

Comment assurer la traçabilité des données numériques ?

Décrire la provenance

Le Web sémantique pour les données culturelles : bilan et perspectives
INHA, 12 décembre 2016

Claire Sibille – Service interministériel des archives de France



Sommaire

1 – Pourquoi fournir des informations de provenance ?

2 – Quels modèles pour la provenance ?

3 – Quelles stratégies de mise en œuvre ?

1 – Pourquoi fournir des informations de provenance ?

2 – Quels modèles pour la provenance ?

3 – Quelles stratégies de mise en œuvre ?

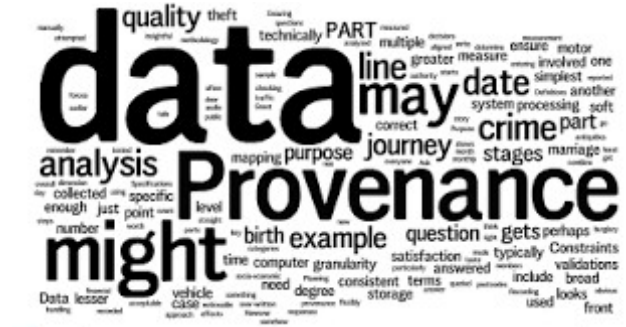
Gt 8 - Positionner le MCC comme expert sur la traçabilité des données numériques

Objectif :

→ Dresser un état de l'art sur les modèles permettant de reconstituer la provenance des données

Participants :

- Roselyne Aliacar, Francisca Cabrera, Ana Teixeira, MCC / SG / DPN
- Maya Khelifi, MCC / SG / DSM
- Alain Mille, LIRIS
- Florent Palluault, MCC/DGMIC/SLL
- Stéphane Reecht, Bibliothèque nationale de France
- Pascal Romain, Département de la Gironde
- Claire Sibille – de Grimoüard, SIAF (animatrice du groupe)
- Gwendoline Stab, Ministère des Affaires étrangères et équipe projet VITAM
- Édouard Vasseur, Ministère de la Défense et équipe projet VITAM



Comprendre comment les données ont évolué...



- ✓ Qui est responsable de leur création ?
 - ✓ Qui en est propriétaire ?
 - ✓ Comment ont-elles été modifiées depuis leur première version ?
 - ✓ Sont-elles affectées par d'autres données ?
 - ✓ Quels outils ont été utilisés pour générer chaque version ?
- Etc.

« La provenance est définie comme l'enregistrement des personnes, des institutions, des entités et des activités qui jouent un rôle dans la production, la modification et la mise à disposition de données. Les informations de provenance font partie des métadonnées contextuelles qui peuvent elles-mêmes devenir importantes en raison de leur propre provenance. » (Groupe d'incubation du W3C sur la provenance)

Toutes les métadonnées ne concernent pas la provenance.
→ Le titre d'un livre est une métadonnée mais ne donne pas d'information sur la provenance
→ La date de création, l'auteur, l'éditeur et les droits sur le livre donnent des informations sur sa provenance.

...mais ce n'est pas facile

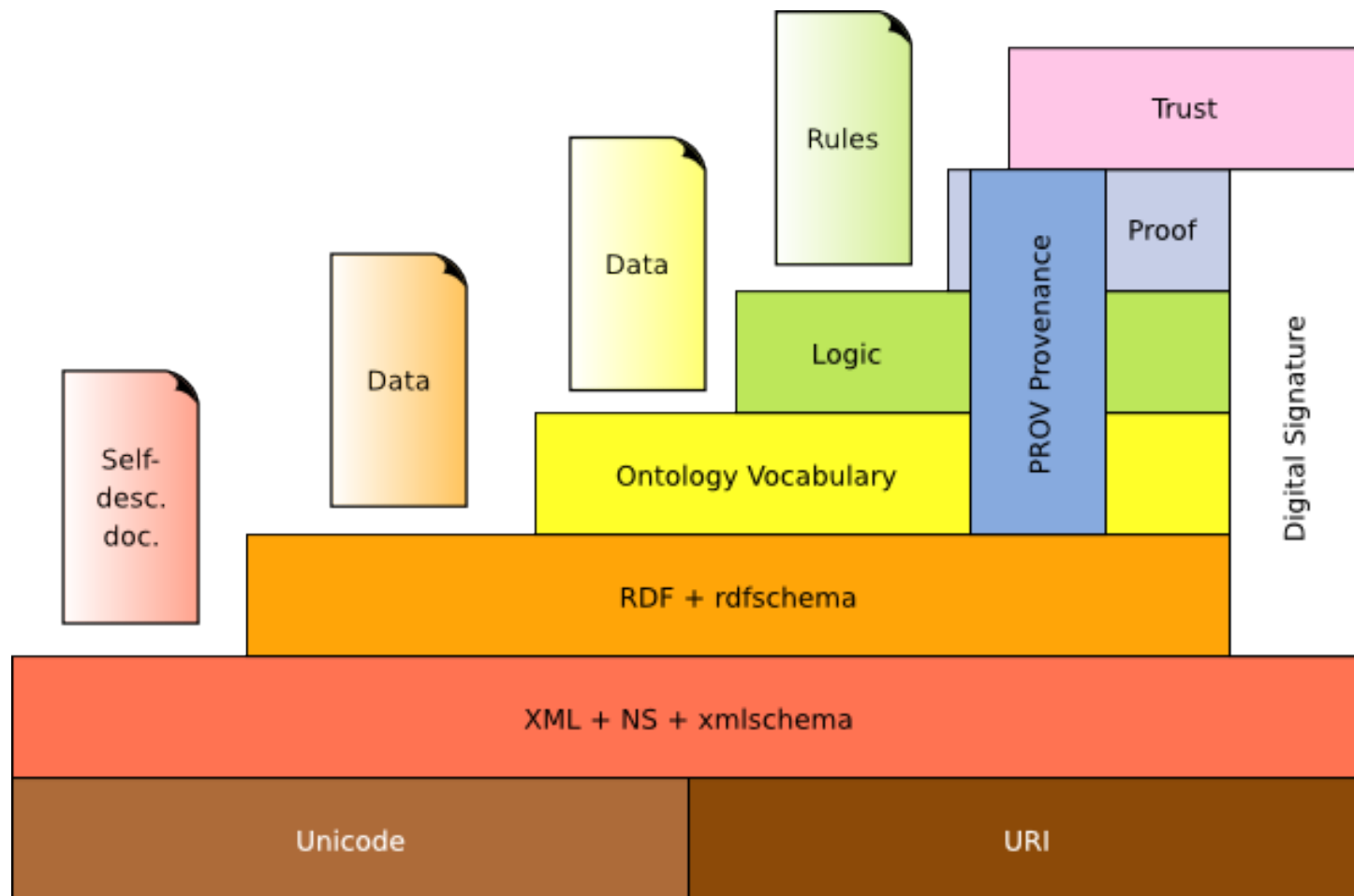
Comment identifier toutes les sources de provenance ?

Les données et modèles de provenance sont hétérogènes

Comment interroger les données de provenance avec des besoins métier différents ?

Cette problématique n'est pas nouvelle (architecture du Web sémantique)

La provenance dans le Web sémantique



Quels enjeux ?



Bénéfices

- Enjeux économiques, de protection des droits.
- Enjeux de gouvernance des données.
- Enjeux de positionnement du MCC comme tiers de confiance.

Risques à ne pas faire

- Frein au développement des industries culturelles.
- Dégradation de la qualité des données.
- Perte de visibilité des institutions culturelles.



Cas d'usage

- Qualité des données (ex. : « data journalisme » ou « journalisme de données »).
- Reproduction de processus (ex. : expérience scientifique).
- Conformité légale (ex. : en cas de conflit entre deux entités sur l'authenticité d'une donnée).

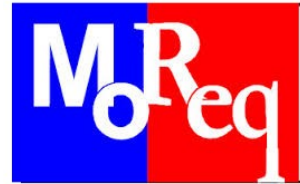
1 – Pourquoi fournir des informations de provenance ?

2 – Quels modèles pour la provenance ?

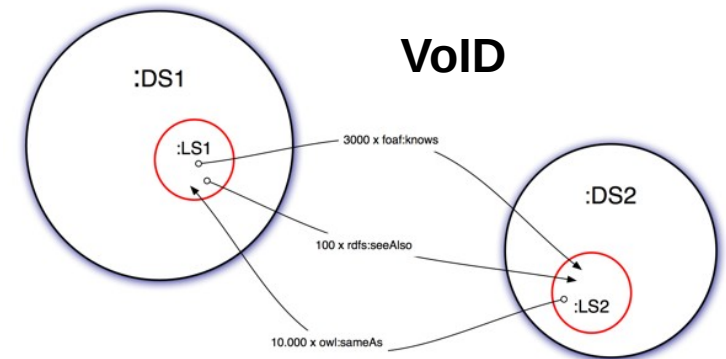
3 – Quelles stratégies de mise en œuvre ?

Les modèles et vocabulaires métier

- Dublin Core
- PREMIS
- ISAD(G), EAD
- Ontologie Provenir
- Open Provenance Model
- Etc.



Des modèles sémantiquement riches, mais difficilement réutilisables dans d'autres contextes



PROV n'est pas seulement un modèle



- PROV Overview (<http://www.w3.org/TR/prov-overview/>)
- PROV Primer (<http://www.w3.org/TR/prov-primer/>)
- PROV Data Model(*) (<http://www.w3.org/TR/prov-dm/>)
- PROV Constraints(*) (<http://www.w3.org/TR/prov-constraints/>)
- PROV Semantics (<http://www.w3.org/TR/prov-sem/>)

Modèle

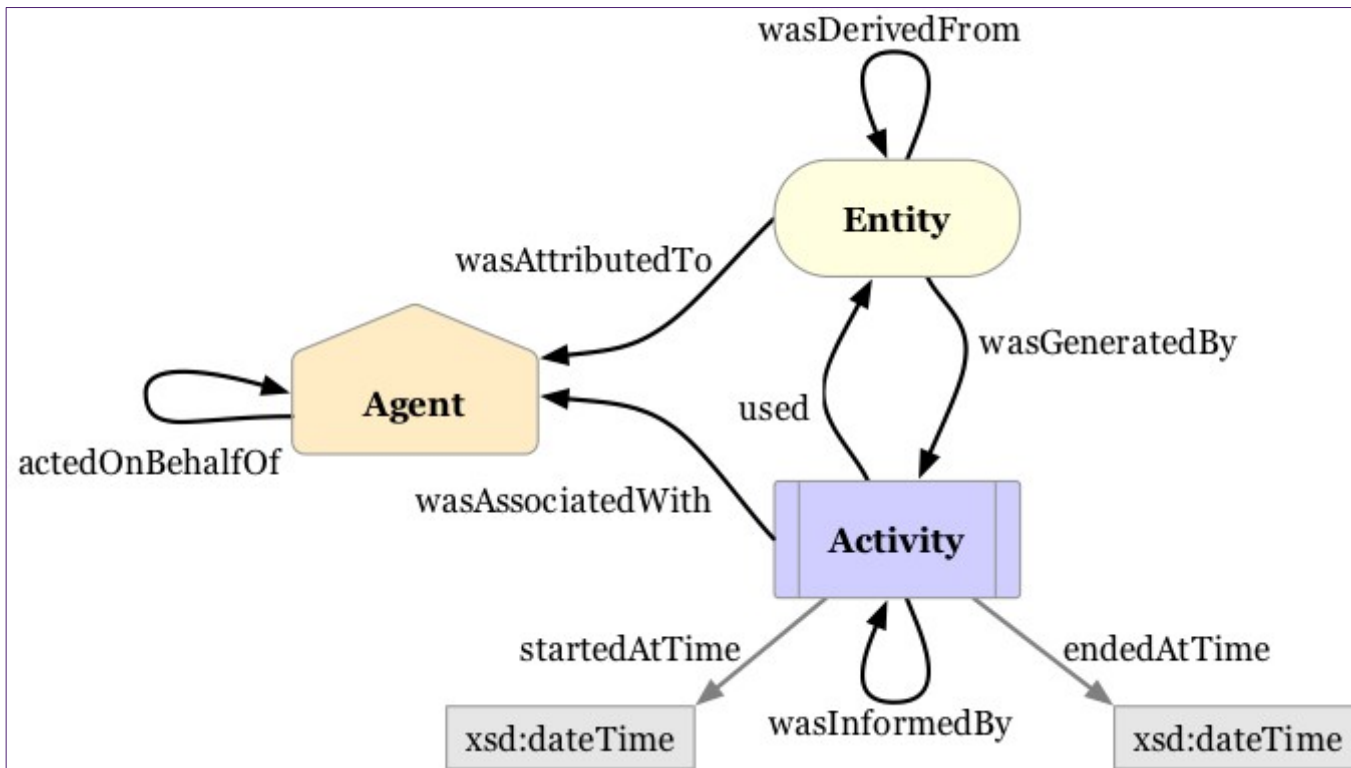
- PROV Notation(*) (<http://www.w3.org/TR/prov-n/>)
- PROV Ontology(*) (<http://www.w3.org/TR/prov-o/>)
- PROV XML Serialization (<http://www.w3.org/TR/prov-xml/>)
- PROV Access and Query (<http://www.w3.org/TR/prov-aq/>)

Sérialisations

- PROV DC Mapping (<http://www.w3.org/TR/prov-dc/>)
- PROV Links (<http://www.w3.org/TR/prov-links/>)
- PROV Dictionary (<http://www.w3.org/TR/prov-dictionary/>)
- PROV Implementations (<http://www.w3.org/TR/prov-implementations/>)

Extensions

PROV - Le modèle générique



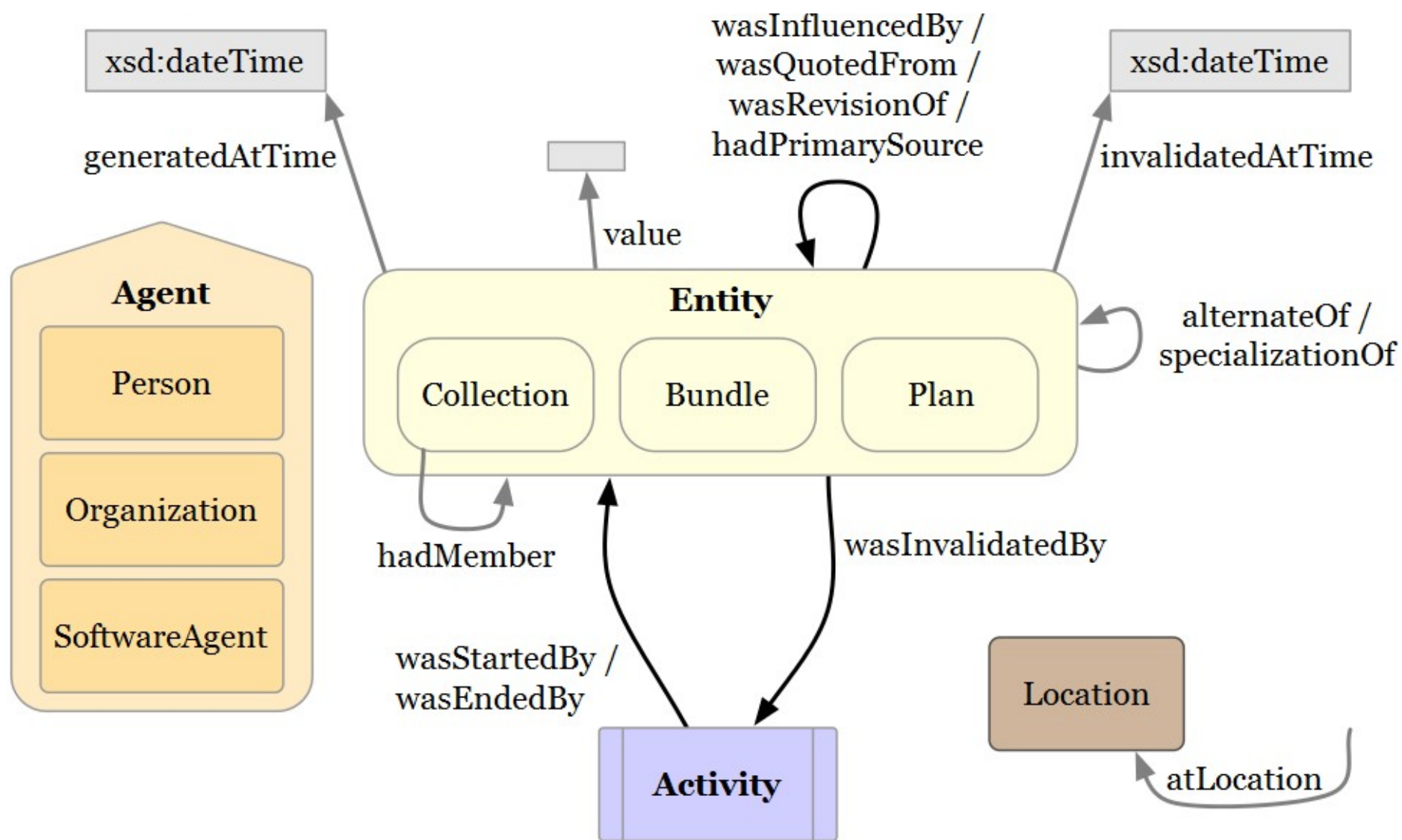
Le schéma ci-contre représente les trois principales entités du modèle PROV ainsi que les relations qu'elles peuvent avoir les unes avec les autres.

PROV – Les interactions entre entités

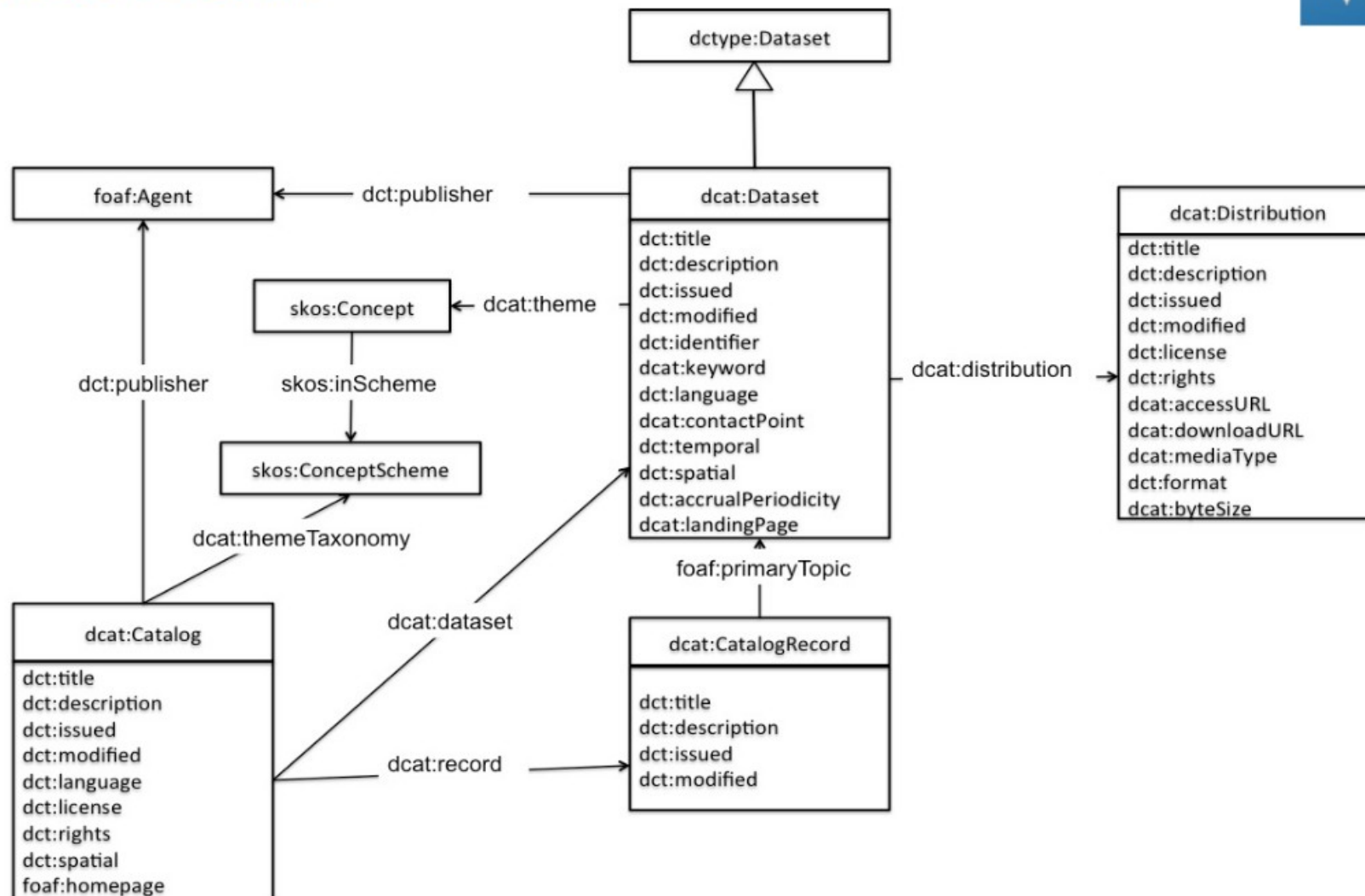


Concepts	Types ou relations	Dénomination
Entité	Types du modèle PROV	Entity
Activité		Activity
Agent		Agent
Génération	Relations du modèle PROV	WasGeneratedBy
Utilisation		Used
Communication		WasInformedBy
Dérivation		WasDerivedFrom
Attribution		WasAttributedTo
Association		WasAssociatedWith
Délégation		ActedOnBehalfOf

PROV – Le modèle étendu



W3C DCAT



1 – Pourquoi fournir des informations de provenance ?

2 – Quels modèles pour la provenance ?

3 – Quelles stratégies de mise en œuvre ?

Recommandations pour les producteurs de données

- ✓ Mettre en place des dispositifs pour auditer les systèmes d'information
- ✓ Choisir un modèle générique, extensible et interopérable
- ✓ Voir aussi les recommandations du **référentiel de bonnes pratiques** élaboré pour les producteurs de données ouvertes en Région Aquitaine
 - Les jeux de données sont accompagnés de liens ou de ressources d'animation et de support
 - L'identité et la qualité du responsable de chaque jeu de donnée est précisée
 - Chaque jeu de données est accompagné d'une fiche descriptive
 - Chaque jeu de données est accompagné d'une date de création
 - Etc.



le référentiel du projet SAEM
SYSTÈME D'ARCHIVAGE ÉLECTRONIQUE MUTUALISÉ

Département de la Gironde • Bordeaux Métropole • Ville de Bordeaux

Notices d'autorité

Vocabulaires

Profils SEDA

Unités d'archives

Search

s'identifier

- **Système d'archivage électronique mutualisé (SAEM)**
 - Un projet partenarial regroupant depuis 2012 le département de la Gironde, Bordeaux Métropole et la ville de Bordeaux
 - Réalisation d'un référentiel de gestion et de structuration des métadonnées des documents électroniques
 - Gestion des événements du cycle de vie
 - Enregistrement des événements liés à chaque entité (implémentation du modèle PROV)



le référentiel du projet SAEM SYSTÈME D'ARCHIVAGE ÉLECTRONIQUE MUTUALISÉ

Département de la Gironde • Bordeaux Métropole • Ville de Bordeaux

[s'identifier](#)[Notices d'autorité](#)[Vocabulaires](#)[Profil SEDA](#)[Unités d'archives](#)

Types de relation > used (a utilisé)

used (a utilisé)

description

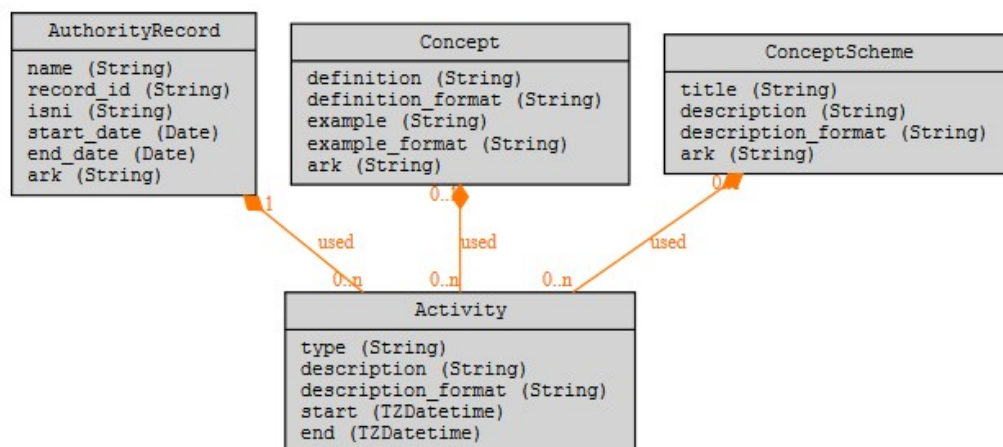
Relate an Activity to an other entity, which is used as the source of the activity (usually for the production of an other entity through the 'generated' relation).

symétrique

non

mise en ligne

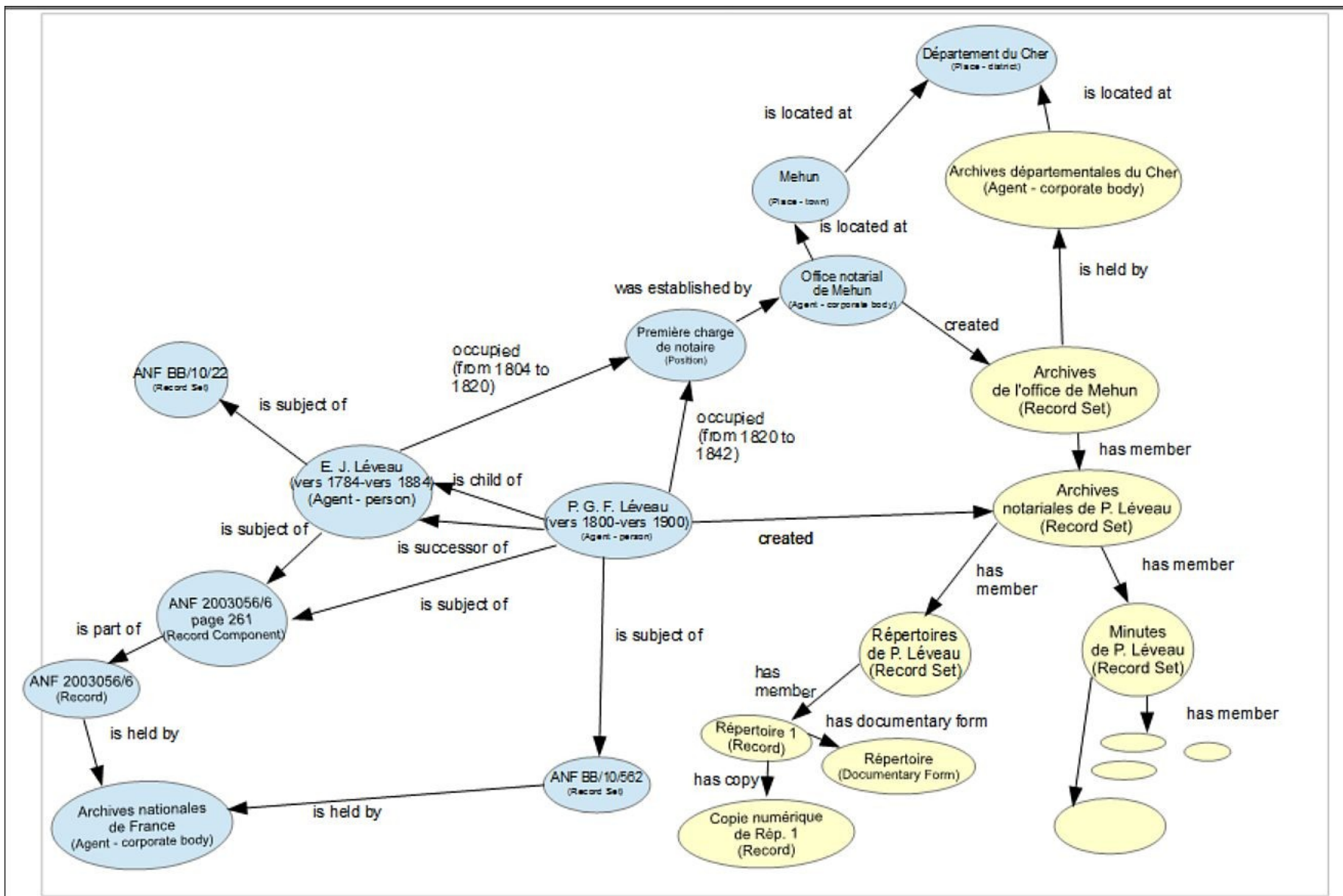
oui



relation (3)	cardinalité	contraintes	options
Activity used AuthorityRecord	0..1 0..n		mise en ligne, composite=objet
Activity used Concept	0..1 0..n		mise en ligne, composite=objet
Activity used ConceptScheme	0..1 0..n		mise en ligne, composite=objet

Type de relation #5316 - créé le 14/11/2016

Records in Context (Ric)



Relation Number	Relation Domain	Relation Name	Relation Range	Inverse Relation Name and Number	Comments
RiC-R231	Agent	wrote	Record Component	was written by (RiC-R96)	
RiC-R232	Agent	accumulated	Record Set	was accumulated by (RiC-R155)	
RiC-R233	Agent	arranged	Record Set	was arranged by (RiC-R157)	
RiC-R234	Agent	assembled	Record Set	was assembled by (RiC-R158)	
RiC-R235	Agent	collected	Record Set	was collected by (RiC-R160)	
RiC-R236	Agent	created	Record Set	was created by (RiC-R161)	
RiC-R237	Agent	is associated with	Record Set	is associated with (RiC-R151)	See also RiC-R247
RiC-R238	Agent	is holder of	Record Set	is held by (RiC-R152)	See also RiC-R248