Modeling the Statistical Process with Linked Metadata

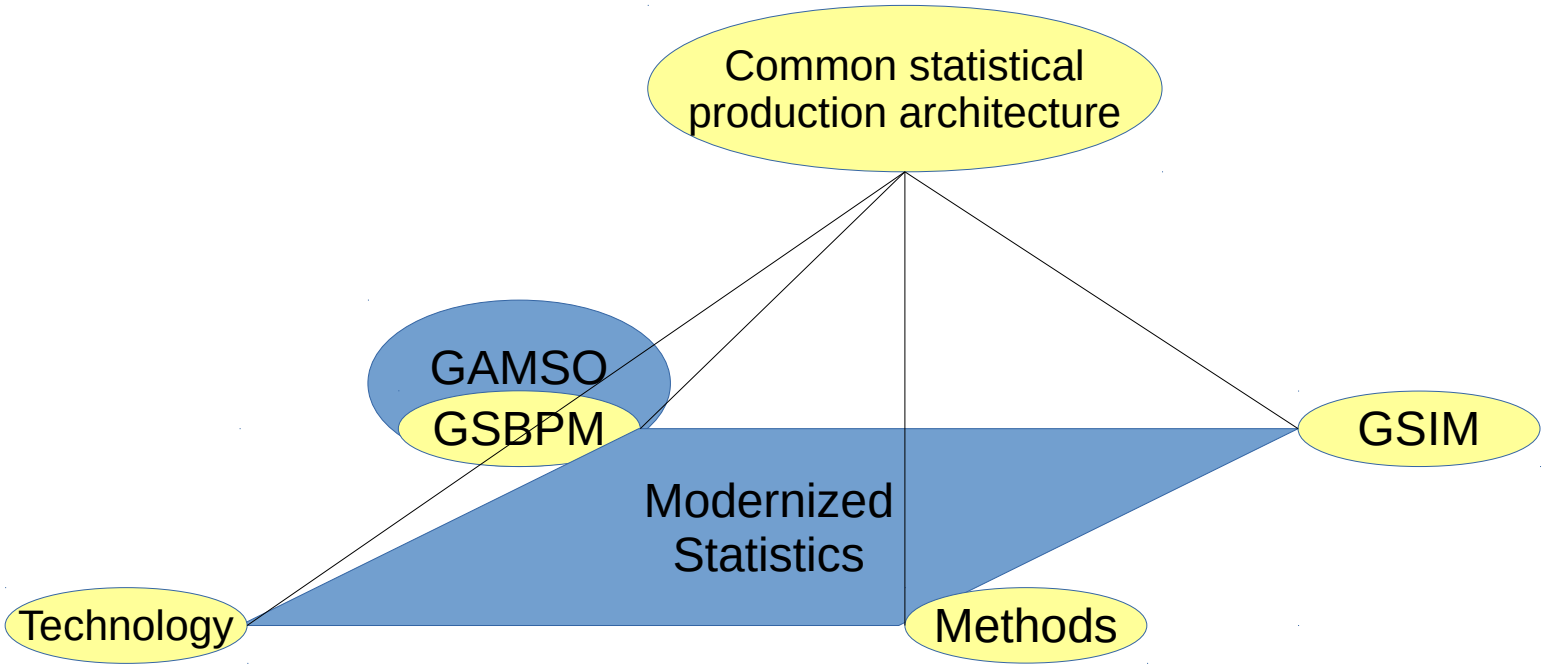
Dan Gillman – BLS
Franck Cotton – INSEE

SemStats

General context: Modernization of official statistics

- OS is challenged on products and processes
- OS needs industrialization on a global scale
  - Implies collaboration, standardization
- International initiative lead by UNECE
  - High-level group on modernization of OS
  - Works on:
    - Models (GSIM/LIM, GSBPM/GAMSO)
    - Enterprise Architecture
    - Big Data

General context: Modernization of official statistics



General context: Linked (open) statistical metadata

- One line of work for UNECE/HLG
- What are LOSM?
  - Vocabularies, concepts
  - Codes and classifications
  - Business models
  - Data description and discovery
  - Quality, methodological, process, provenance, etc.

General context: Linked (open) statistical metadata

- Why create LOSM?
  - Identify
  - Model
  - Expose
  - Link
  - Activate

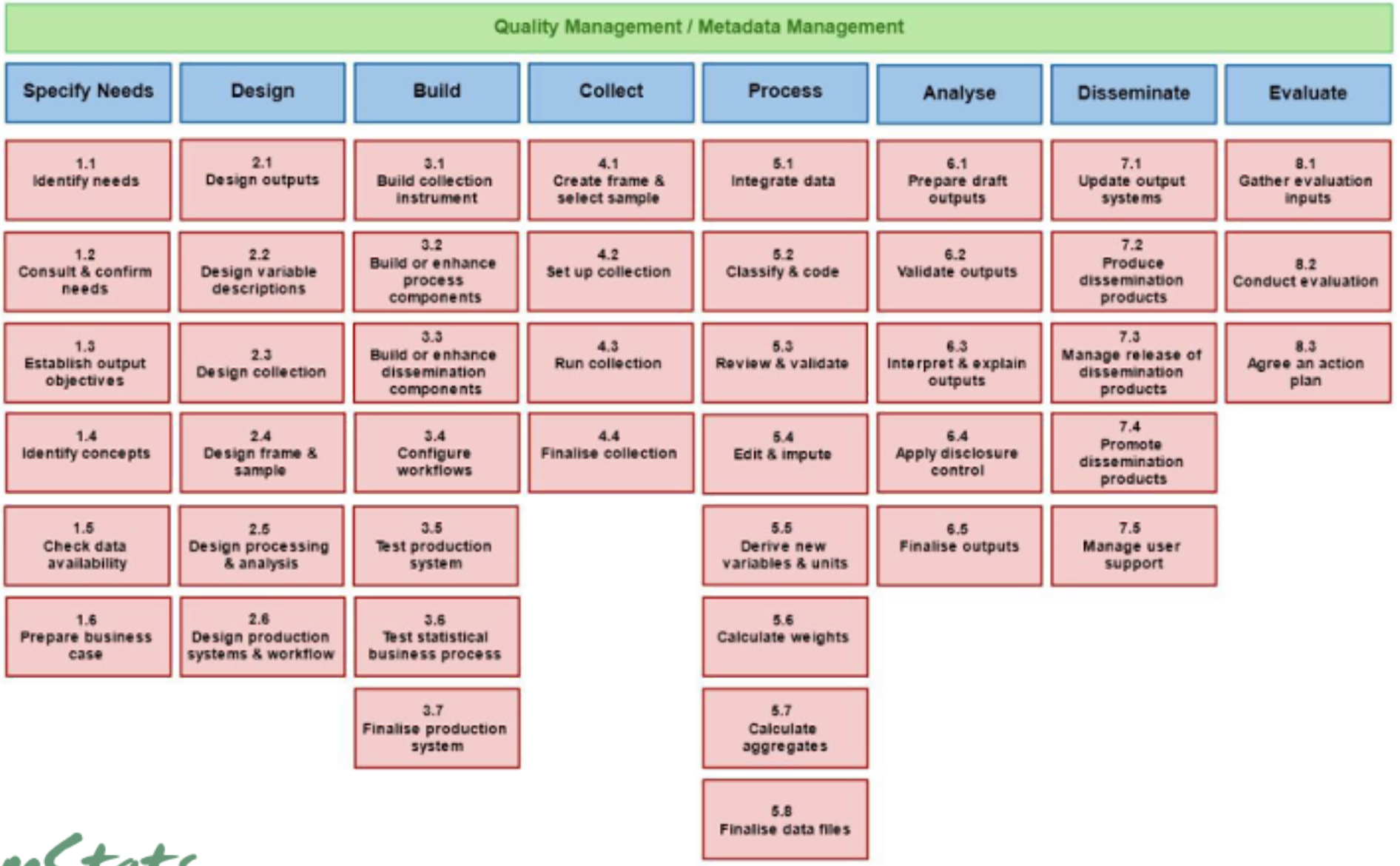
Focus on business models

- Main models
  - GSBPM: Generic Statistical Business Process Model
  - GAMSO: Generic Activity Model for Statistical Organizations
  - GSIM: Generic Statistical Information Model
  - LIM: Logical Information Model
- Published and maintained by the UNECE

Meet the GSBPM

- Current version v5.0 (released December 2013)
- 8 phases divided into 44 sub-processes
- 2 over-arching processes
  - Metadata management
  - Quality management

Meet the GSBPM



Semantizing the GSBPM

- Objectives: re-use existing work, keep it simple
- A lot of work exists on SBPM
  - Ontologies for BPMN and BPEL
  - OWL-S: Semantic Markup for Web Services
  - Service description, discovery, composition, mediation, etc.
- Less on representing process / activities *per se*
  - LOV does not help
  - Fundational ontologies (eg DOLCE) are too general

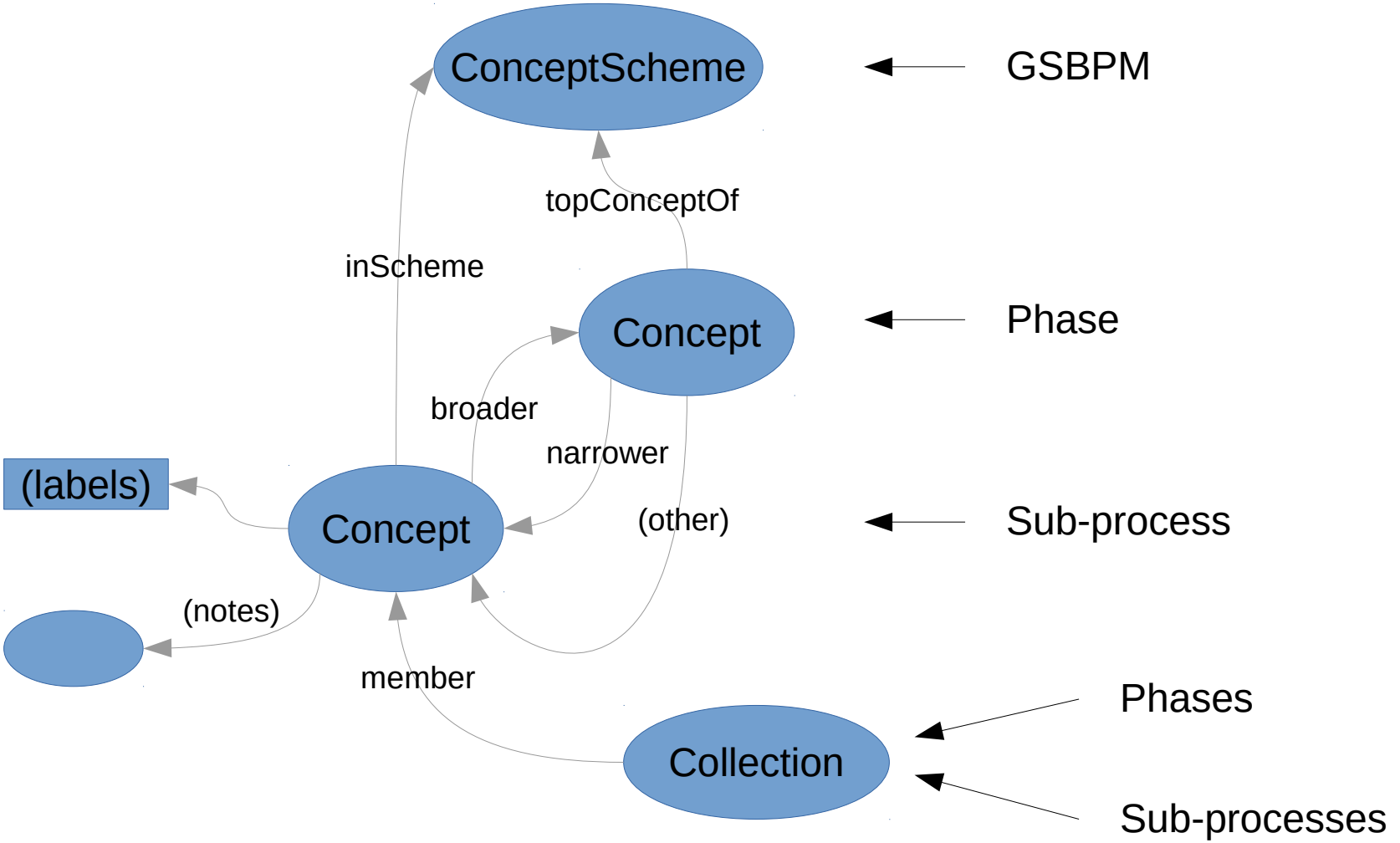
Semantizing the GSBPM

- Problem: the GSBPM is not a BPM
  - Very coarse-grain
  - No precise dependencies or sequencing between sub-processes
  - No description of flows
  - Rather a taxonomy of statistical activities

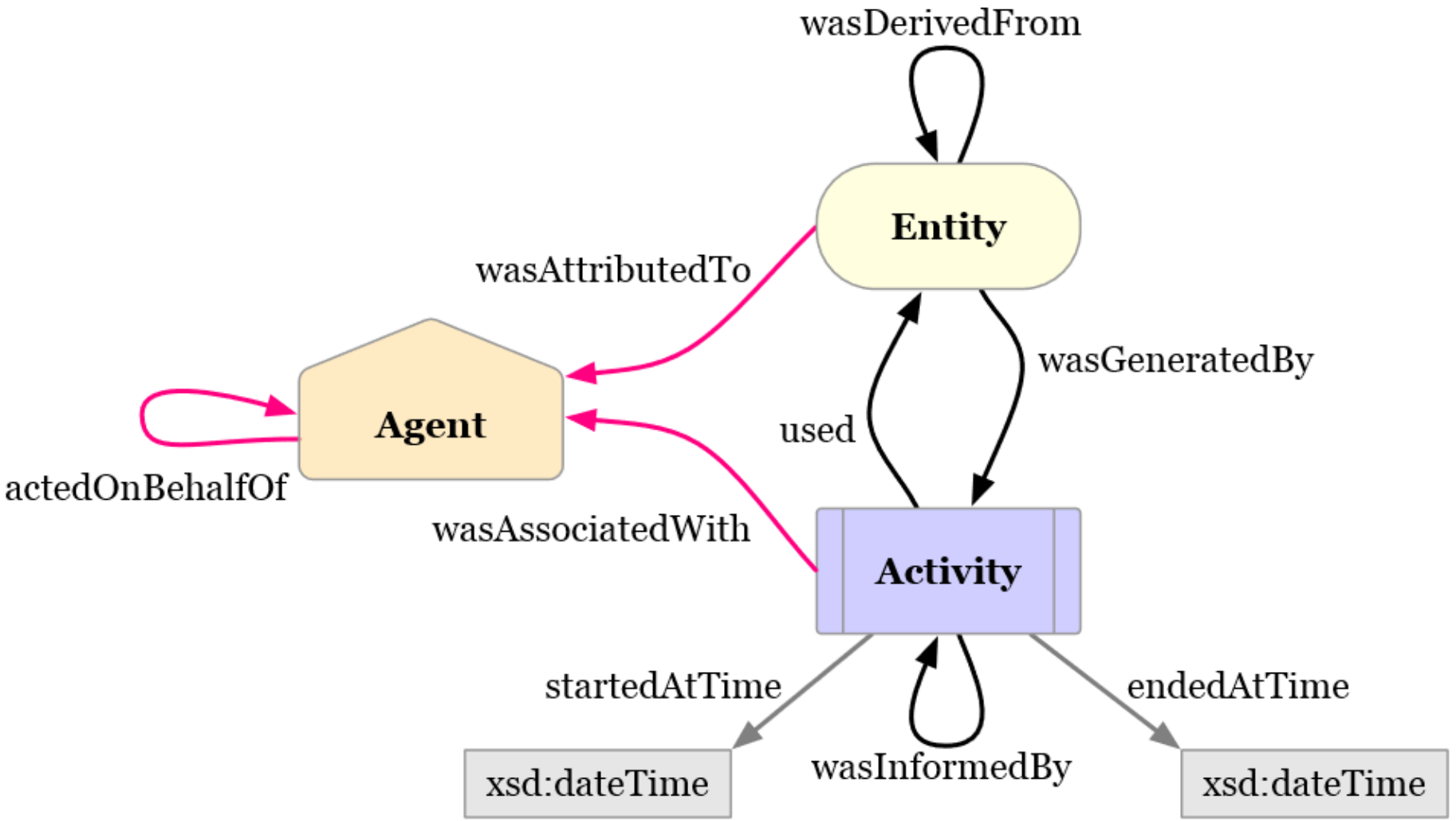
Semantizing the GSBPM

- A taxonomy (→ SKOS)…
- … of statistical activities (→ PROV)
- SKOS will allow for representing the structure
  - Possibilities to use extensions (XKOS)
- PROV will provides additional semantics:
  - Links between phases or sub-processes
  - Links with GSIM objects (PROV entities)
  - Who is responsible for the process or its outputs

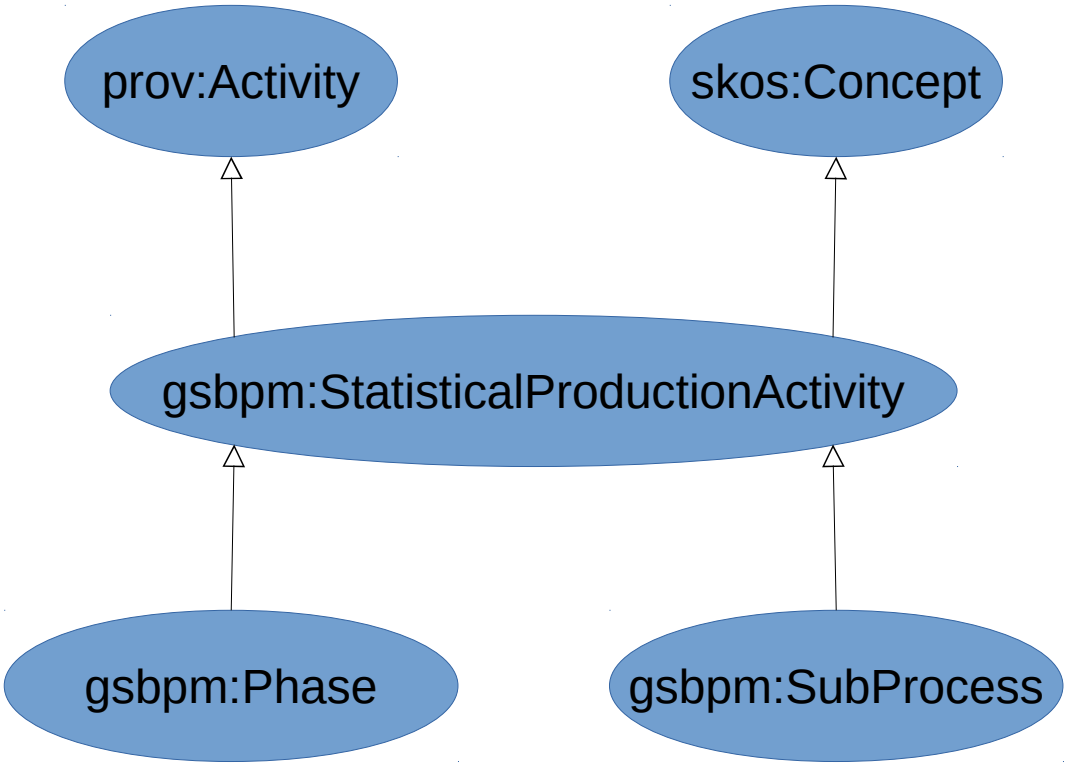
The SKOS basic model



The PROV model (starting points)



Vocabulary base structure



Vocabulary details

```
############################
# Ontology
############################

<http://rdf.unece.org/models/gsbpm>
        a                               voaf:Vocabulary , owl:Ontology ;
        cc:license                      <http://creativecommons.org/licenses/by/3.0/> ;
        dc:rights                       "Copyright © 2015 INSEE" ;
        dcterms:creator                 [ a             foaf:Person ;
                                          foaf:name     "Franck Cotton"
                                        ] ;
        dcterms:description             "Vocabulaire pour la représentation du GSBPM en RDF"@fr ,
                                        "Vocabulary for the representation of the GSBPM as RDF"@en ;
        dcterms:issued                  "2015-06-10"^^xsd:date ;
        dcterms:modified                "2015-06-10"^^xsd:date ;
        dcterms:publisher               <http://dbpedia.org/resource/United_Nations_Economic_Commission_for_Europe> ;
        dcterms:title                   "Vocabulaire GSBPM"@fr , "GSBPM Vocabulary"@en ;
        vann:preferredNamespacePrefix   "gsbpm" ;
        vann:preferredNamespaceUri      gsbpm: ;
        voaf:classNumber                3 ;
        voaf:propertyNumber             0 ;
        owl:versionInfo                 "Version 5.0" .
```

Vocabulary details

```
#############################
# Classes
#############################

gsbpm:StatisticalProductionActivity
        a                   owl:Class , rdfs:Class ;
        rdfs:isDefinedBy    <http://rdf.unece.org/def/gsbpm#> ;
        rdfs:label          "Activité de production statistique"@fr , "Statistical production activity"@en ;
        rdfs:subClassOf     skos:Concept , prov:Activity .

gsbpm:Phase   a             owl:Class , rdfs:Class ;
        rdfs:isDefinedBy    <http://rdf.unece.org/def/gsbpm#> ;
        rdfs:label          "Phase du GSBPM"@fr , "GSBPM phase"@en ;
        rdfs:subClassOf     gsbpm:StatisticalProductionActivity .

gsbpm:SubProcess   a        owl:Class , rdfs:Class ;
        rdfs:isDefinedBy    <http://rdf.unece.org/def/gsbpm#> ;
        rdfs:label          "Sous-processus du GSBPM"@fr , "GSBPM sub-process"@en ;
        rdfs:subClassOf     gsbpm:StatisticalProductionActivity .
```

Vocabulary details

```
############################
# Individuals
############################

igsbpm:1  a                 gsbpm:Phase ;
        skos:definition     "This phase is triggered when a need for new statistics is identified, or feed
        includes all activities associated with engaging customers to identify their detailed statistical
        business cases to meet these needs.\nIn this phase the organisation:\n- identifies the need for t
        needs of the stakeholders;\n- establishes the high level objectives of the statistical outputs;\n
        data are required;\n- checks the extent to which current data sources can meet these needs;\n- pr
        statistics.\nThis phase is broken down into six sub-processes. These are generally sequential, fr
        iterative. The sub-processes are:\n"@en ;
        skos:narrower       igsbpm:1.4 , igsbpm:1.5 , igsbpm:1.2 , igsbpm:1.3 , igsbpm:1.1 , igsbpm:1.6 ;
        skos:notation       "1" ;
        skos:prefLabel      "Specify Needs"@en ;
        skos:topConceptOf   igsbpm:gsbpm .

igsbpm:1.1  a               gsbpm:SubProcess ;
        skos:broader        igsbpm:1 ;
        skos:definition     "This sub-process includes the initial investigation and identification of what
        It may be triggered by a new information request, an environmental change such as a reduced budge
        the process, or from other processes, might provide an input to this sub-process. It also include
        international) statistical organisations producing similar data, and in particular the methods us
        specific needs of different user communities, such as the disabled, or different ethnic groups.\n
        skos:inScheme       igsbpm:gsbpm ;
        skos:notation       "1.1" ;
        skos:prefLabel      "Identify Needs"@en .
```

Future work

- Identify
  - Define URI scheme
- Model
  - Introduce PROV information
- Expose
  - Clickable GSBPM
  - Translations

Future work

- Link
  - Glossaries
  - Local refinements
  - GSIM
  - Attach detailed process models
- Activate
  - Quality indicators
  - CSPA Service definitions and descriptions

Modeling the Statistical Process with Linked Metadata

Thank you

SemStats