Métadonnées statistiques et RDF

Franck Cotton – Institut National de la Statistique et des Études Économiques

APSEM 2018
ENSEEIHT – Toulouse – 12-15 novembre 2018

Agenda

Données et métadonnées

Description de la structure des données

Description des jeux de données

- Qualifier
- Découvrir
- Sourcer

Données et métadonnées

Données et métadonnées

479 638

Données et métadonnées

479 638 personnes

Données et métadonnées

479 638 personnes

Population totale de Toulouse

Données et métadonnées

479 638 personnes

Population totale de Toulouse

En 2015

Données et métadonnées

479 638 personnes

Population totale de Toulouse

En 2015

Source : Insee, Recensement de la population 2015 en géographie au 01/01/2017

Données et métadonnées

479 638 personnes

Population totale de Toulouse

En 2015

Source : Insee, Recensement de la population 2015 en géographie au 01/01/2017

Méthode : voir le décret n° 2003-485 du 5 juin 2003 relatif au recensement de la population

Données et métadonnées

Sans métadonnées, la donnée n’a pas de sens

Comprendre la donnée, c’est comprendre les métadonnées

Nombreux types de métadonnées :

- structurelles
- descriptives
- provenance
- qualité
- licence
- etc.

Description de la structure des données

Description de la structure des données

On se limite aux données multidimensionnelles

Principaux vocabulaires RDF utilisés

W3C Data Cube

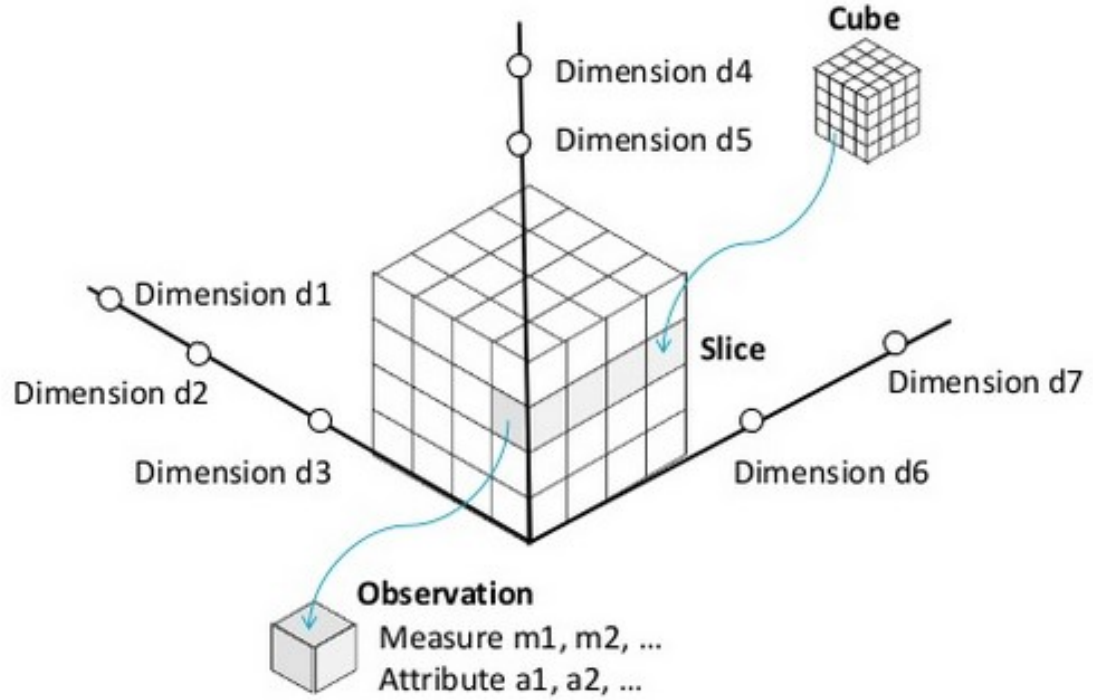
- Publication de données multidimensionnelles sour forme de Linked Data
- Recommandation du W3C
- Basée sur le modèle d'information de SDMX (ISO 17369:2013)

SKOS / XKOS

- Représentation des structures de concepts et listes de codes associées aux données
- SKOS est une recommendation du W3C
- XKOS, standard de l'Alliance DDI, étend SKOS pour les besoins statistiques

Description de la structure des données

Modèle SDMX



Le modèle couvre la description du jeu de données (Data Set) et la définition de la structure de données (DSD)

Description de la structure des données

Transposition en RDF Data Cube

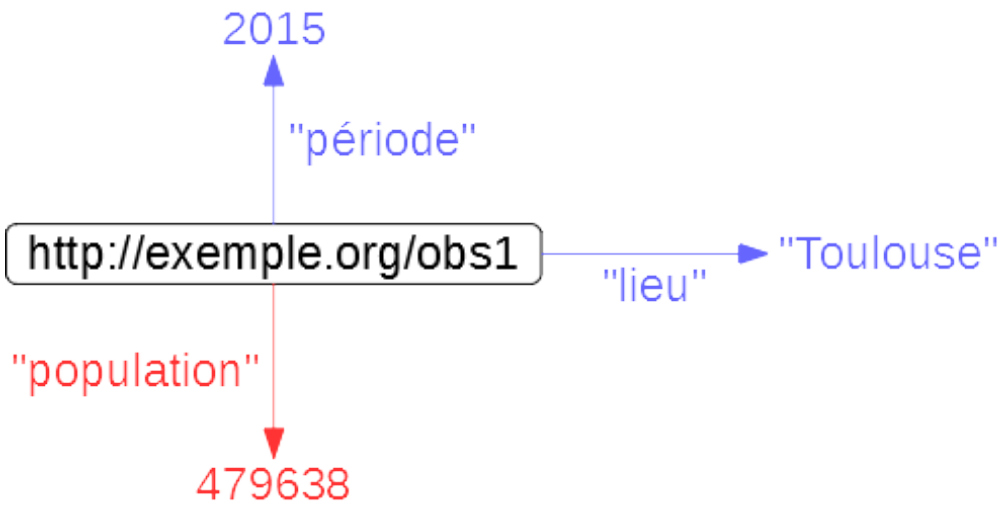
Les objets centraux du modèle se transposent en classes RDFS

- qb:DataSet
- qb:Slice
- qb:Observation
- qb:DataStructureDefinition
- etc.

Description de la structure des données

Transposition en RDF Data Cube

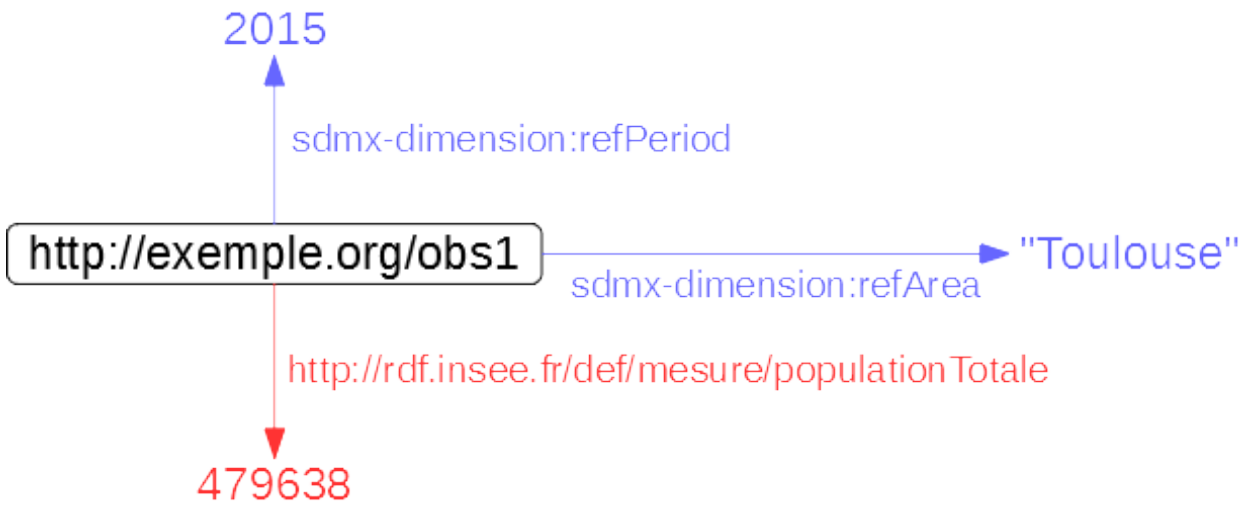
Les composants de la structure de données (dimensions, mesures et attributs) se traduisent par des propriétés RDFS



Description de la structure des données

Transposition en RDF Data Cube

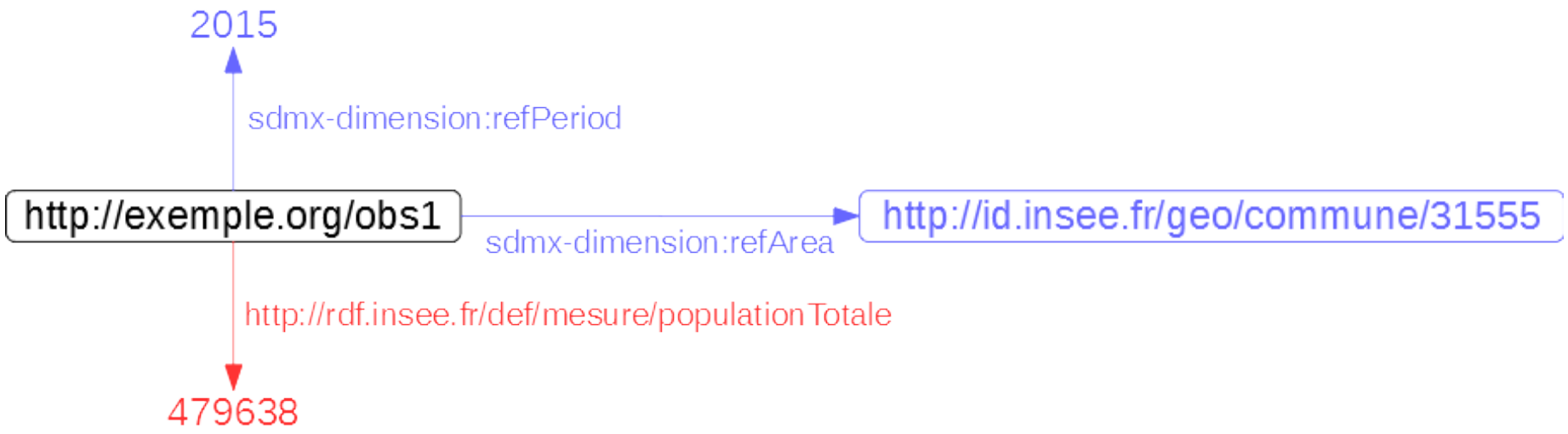
Data Cube prédéfinit des propriétés RDFS correspondant aux concepts SDMX les plus usuels



Description de la structure des données

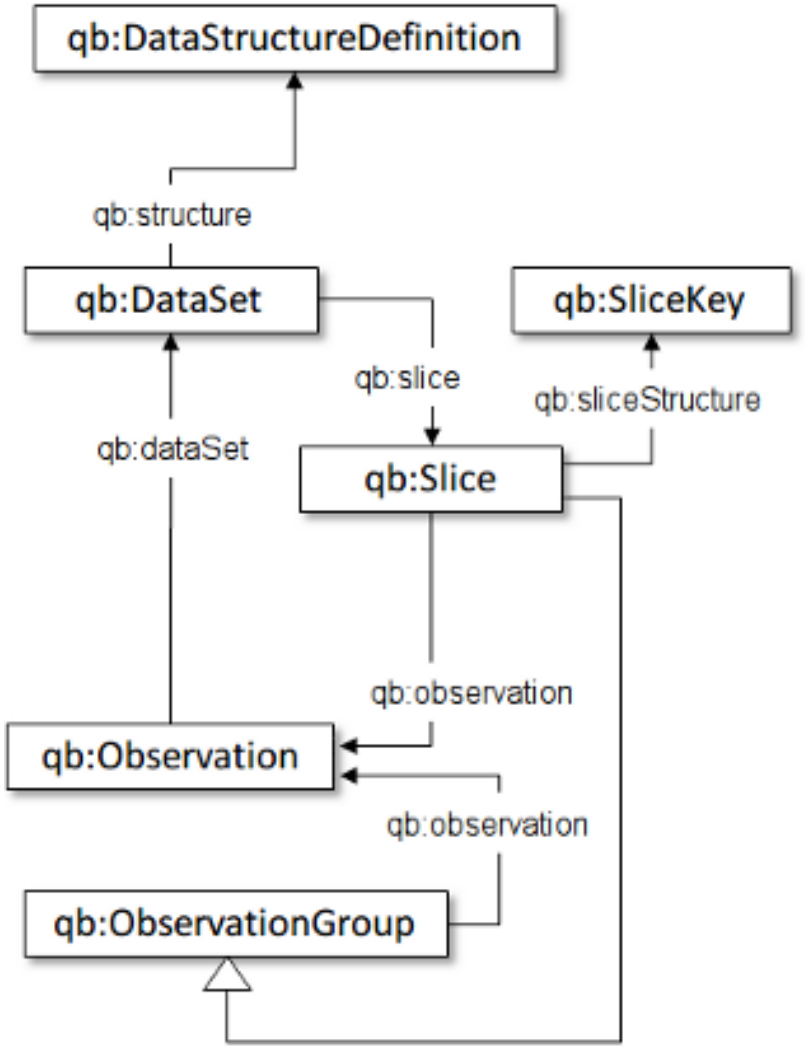
Transposition en RDF Data Cube

On bénéficie de toutes les possibilités de RDF



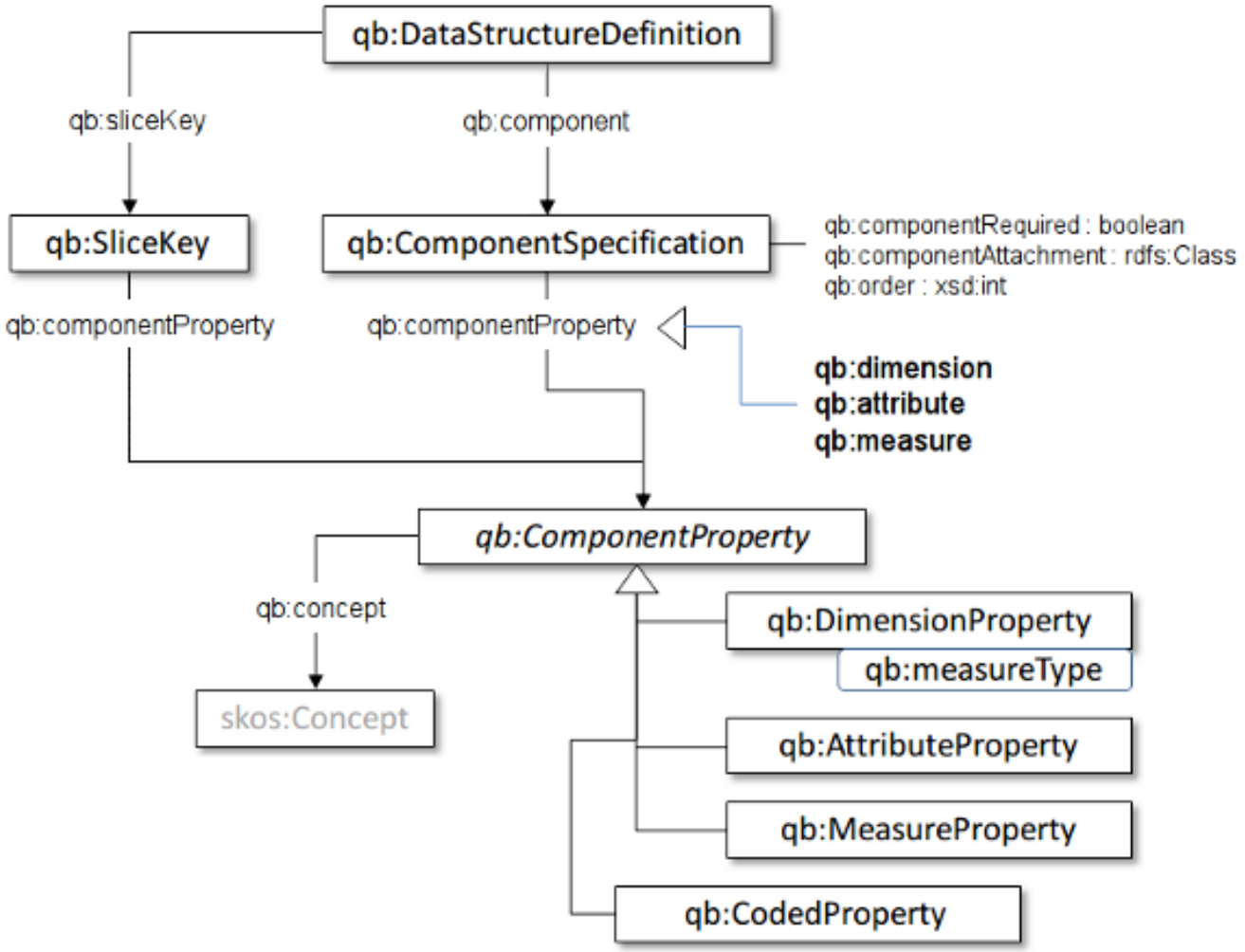
Data Cube – Modèle

Data Set



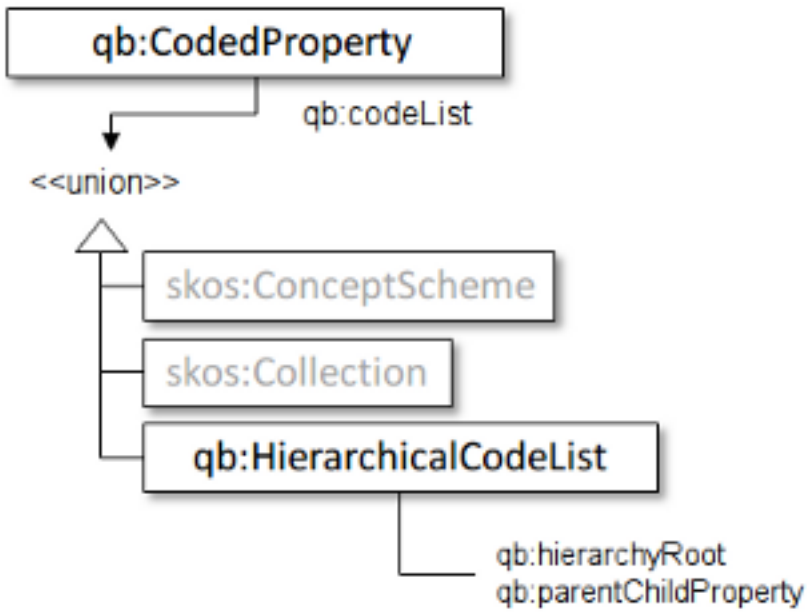
Data Cube – Modèle

Définition de Structure de Données



Data Cube – Modèle

Composants codés



Data Cube – Exemple

```
eg:dsd-pl a qb:DataStructureDefinition ;
    rdfs:label    "Populations légales par commune et année"@fr ;
    # Dimensions
    qb:component [ qb:dimension sdmx-dimension:refArea;      qb:order 1 ] ;
    qb:component [ qb:dimension sdmx-dimension:refPeriod;    qb:order 2 ] ;
    # Mesure
    qb:component [ qb:measure insee-mesure:populationTotale ] .

eg:dataset-pl2015 a qb:DataSet ;
    rdfs:label    "Recensement de la population 2015 - populations légales"@fr ;
    qb:structure  eg:dsd-pl .

eg:obs1 a qb:Observation ;
    qb:dataSet                      eg:dataset-pl2015 ;
    sdmx-dimension:refArea          <http://id.insee.fr/geo/commune/31555> ;
    sdmx-dimension:refPeriod        "2015-01-01"^^xsd:date ;
    insee-mesure:populationTotale   479638 .

eg:obs1 a qb:Observation ;
    qb:dataSet                      eg:dataset-pl2015 ;
    sdmx-dimension:refArea          <http://id.insee.fr/geo/commune/67482> ;
    sdmx-dimension:refPeriod        "2015-01-01"^^xsd:date ;
    insee-mesure:populationTotale   281512 .
```

Data Cube et SKOS / XKOS

Les composants d'une DSD sont associés à des concepts

Lieu, période, commune, année, activité économique…

Population totale, légale, comptée à part, chiffre d'affaires…

Statut d'observation, unité de mesure, précision…

→ Ces concepts sont modélisés en utilisant SKOS

Les valeurs d'une propriété peuvent être

Des types simples (entier, flottant, chaîne…) : effectif, âge moyen, EBE…

Des modalités de listes de codes : classe d'âge, activité économique…

→ Ces listes de codes sont modélisées en utilisant SKOS ou XKOS

SKOS / XKOS

SKOS (Simple Knowledge Organization System)

- Modèle pour représenter les systèmes d'organisation des connaissances en RDF
  - Thesaurus, listes de codes, taxonomies, vocabulaires, nomenclatures, etc.
- Très utilisé
- Simple and extensible

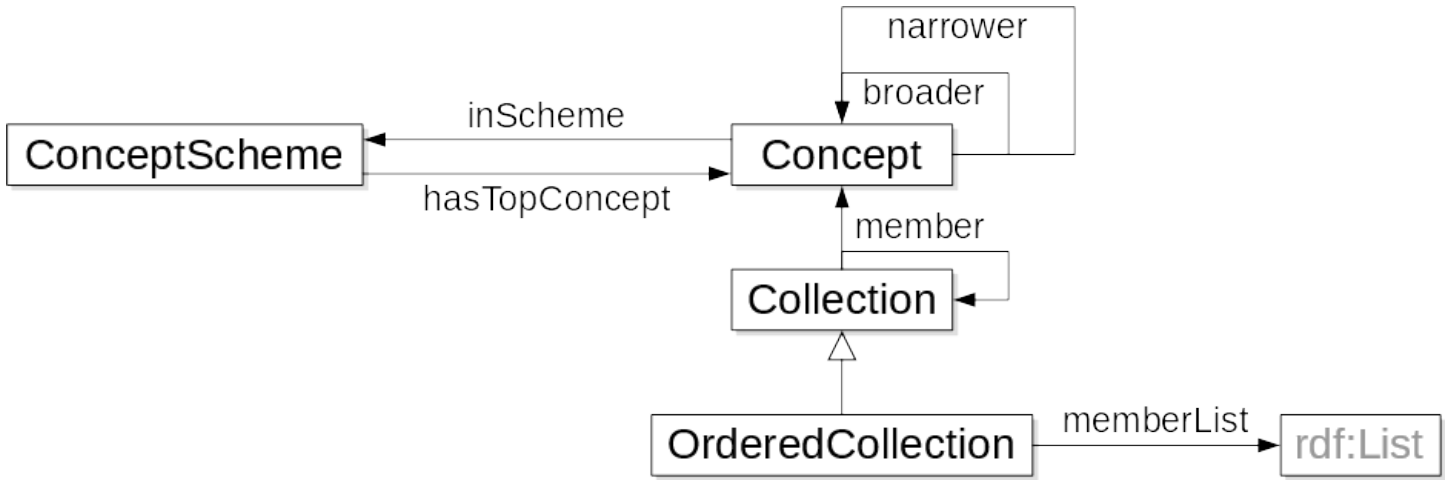
XKOS étend SKOS pour modéliser les nomenclatures statistiques

- Niveaux
- Notes explicatives
- Correspondances

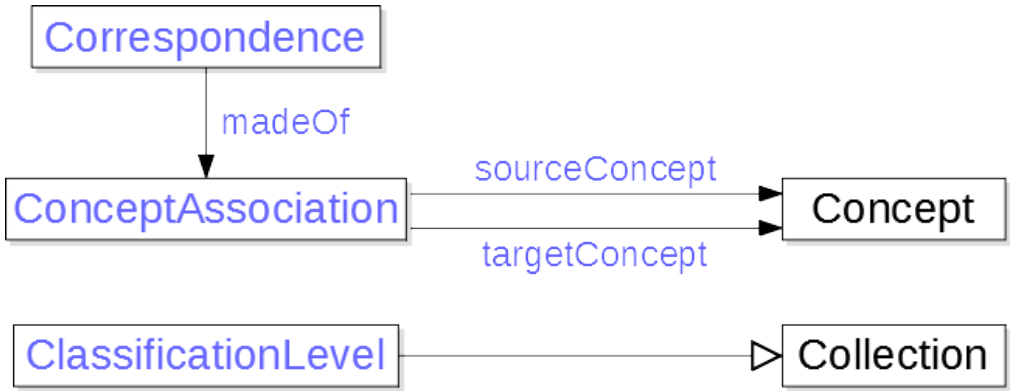
SKOS / XKOS – Modèle

Structures de concepts

Base SKOS



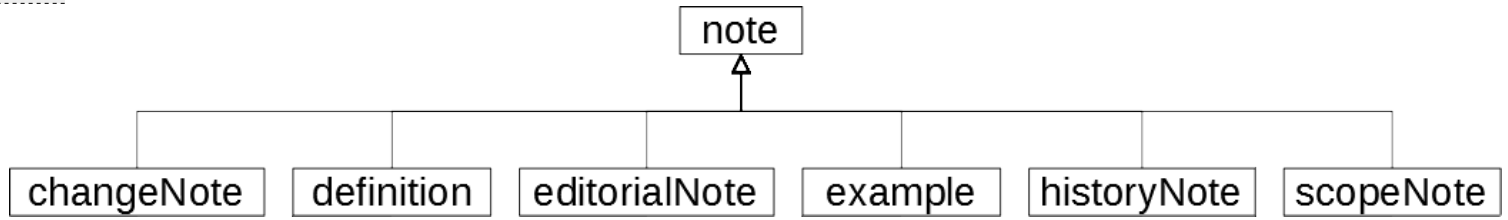
Ajouts XKOS



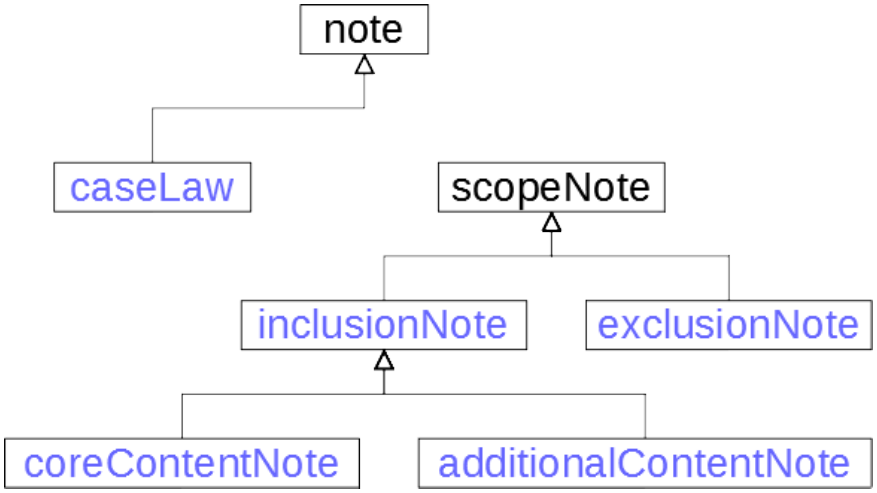
SKOS / XKOS – Modèle

Annotations

Base SKOS



Ajouts XKOS

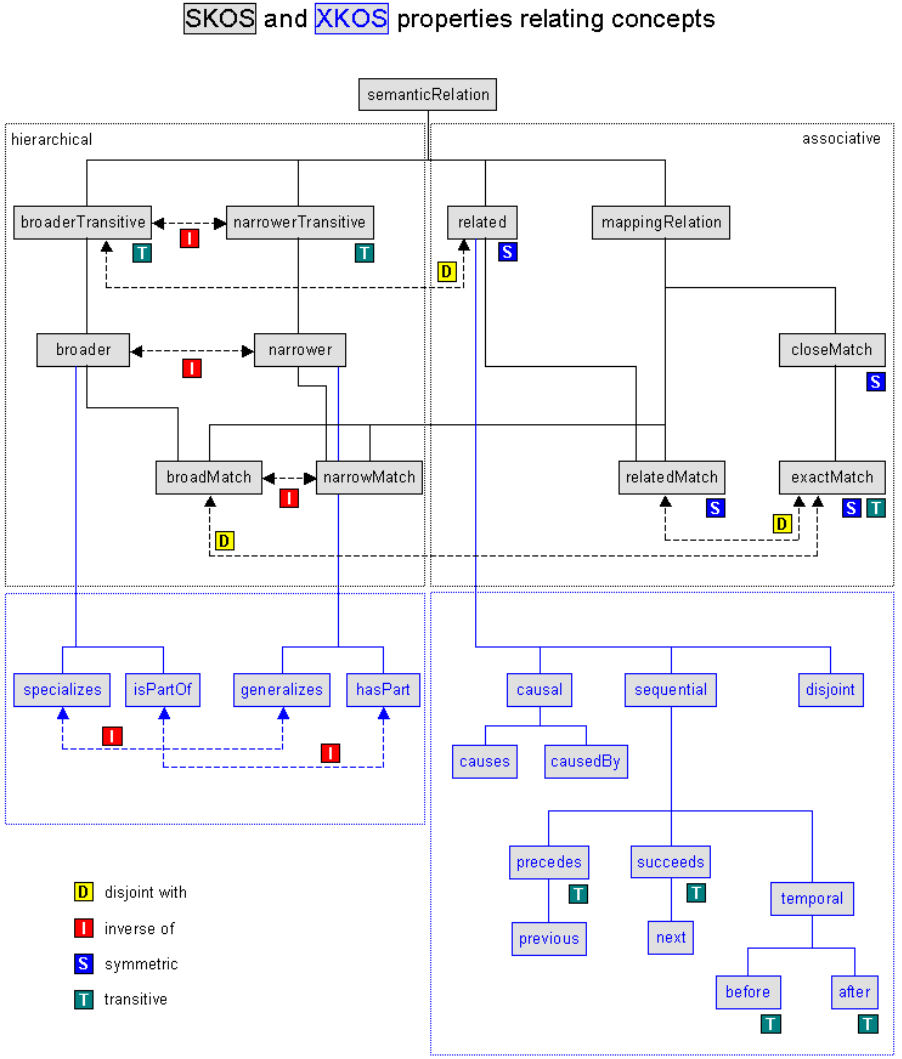


SKOS / XKOS – Modèle

Propriétés sémantiques

Base SKOS

Ajouts XKOS



Description de la structure des données

Conclusions

Il existe des vocabulaires RDF éprouvés pour :

- La publication des données multidimensionnelles
- La description de leur structure
- La formalisation des concepts associés

Des outis génériques ont été produits

- Conversion
- Exploration / visualisation

→ C’est le minimum pour publier des données compréhensibles

Description des jeux de données

Description des jeux de données

Décrire la structure des jeux de données est un premier pas

D'autres types de métadonnées sont nécessaires pour pouvoir utiliser les données

- Qualification des données et des jeux de données
- Description des données pour en améliorer la « découvrabilité »
- Spécifier d'où viennent les données et comment elles ont été produites

Description des jeux de données : qualifier

Description des jeux de données : qualifier

Exemples

- Indicateurs de qualité
- Statut de la donnée (provisoire, estimée, finale…)

Différentes solutions

- Outils généraux (ex: annotations)
- Outils de base : attributs Data Cube
- Vocabulaires et modèles dédiés
  - DQV : Data Quality Vocabulary
  - SDMX / SIMS : standard de *quality reporting* d'Eurostat → n'existe pas en RDF

Data Quality Vocabulary

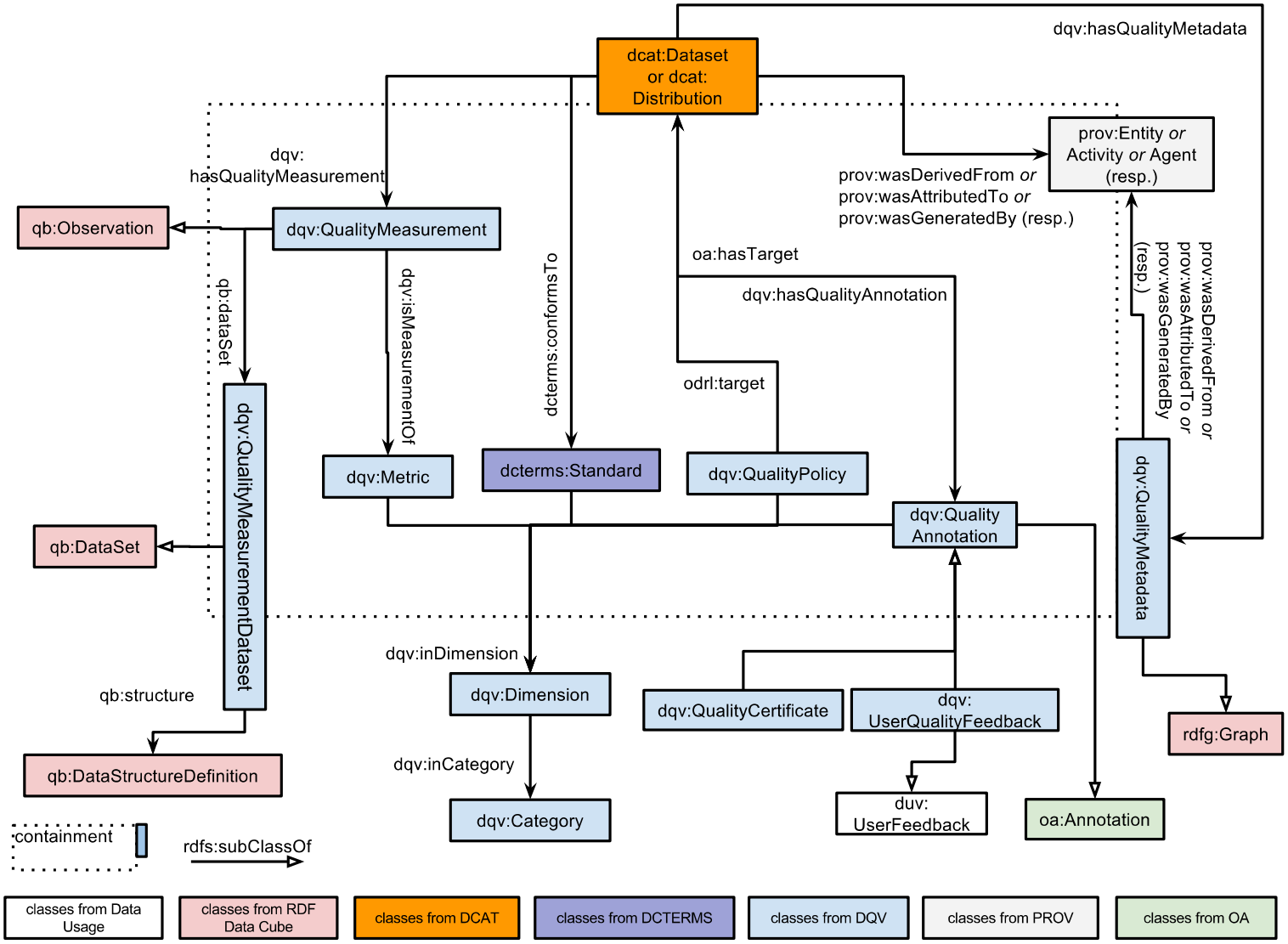
Définit des notions de base de qualité

- Catégories et dimensions
- Métriques et mesures
- Politiques, standards, certifications
- User feedback

Spécifie les jeux de métadonnées de qualité

- Articulation avec Data Cube et DCAT

Data Quality Vocabulary – Modèle



SDMX / SIMS

SIMS (Single Integrated Metadata Structure)

- Standard d'Eurostat pour le reporting sur la qualité
- Basé sur la partie « métadonnées » du modèle SDMX (SDMX-MM)
  - On ne modélise plus un cube de valeurs mais une hiérarchie d'attributs
  - Metadata Set et MetadataStructureDefinition

Travail en cours pour convertir SDMX-MM en vocabulaire RDF

- Selon les principes suivis par Data Cube
- Sera également utilisable pour d'autres utilisations de SDMX-MM
  - Reporting économique sur les indicateurs de développement durable

SDMX / SIMS : structure

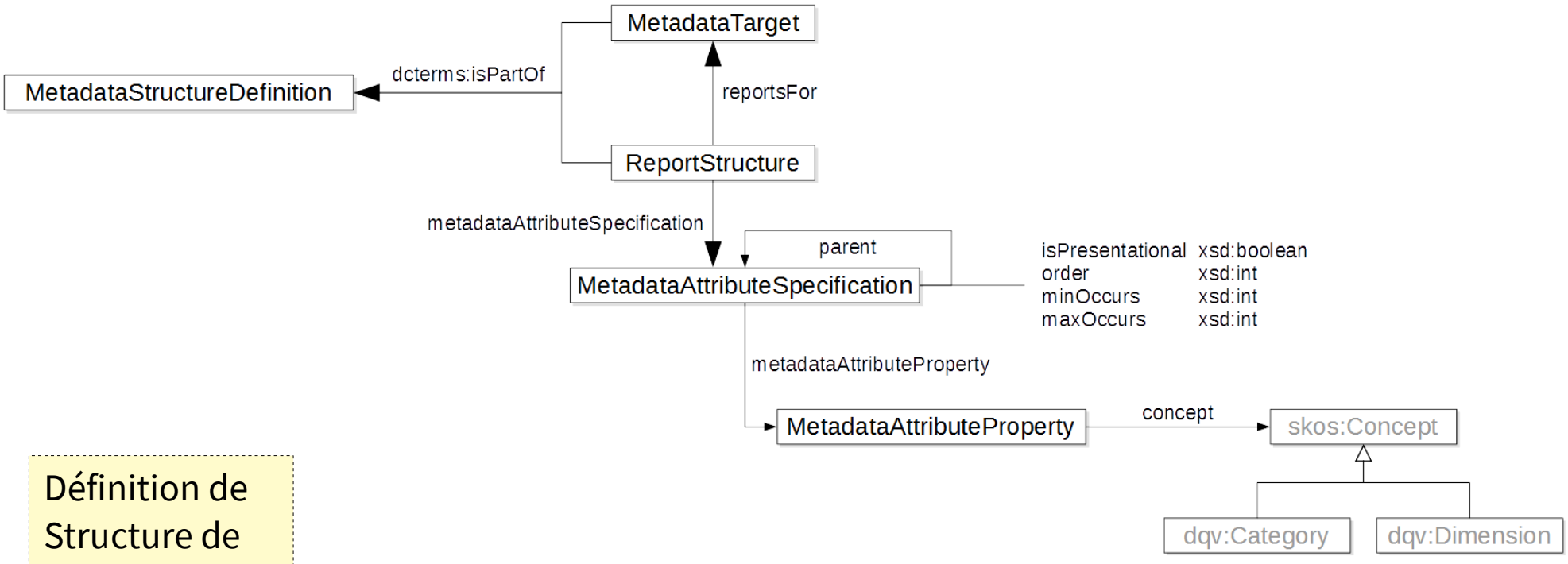
Item No	Concept name
S.1	**Contact**
S.1.1	Contact organisation
S.1.2	Contact organisation unit
S.1.3	Contact name
S.1.4	Contact person function
S.1.5	Contact mail address
S.1.6	Contact email address
S.1.7	Contact phone number
S.1.8	Contact fax number
S.2	**Metadata update**
S.2.1	Metadata last certified
S.2.2	Metadata last posted
S.2.3	Metadata last update
S.3	**Statistical presentation**
S.3.1	Data description
S.3.2	Classification system
S.3.3	Sector coverage
S.3.4	Statistical concepts and definitions
S.3.5	Statistical unit
S.3.6	Statistical population
S.3.7	Reference area
S.3.8	Time coverage
S.3.9	Base period
S.4	**Unit of measure**
S.5	**Reference period**
S.6	**Institutional mandate**
S.6.1	Legal acts and other agreements
S.6.2	Data sharing
S.7	**Confidentiality**
S.7.1	Confidentiality - policy
S.7.2	Confidentiality - data treatment
S.8	**Release policy**
S.8.1	Release calendar
S.8.2	Release calendar access
S.8.3	User access
S.9	**Frequency of dissemination**
S.10	**Accessibility and clarity**
S.10.1	News release
S.10.2	Publications
S.10.3	On-line database

Item No	Concept name
S.10.3.1	AC1. Data tables - consultations
S.10.4	Micro-data access
S.10.5	Other
S.10.5.1	AC 2. Metadata - consultations
S.10.6	Documentation on methodology
S.10.6.1	AC 3. Metadata completeness - rate
S.10.7	Quality documentation
S.11	**Quality management**
S.11.1	Quality assurance
S.11.2	Quality assessment
S.12	**Relevance**
S.12.1	User needs
S.12.2	User satisfaction
S.12.3	Completeness and R1. Data completeness - rate for U
S.12.3.1	R1. Data completeness - rate for P
S.13	**Accuracy and reliability**
S.13.1	Overall accuracy
S.13.2	Sampling error and A1. Sampling errors - indicators for U
S.13.2.1	A1. Sampling errors - indicators for P
S.13.3	Non-sampling error and A4. Unit non-response - rate for U and A5. Item non-response - rate for U
S.13.3.1	Coverage error
S.13.3.1.1	A2. Over-coverage - rate
S.13.3.1.2	A3. Common units - proportion
S.13.3.2	Measurement error
S.13.3.3	Non response error
S.13.3.3.1	A4. Unit non-response - rate for P
S.13.3.3.2	A5. Item non-response - rate for P
S.13.3.4	Processing error
S.13.3.5	Model assumption error
S.14	**Timeliness and punctuality**
S.14.1	Timeliness and TP2. Time lag - final results for U
S.14.1.1	TP1. Time lag - first results for P
S.14.1.2	TP2. Time lag - final results for P
S.14.2	Punctuality and TP3. Punctuality - delivery and publication for U
S.14.2.1	TP3. Punctuality - delivery and publication for P
S.15	**Coherence and comparability**
S.15.1	Comparability - geographical
S.15.1.1	CC1. Asymmetry for mirror flows statistics - coefficient
S.15.2	Comparability - over time and CC2. Length of comparable time series for U
S.15.2.1	CC2. Length of comparable time series for P

Item No	Concept name
S.15.3	Coherence- cross domain
S.15.3.1	Coherence - sub annual and annual statistics
S.15.3.2	Coherence- National Accounts
S.15.4	Coherence - internal
S.16	**Cost and burden**
S.17	**Data revision**
S.17.1	Data revision - policy
S.17.2	Data revision - practice and A6. Data revision - average size for U
S.17.2.1	A6. Data revision - average size for P
S.18	**Statistical processing**
S.18.1	Source data
S.18.2	Frequency of data collection
S.18.3	Data collection
S.18.4	Data validation
S.18.5	Data compilation
S.18.5.1	A7. Imputation - rate
S.18.6	Adjustment
S.18.6.1	Seasonal adjustment
S.19	**Comment**

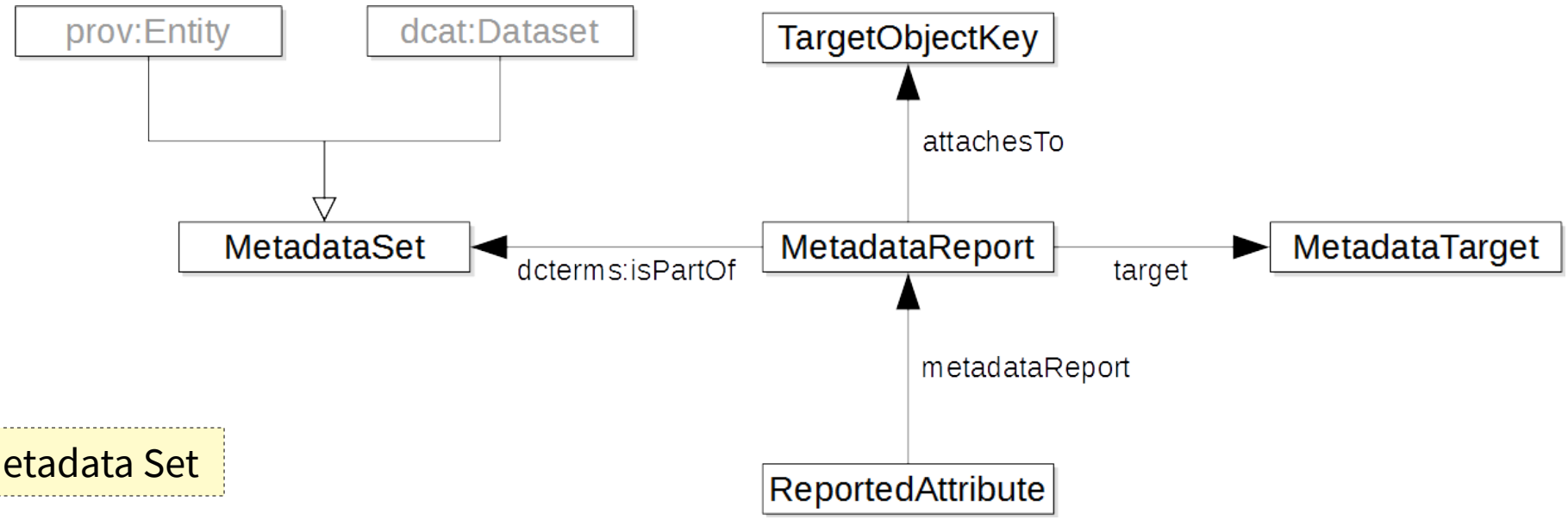
Eurostat

SDMX-MM – Modèle



Définition de Structure de Métadonnées

SDMX-MM – Modèle



Description des jeux de données : découvrir

Description des jeux de données : découvrir

Différentes solutions

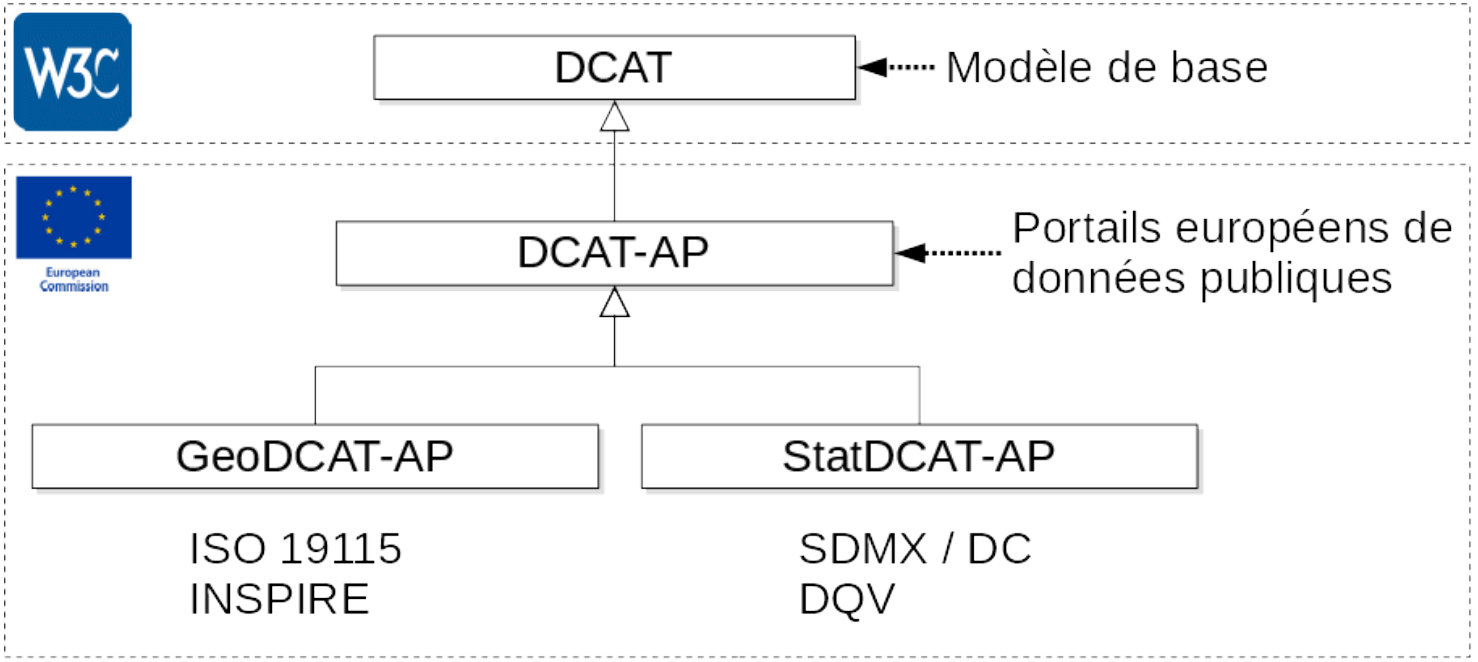
- Simples (Dublin Core)
- Ou plus complexes (DDI Disco)

Un profil DCAT spécialisé

- StatDCAT-AP
  - Produit du programme ISA² de la Commission Européenne
  - Basé sur la recommandation W3C DCAT
    - Un modèle pour les catalogues de données (liées ou pas) publiées sur le web
  - Adapté aux jeux de données statistiques
  - Extension du profil DCAT-AP (au même titre que GeoDCAT-AP)

Description des jeux de données découvrir

La famille DCAT

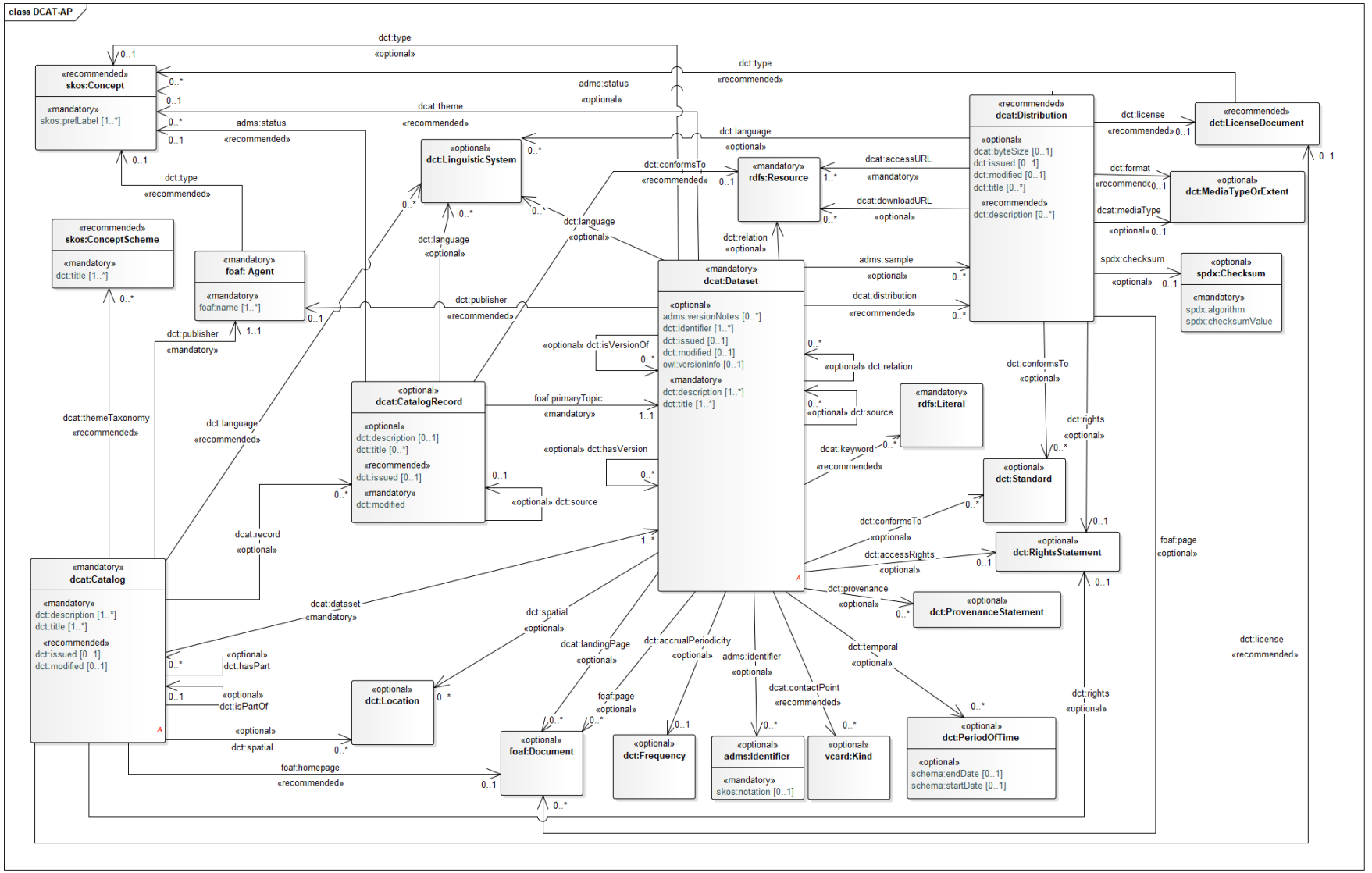


Description des jeux de données : découvrir

DCAT (Data Catalog Vocabulary)

- Utilisé par la plupart des portails de données ouvertes (European Data Portal, data.gov…)
- Utilisé également par des moteurs de recherche (Google Dataset Search)
- Modèle simple
  - Un catalogue contient des jeux de données
  - Un jeu de données peut avoir différentes distributions
- Actuellement en cours de révision
  - Par exemple, addition d'une classe DataService, alignement sur schema.org, etc.

DCAT-AP – Modèle



StatDCAT-AP

Définit des éléments de métadonnées qui permettent d'améliorer la « découvrabilité » des données

Ajoute notamment des information sur les dimensions et attributs

Les valeurs sont des instances de Data Cube DimensionProperty et AttributeProperty respectivement

Ajoute des annotations de qualité

Conformes au W3C Data Quality Vocabulary

Description des jeux de données : sourcer

Description des jeux de données : sourcer

Exemples

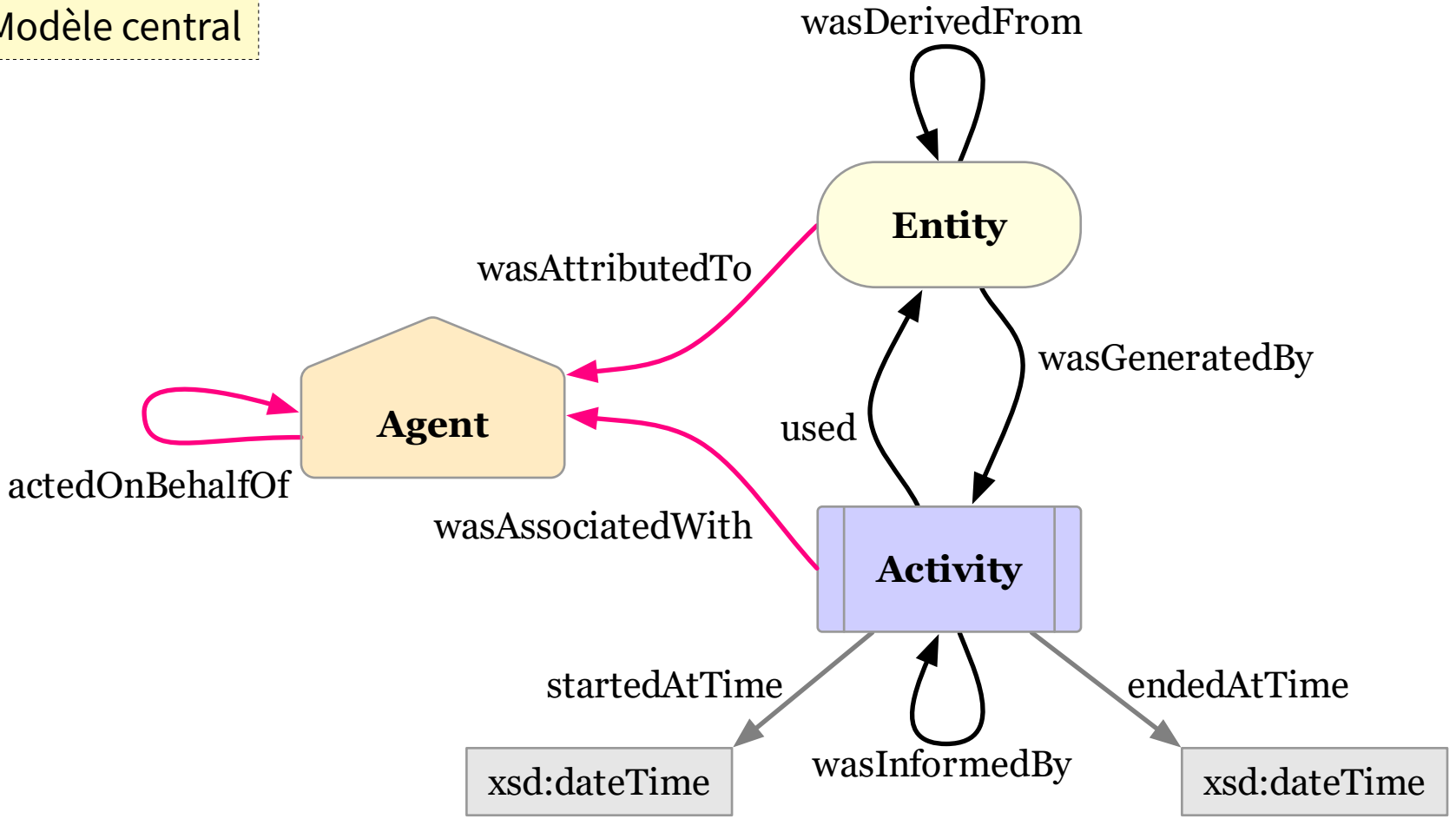
- Éditeur des données
- Traitements de production du jeu de données

Différentes solutions

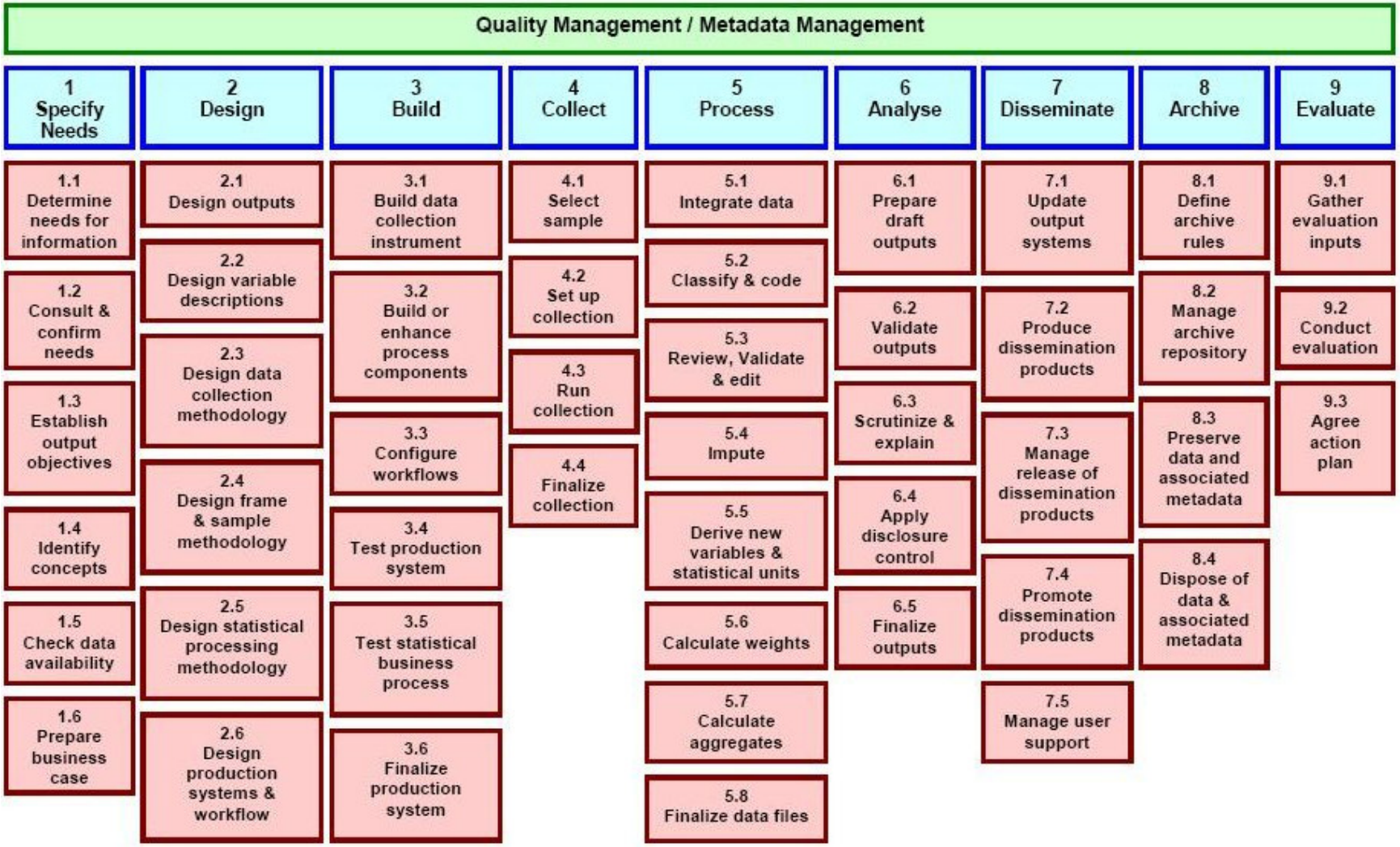
- Vocabulaire généraux
  - Base : Dublin Core, DCAT
  - Avancé : PROV-O
- Modèle métier
  - GSPBM : Generic Statistical Business Process Model
  - À transposer en RDF et à articuler avec PROV-O

PROV-O – Modèle

Modèle central



GSBPM – Modèle

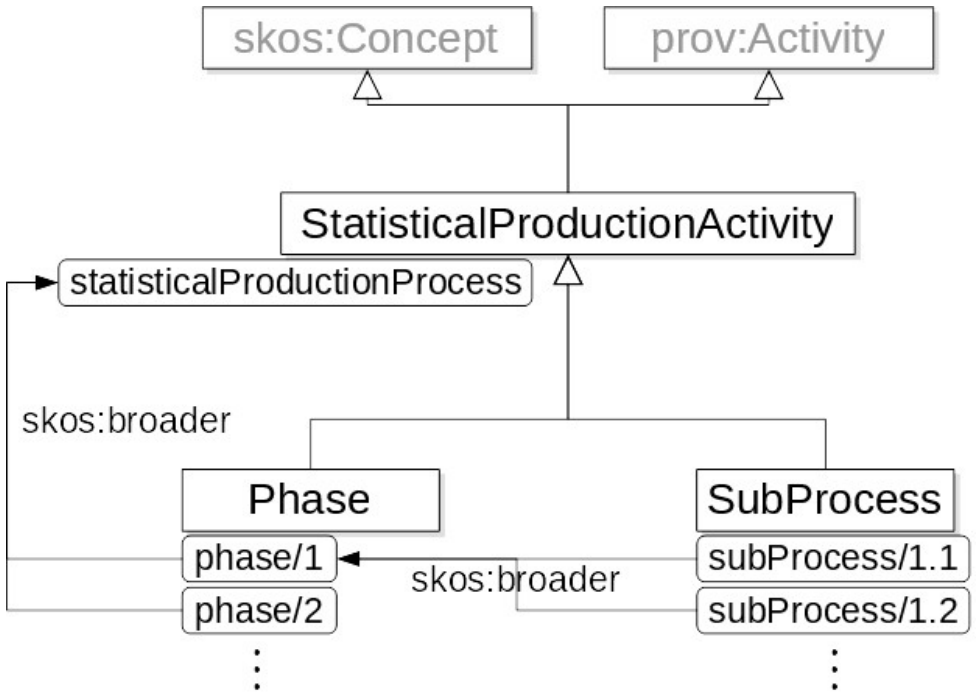


Articulation GSBPM - PROV-O

Définir une spécialisation de prov:Activity pour les activités de production statistique

Raffiner la normalisation internationale sur les activités statistiques

Activité de long terme conduite par l'UNECE



Description des jeux de données

Conclusions

Il reste un travail considérable pour définir des standards RDF permettant de décrire correctement les données statistiques

Ce travail est nécessaire pour publier des données interprétables

La statistique officielle s'installe (lentement) sur le web des (méta)données

Merci

Questions?