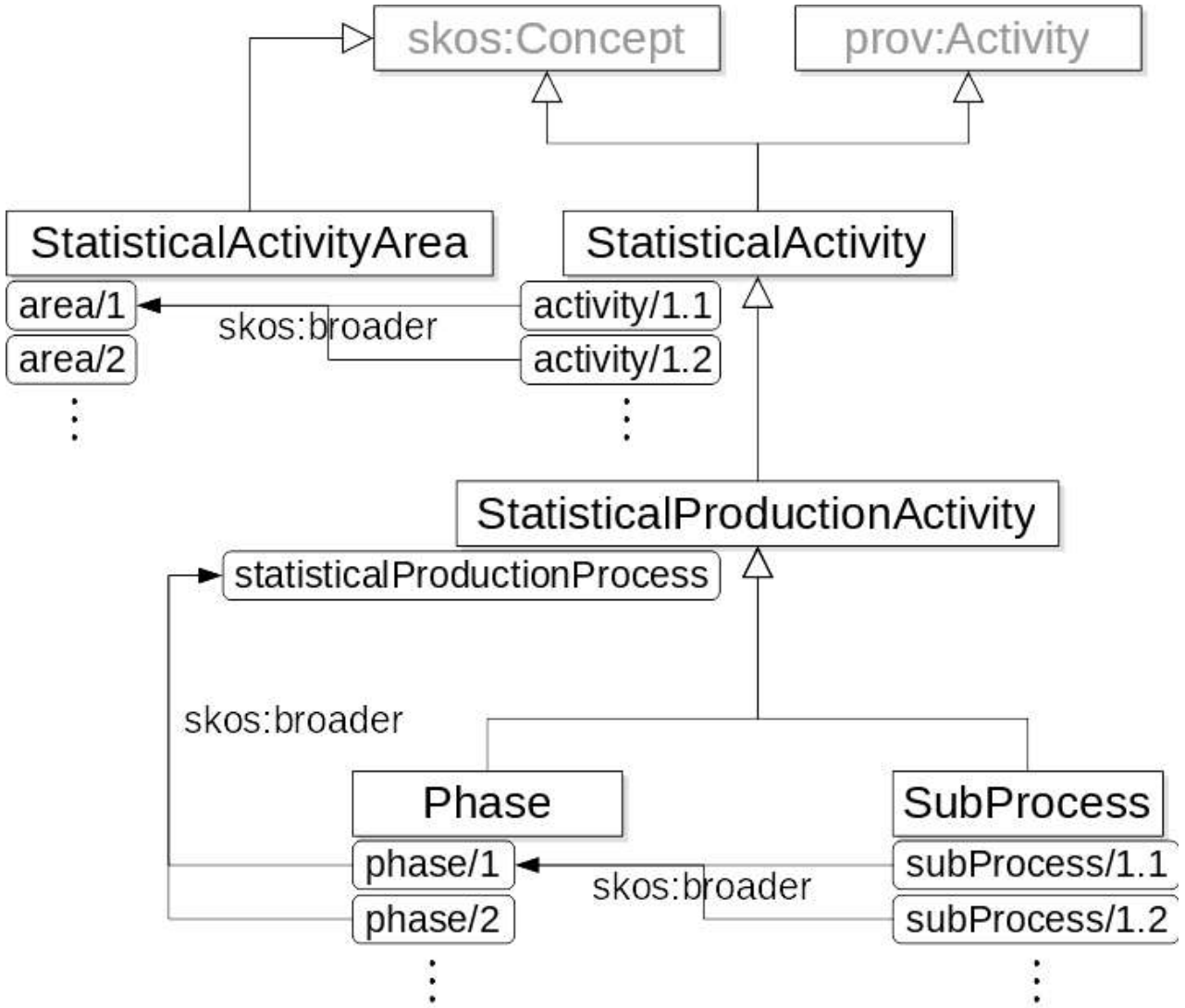
- Are GSBPM phases and sub-processes of the same nature as GAMSO activities (aka: “Activity vs. Process”)?
- Can we (should we) distinguish statistical activities and non-statistical activities in GAMSO?
- How can we link GSBPM overarching processes (eg: Metadata Management) with GAMSO activities (eg: Manage Metadata)
- If GSBPM is production area in GAMSO, why not call it Statistical Production Process?

What is in there?

- Domains covered: activities, products, organizations
- Activity and StatisticalActivity classes (link to PROV – W3C provenance ontology)
  - GSBPM phases and subprocesses are instances of (Statistical)Activity
  - GAMSO activities are instances of Activity or StatisticalActivity
- Concept schemes for GAMSO and GSBPM
  - Representing the structure of both models
  - Linked by partitive relation (“semi-unified view”)

What is in there?

Example of constructs in the “Activity” area

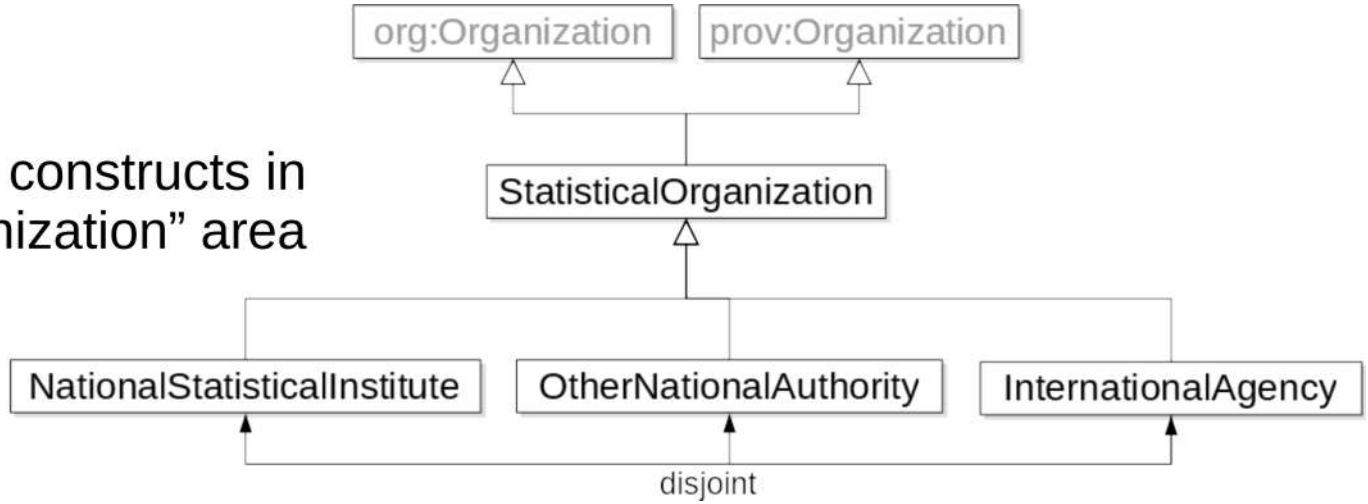


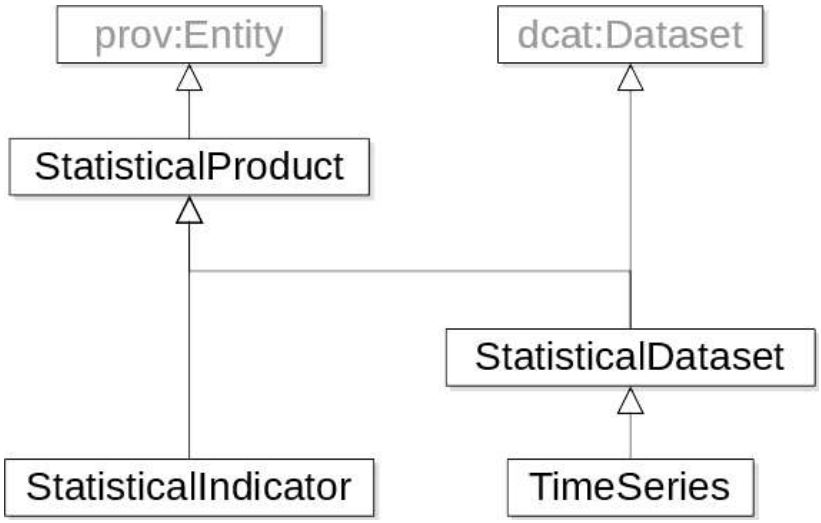
What is in there?

- StatisticalOrganization (link to ORG – W3C organization ontology and PROV)
  - Could allow to represent CSPA roles
- StatisticalProduct and StatisticalDataset (link to PROV and DCAT)
  - Do we need more detail (Indicator, TimeSeries, Publication…)?
- Elements from GSIM
  - StatisticalProgram, StatisticalProgramCycle
  - ProcessInput / ProcessInputSpecification?
- Not much from CSPA
  - Should we add StatisticalService?

What is in there?

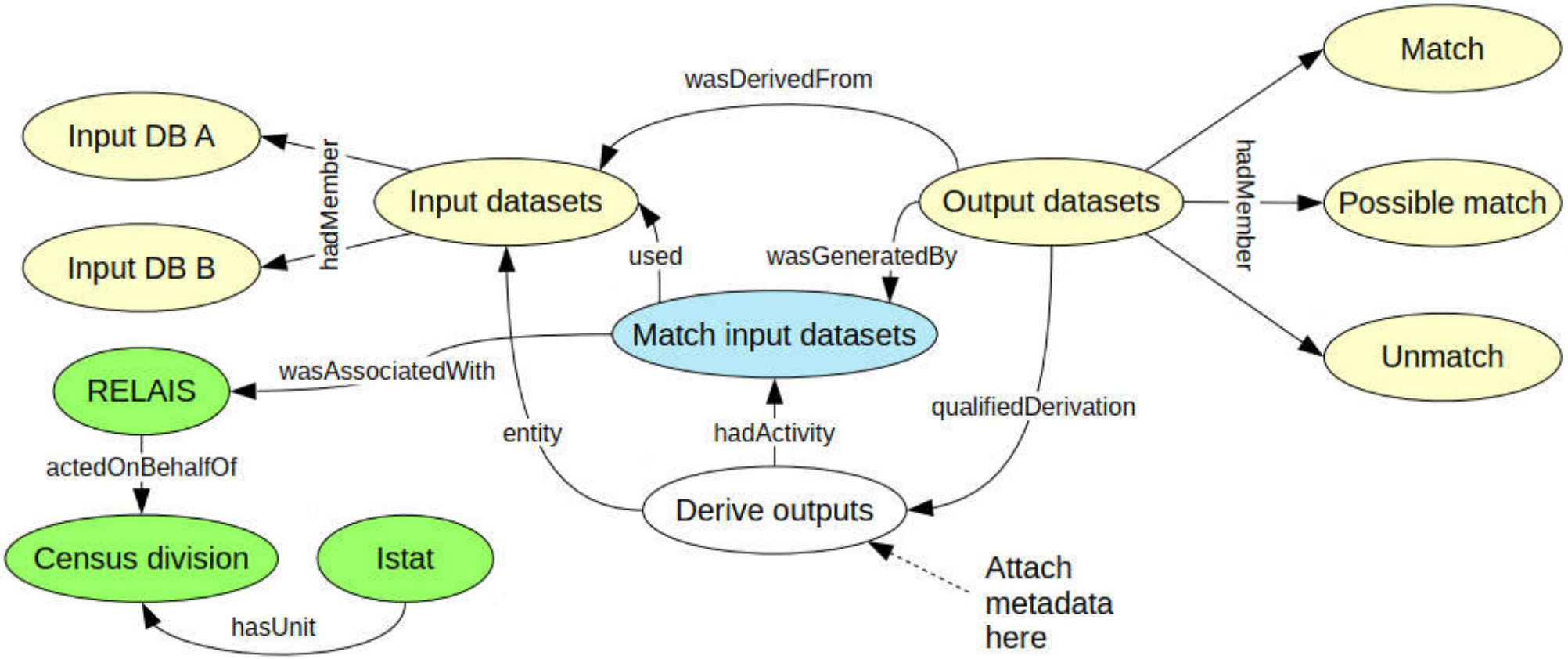
Example of constructs in the “Organization” area





Example of constructs in the “Product” area

Example use case: record linkage



Example use case: programmatic use

- COOS constructs are RDF, so:
  - They can be queried via SPARQL
    - Give me the GSIM objects whose name ends with 'specification'
  - They can be directly used in web applications
    - Example: GSIM explorer (clickable GSIM on steroids)
    - Example: List of NSIs
  - They are natively multilingual

Next steps

- Finish first version of ontology and write specification paper
- Finalize record linkage use case
- Present to Supporting Standards group
- Organize public review
- Publish COOS

Thank you

Any questions?