Documentation – Base de données de Varagnes

Documentation fournie par Thomas Chaineux, en fin de stage à l’Association des Amis de la Fondation Seguin (3 avril – 28 juillet 2023)

**Table des matières**

[1) Omeka-S et le RDF 2](#__RefHeading___Toc944_2863259247)

[2) Les données d’inventaire 4](#__RefHeading___Toc946_2863259247)

[3) Les données d’indexation 5](#__RefHeading___Toc948_2863259247)

[4) Les *resource templates* 6](#__RefHeading___Toc950_2863259247)

[5) Modélisation des données par *template* 7](#__RefHeading___Toc952_2863259247)

[5.1) Notes sur des données communes 8](#__RefHeading___Toc954_2863259247)

[5.2) Liens entre items 8](#__RefHeading___Toc956_2863259247)

[5.3) Modélisation de données d'inventaire: Archives 10](#__RefHeading___Toc958_2863259247)

[5.4) Modélisation de données d'inventaire: Patrimoine artistique 10](#__RefHeading___Toc974_2863259247)

[5.5) Modélisation de données d’indexation: Lieux 11](#__RefHeading___Toc1001_2863259247)

[5.6) Modélisation de données d'inventaire: Instruments 12](#__RefHeading___Toc976_2863259247)

[5.7) Modélisation de données d'inventaire: Minéralogie 12](#__RefHeading___Toc995_2863259247)

[5.8) Modélisation de données d’inventaire/d’indexation: Livres 13](#__RefHeading___Toc997_2863259247)

[5.9) Modélisation de données d’indexation: Personnes 14](#__RefHeading___Toc999_2863259247)

[5.10) Modélisation de données d’indexation: Sociétés 15](#__RefHeading___Toc1003_2863259247)

[5.11) Modélisation de données d’indexation: Thématiques 15](#__RefHeading___Toc1005_2863259247)

[5.12) Modélisation de données de médias: Images 16](#__RefHeading___Toc1018_2863259247)

[6) Visualiser les données 16](#__RefHeading___Toc1423_2863259247)

[6.1) Généralités: les fiches d’item 16](#__RefHeading___Toc1425_2863259247)

[6.2) Parcourir les collections 17](#__RefHeading___Toc1427_2863259247)

[6.2.1) Via les *item sets* de données d’indexation 17](#__RefHeading___Toc1429_2863259247)

[6.2.2) Via les *item sets* d’inventaire (hors Archives) 18](#__RefHeading___Toc1431_2863259247)

[6.2.3) Le cas des archives 19](#__RefHeading___Toc1453_2863259247)

[6.2.4) Formulaire de recherche 20](#__RefHeading___Toc3568_2842438568)

[7) Mises à jour de la base 21](#__RefHeading___Toc3570_2842438568)

[7.1) Mise à jour manuelle 21](#__RefHeading___Toc3572_2842438568)

[7.1.1) Modification, une *ressource* à la fois 21](#__RefHeading___Toc3574_2842438568)

[7.1.2) Modification par « batch » 22](#__RefHeading___Toc3576_2842438568)

[7.1.3) Création de nouveau *items* 23](#__RefHeading___Toc3578_2842438568)

[7.2) Mise à jour de masse – CSV Import 23](#__RefHeading___Toc3580_2842438568)

[8) Exporter les données 25](#__RefHeading___Toc3986_2842438568)

[9) Perspectives d’amélioration 25](#__RefHeading___Toc3988_2842438568)

[9.1) Normaliser les données au maximum 25](#__RefHeading___Toc3990_2842438568)

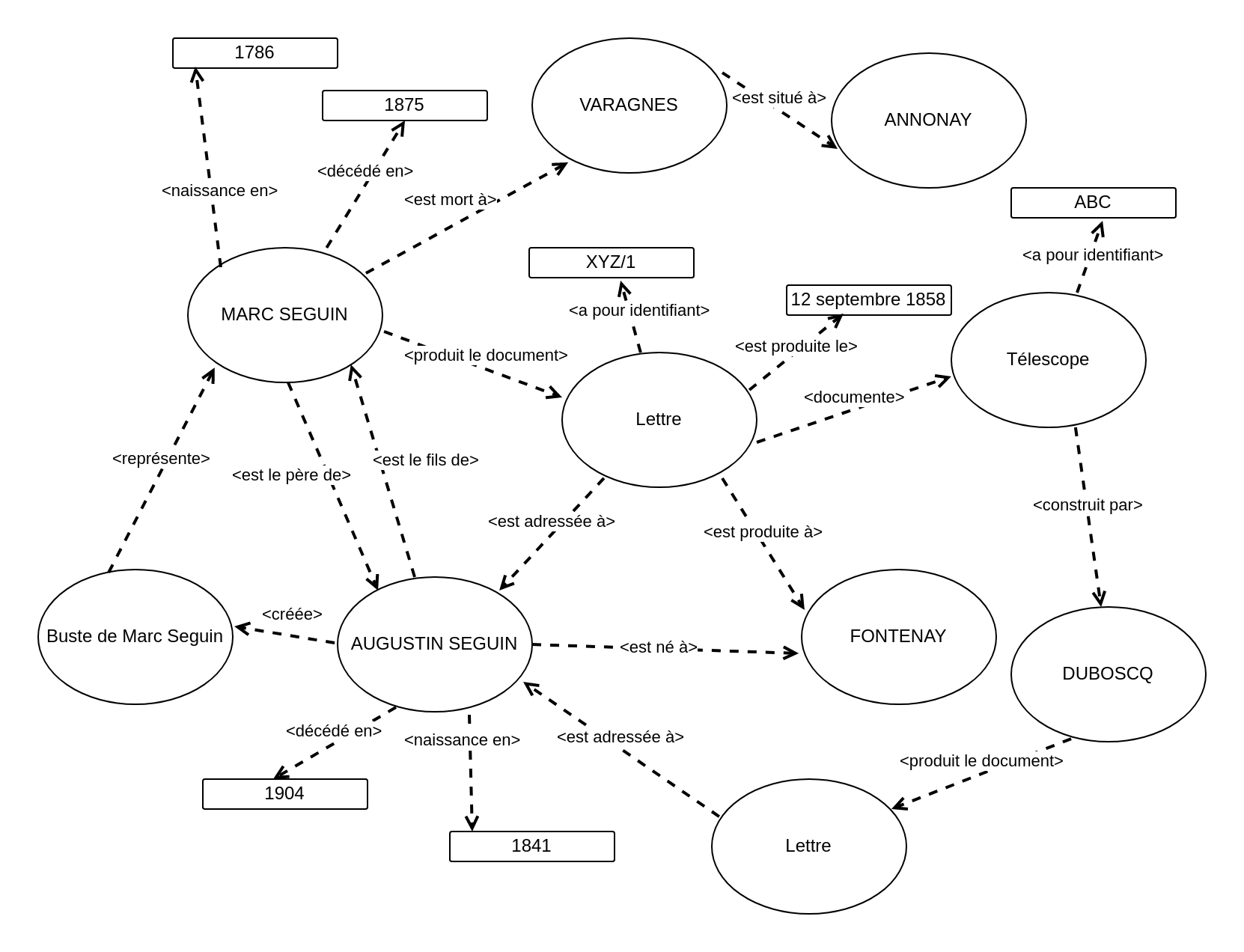
[9.2) Poursuivre le travail d’indexation 26](#__RefHeading___Toc3992_2842438568)

# 1) Omeka-S et le RDF

La base de données est crée avec le logiciel Omeka-S, logiciel open source réunissant des outils de création de base de données et de création de contenu web. Bien que les données soient stockées en SQL en arrière-plan, le modèle logique est un modèle en **graphe**, qui utilise le format RDF.

Le RDF a la particularité de faire appel a des ***triplets***, qui seront utilisés pour décrire une ressource. Ils s’articulent de manière relativement simples : ils sont constitués d’un ***sujet***, d’un *prédicat* (nous utiliserons plutôt le terme de ***propriété***) et d’un ***objet***, de même qu’une phrase simple comporte un sujet, un verbe et un complément (ex. : « Marc Seguin (sujet) a pour enfant (propriété) Augustin Seguin (objet) »). La multiplication des triplets ayant un même sujet va permettre d’accumuler les informations sur ce sujet. De plus, une ressource peut également être un objet, ce qui va compléter les informations que nous avons sur lui.

Voir cet exemple fictif, qui illustre cette modélisation des connaissances en graphe :



Dans cet exemple, chaque bulle est une information qu’il est possible de documenter avec autant de triplets que nécessaire. Les propriétés vont les relier à d’autres bulles (pouvant elles aussi recevoir des propriétés) ou renseigner de l’information brute (les rectangles : des dates, des numéros d’inventaire, mais aussi des commentaires, des descriptions physiques, des dimensions…).

Potentiellement, toute les bulles du schéma ci-dessus peuvent être représentés par une ressource. En pratique, nous devons nous limiter au champ des collections, car les connexions de ce type peuvent potentiellement organiser le savoir du monde entier. Par exemple, nous avons inclus le nom du fabricant Jules Duboscq: il apparaît d’une manière ou d’une autre dans les collections, mais ses éléments biographiques sont moins essentiels à la compréhension de celles-ci. Nous pouvons tout de même l’identifier de manière sûre dans des référentiels extérieurs qui centralisent des données biographiques, comme par exemple Wikidata, le VIAF, les autorités de la BnF… Les liens vers ces fiches d’autorités sont prévus pour être permanents : ce sont les **URIs**.

Jules Duboscq sur : VIAF : <http://viaf.org/viaf/95912279>

Wikidata : <https://www.wikidata.org/wiki/Q165282>

Autorités BnF : <http://ark.bnf.fr/ark:/12148/cb14888760g>

Quant aux propriétés, elles sont constituées en « langages » adaptés à différents usages : ces langages sont appelés **ontologies**. Les critères pour décrire (par exemple) une personne ne sont pas les mêmes que pour une pièce d’archive, et la pièce d’archive se décrit elle-même de manière différente d’un instrument scientifique. Les ontologies ont pour but de couvrir un champ de description spécifique. En pratique, il nous faudra aller « piocher » dans plusieurs ontologies différentes pour couvrir tous nos besoins de description.

Une ontologie ne définit pas uniquement les propriétés : elle délimite également la nature des sujets et des objets qui peuvent être reliés entre eux par les propriétés. Pour prendre un exemple simple, la propriété <est né à> exige que le sujet soit une personne, et que l’objet soit un lieu. La nature des sujets est appelée une **classe**. Une ontologie définit obligatoirement les classes, afin de structurer l’usage de ses propriétés.

Omeka-S utilise également quelques termes qui lui sont propres, et qu’il convient de citer ici :

- ***item***: il s’agit de l’unité essentielle des données d’Omeka-S. Chaque ressource est créée en tant qu’*item*, quelque soit sa classe ou ce qu’elle décrit. Par exemple : la fiche de Marc Seguin, la fiche de la pièce d’archive A25/1, de l’instrument SEGUIN-0244…

- ***item set***: c’est un regroupement d’*items*, qui sert donc à les ordonner selon de grandes catégories. Chacun est un type de fiche, à laquelle sont liés les *items* qui en dépendent.

- ***média***: des fichiers multimédias (photos, vidéos, sons, …) qui sont liés à des *items* déjà existant.

- ***resource***: c’est le terme générique pour désigner indistinctement un *item, item set* ou un *media*.

# 2) Les données d’inventaire

Les données compilées au fur et à mesure des inventaires doivent donc être modélisées selon ces principes de triplets. On le comprend, le **RDF implique des données très structurées** : il est essentiel que le sujet d’une propriété <est créé à> soit un lieu, ou que <numéro d’édition> indique bien s’il s’agit de la 1ère, 2ème ou 3ème édition d’un ouvrage. Dans le cas du patrimoine de Varagnes, le défi est de trouver le bon équilibre entre des données qui peuvent être exprimées de la même manière, et celles qui doivent être exprimées selon des propriétés adaptées à des ressources spécifiques. Le fait que les données représentent des objets très divers et qu’elles aient été produites par des personnes différentes dans des cadres différents complexifient la tâche. Pour rappel, les différents inventaires sont :

1) un fichier Excel d’inventaire des archives des sous-fonds A et B (Archives des générations de Marc et Augustin Seguin), dressé par Lisa Lafontaine

2) un fichier Excel avec l’inventaire des cartes et plans, dressé par Marie-Françoise Crouch

3) un fichier Excel avec l’inventaire des factures, dressé par Lisa Lafontaine (avec l’appui de Marie-Françoise Crouch)

4) un fichier Excel avec l’inventaire des plaques de verre photographique, dressé par Lisa Lafontaine

5) le fichier XML d’un inventaire du patrimoine scientifique et technique réalisé par Quentin Rodriguez, assistant de conservation du patrimoine, et Nathalie Vidal, attachée de conservation du patrimoine et responsable du département histoire des sciences et techniques. Des photos en basse définition sont disponibles et sont/seront insérées dans la Base.

6) un fichier Excel avec l’inventaire artistique et mobilier, dressé par Elizabeth Dandel de l’agence Art2 Conseil, destiné à appuyer la demande de classement de l’ensemble des collections et objets de la famille Seguin au domaine de Varagnes. Des photos sont disponibles, mais les noms de fichiers ne permettent pas de retrouver directement les pièces qu’elles montrent (travail de recoupement à prévoir).

7) un fichier Excel d’un inventaire de pièces de minéralogie, dressé par Taya Flaherty. Là aussi, des photos sont disponibles, mais les noms de fichiers ne permettent pas de retrouver directement les pièces qu’elles montrent.

8) un fichier Excel des ouvrages présents en bibliothèque, dressé par Lisa Lafontaine et Marie-Françoise Crouch.

Les fichiers 2, 3 et 4 ont été fusionnés avec le 1er. Ils sont donc inclus dans les « Archives » de manière générale (tant dans la base de données que dans le fichier EAD[[1]](#footnote-2)), tout comme ils avaient été inclus dans l’inventaire analytique des archives, sorti en PDF par Lisa Lafontaine.

**Nous recommandons de ne plus maintenir ces inventaires à jour** et de ne les conserver qu’à titre de consultation. En effet, les données de ces inventaires ont été transformées et croisées de nombreuses fois pour être intégrées dans la base.

Dans Omeka-S, les données d’inventaires sont contenues dans des *item sets* qui ont « Collection » définie comme classe :



# **3) Les données d’indexation**

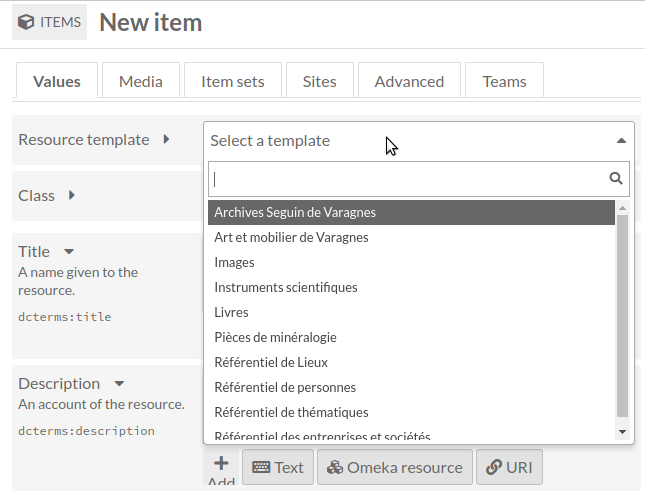
Dans la base d’Omeka-S, nous n’allons pas rencontrer uniquement des données de description des pièces d’inventaires. En effet, nous avons également mis en place un système qui créé des fiches de références qui serviront à leur indexation. Ces fiches de références sont de plusieurs types :

|  |  |  |
| --- | --- | --- |
| **Nom du référentiel** | **Description** | **Nbre de fiches** |
| Famille et collaborateurs | Référentiel des personnes, qu’il s’agisse de la famille, des collaborateurs ou des personnes illustres rencontrées dans les archives. La liste a été établie avec Jean-Marc, et sert à indexer les auteurs et les destinataires notables. | 59 |
| Sociétés et industries autour des Seguin | Les sociétés (y compris sociétés savantes) qui sont intervenues dans l’histoire des Seguin. Servent à indexer les documents qui témoignent de leurs activités, et le patrimoine technique produit par elles. | 26 |
| Ouvrages, contributions et travaux de Marc Seguin | Les écrits de Marc Seguin, qu’ils aient été publiés ou non. Il s’agit bien de l’œuvre intellectuelle et non d’un ouvrage physique. Permet d’indexer un document lorsqu’il s’agit de notes préparatoires à l’ouvrage, lorsqu’il s’agit d’un exemplaire, ou lorsqu’il le documente de manière plus générale | 20 |
| Lieux spécifiques aux Seguin | Les lieux trop spécifiques aux Seguin pour apparaître dans les référentiels géographiques. On y retrouve Varagnes, l’Observatoire, le Laboratoire, ainsi que Saint-Marc. Sert à l’indexation par lieu de réception. Varagnes, l’observatoire et la laboratoire sont aussi des sujets de Thématiques. | 4 |
| Thématiques | Thématiques générales rencontrées dans les archives, sur base du travail d’indexation de Jean-Marc. Les sujets sont larges : disciplines scientifiques, commémorations, voyages, art… Les Thématiques « Buire » et « Gnome et Rhône » centralisent les documents liés à ces entreprises qui ont connu de nombreuses transformations au cours de leur histoire. | 59 |

Dans Omeka-S, ces données sont contenues dans des *item sets* qui ont « Dataset » pour classe :



# 4) Les *resource templates*

La modélisation des données peut être complexe, mais heureusement Omeka-S permet de créer des ***resource templates***: ce sont en quelque sorte des « formulaires », qui contiennent tout une série de propriétés que nous avons présélectionnées pour répondre aux besoin de description d’un type de ressource. Cela facilite la création de données, mais permet aussi de s’assurer que deux *items* similaires soient décrit de manière similaire. Nous pouvons également paramétrer un affichage alternatif d’une propriété (pour afficher «Est préparatoire à » plutôt que *rico:isDraftOf*, par exemple). La sélection du *template* se fait à la création de la fiche :

Les différents *resource* *templates* sont visibles dans la section homonyme d’Omeka-S :



A chaque ajout dans la base de données, le *resource template* correspondant au type de pièce **doit** être utilisé. Cela permet de normaliser la description au maximum (bien que la plupart des champs ne sont pas obligatoires). Chaque *template* décrivant un type de ressource, et chaque type de ressource ayant son *item set*, la plupart des *templates* correspondent donc à un *item set* (à l’exception des images et des livres) :

|  |  |  |
| --- | --- | --- |
| **Type de données** | ***Resource template*** | ***Item set*** |
| Item (données d’inventaire) | Archives Seguin de Varagnes | Archives Seguin de Varagnes |
| Item (données d’inventaire) | Art et mobilier de Varagnes | Patrimoine artistique et mobilier de Varagnes |
| Item (données d’inventaire) | Instruments scientifiques | Instruments scientifiques et patrimoine technique de Varagnes |
| Item (données d’inventaire) | Pièces de minéralogie | Collection de minéralogie de Varagnes |
| Item (données d’inventaire et donnée de collection) | Livres | Ouvrages, contributions et travaux de Marc Seguin /  Bibliothèque de Varagnes |
| Item (données d’indexation) | Référentiel de Lieux | Lieux spécifiques aux Seguin |
| Item (données d’indexation) | Référentiel de personnes | Famille et collaborateurs |
| Item (données d’indexation) | Référentiel de thématiques | Thématiques |
| Item (données d’indexation) | Référentiel des entreprises et sociétés | Sociétés et industries autour des Seguin |
| Médias | Images | [aucun – se rattache aux items] |

Il est bien sûr possible de créer de nouveaux *templates*, mais chaque ressource ne pourra suivre qu’un seul *template* à la fois.

|  |
| --- |
| **Note importante**  Ces *resource template* ne sont pas figés et peuvent être édités si, lors d’accroissements futurs, une propriété paraît intéressante pour décrire une ressource mais qu’elle n’apparaît pas dans le *template.*  Lors de la création d’une ressource, il ne faut en tout cas **jamais** ajouter une propriété directement sur la page de création de la ressource. Il faut d’abord mettre le *template* à jour en incluant la propriété souhaitée, et puis seulement créer la pièce selon le *template* désormais mis à jour. De cette manière, nous nous assurons que les mêmes propriétés seront utilisées pour les mêmes données.  Attention également à ne pas multiplier des propriétés à usage unique dans le *template*. Avant de l’éditer, il convient de vérifier si la donnée ne peut pas rentrer dans une propriété déjà existante. |

# **5) Modélisation des données par *template***

Nous allons maintenant montrer les propriétés retenues pour chaque *template*. Le nom simplifié de celui-ci est au centre, dans le cadre noir ; suivent, dans les cadres gris, le nom des propriétés et une note sur la classe ; puis, le nom d’affichage de la propriété (soit tel qu’elle apparaît sur l’interface d’Omeka-S, indépendamment de sa dénomination informatique) ; et enfin, le type de données que cette propriété attend.

Le type de données attendus peut être :

- TXT : un champ texte.

- Item OS : un lien vers un *item* créé précédemment dans Omeka-S.

- URI : une URI, avec, entre parenthèses, une spécification sur le référentiel utilisé.

- Date (ISO 8601) : une date, mise en forme selon ce format standard[[2]](#footnote-3).

## ***5.1) Notes sur des données communes***

Si chaque *template* détermine les propriétés à utiliser pour chaque type de ressource, certaines propriétés sont communes à plusieurs d’entre eux. Dans la plupart des cas, il ne faudra pas se poser trop de questions, et simplement suivre les propriétés du *template*. Nous avons cependant certaines remarques.

Tout *item* **doit** posséder un identifiant **unique** qui est l’objet de la propriété *dcterms:identifier*. C’est le meilleur moyen de désambiguïser une pièce d’une autre. Pour les *items* d’inventaire, il s’agit de son numéro d’inventaire ; pour les *items* d’indexation, c’est un indentifiant que nous avons construit en le faisant commencer par REF.

Chaque *resource template* spécifie également l’objet de la propriété qui servira de titre, qui sera affiché dans la base. Pour les archives, il s’agit de la propriété *dcterms:description*; pour les autres ressources, il s’agit de la propriété *dcterms:title*. Cela peut être pris en compte dans la manière de formuler les objets de ces deux propriétés. **La propriété *dcterms:title* est donc obligatoire dans l’essentiel des cas**.

Toutes les dates relatives à l’origine de la pièce (date de production d’une pièce d’archives, date de fabrication d’un instrument, date de publication d’un livre, date de création d’une œuvre) sont reprises sous la propriété *dcterms:created*. Il est préférable de garder cette propriété spécifiquement pour décrire ces données. Les autres éléments de dates prennent d’autres propriétés, afin de les différencier de la création des pièces et ainsi de ne pas « polluer » les résultats de recherche dans les collections avec des données non-historiques. Ainsi, des dates d’inventaire sont reprises sous la propriété *dcterms:dateSubmitted*, celles de prise de vue d’une photo avec *dcterms:issued*, etc.

Une précision également sur l’emploi des duos de propriétés *rico:isOrWasIncludedIn*/ *rico:includesOrIncluded* d’une part, et *dcterms:isPartOf*/*dcterms:hasPart*. Si les deux peuvent être employés de manière similaire, nous avons opéré une distinction qui devra être conservée pour de futurs ajouts :

- *rico:isOrWasIncludedIn*/ *rico:includesOrIncluded* est utilisé pour décrire des niveau de classements, résultat d’une opération de hiérarchisation intellectuelle :

<dossier d’archives> *rico:includesOrIncluded* <pièce d’archives>

- *dcterms:isPartOf*/*dcterms:hasPart* est utilisé pour décrire des liens physiques entre pièces :

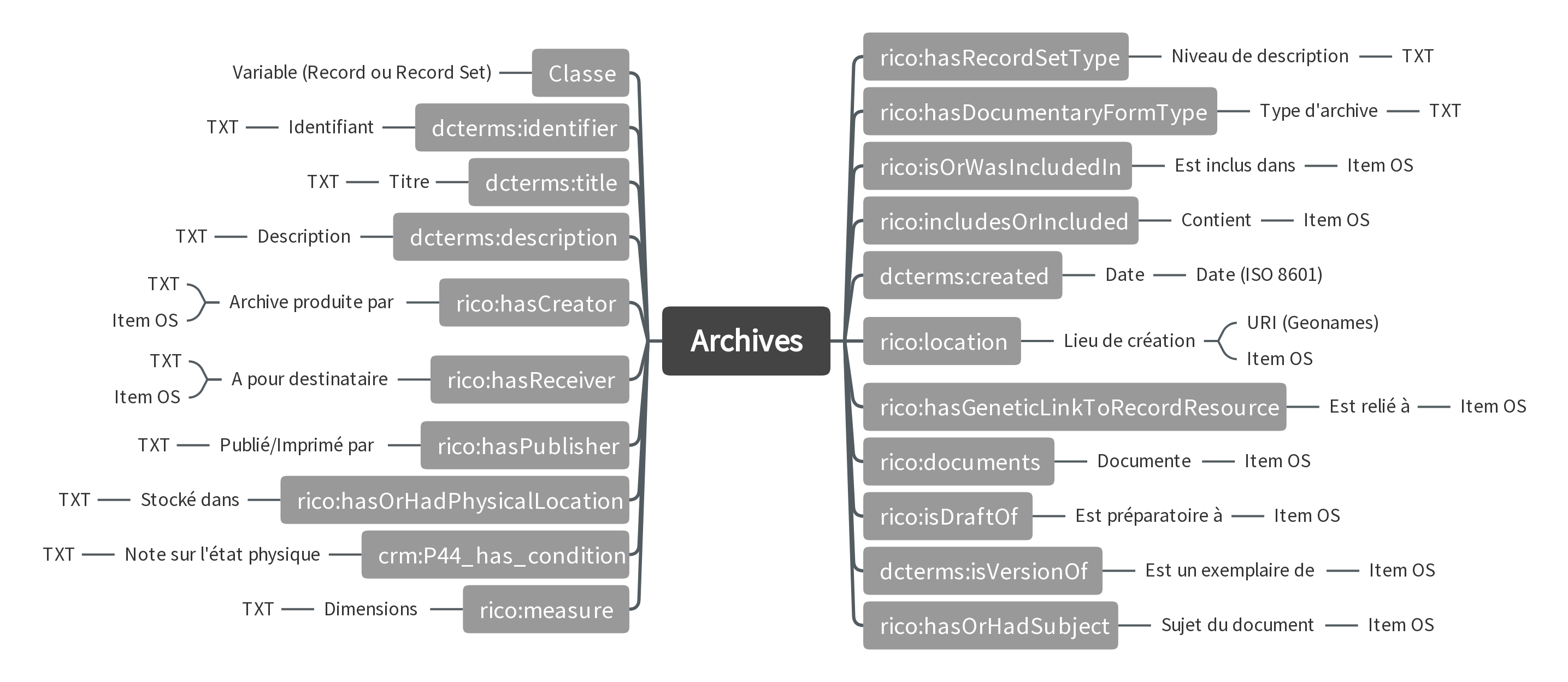
<instrument scientifique1> *dcterms:hasPart* <instrument scientifique2>

## ***5.2) Liens entre* items**

Comme nous l’avons vu plus haut, des liens entre *items* peuvent être dressés dans Omeka-S. Le schéma suivant couvre tous les cas où un *item* renvoie à un autre, avec indication de la propriété. Les données d’indexation sont en fond jaune pâle ; les données d’inventaire sont en fond bleu.



## ***5.3) Modélisation de données d'inventaire: Archives***

 La *Classe* est soit *rico:Record Set* (pour tout *item* hormis les pièces), soit *rico:Record* pour les pièces d’archives.

La valeur de *dcterms:description* est utilisée comme titre. Cette valeur est donc **obligatoire**, sous peine de voir s’afficher une série d’*items* nommés « [Untitled] » dans la base de données. La valeur de *dcterms:title* est réservée lorsqu’un titre est mentionné sur la pièce (titre de livre, en-tête de document, …).

Les valeurs de *rico:isOrWasIncludedIn* et *rico:includesOrIncluded* sont des liens vers les *items* d’archives hiérarchiquement supérieurs et inférieurs, respectivement. Si ceux-ci existent, cette valeur devient **obligatoire**.

## *5.4) Modélisation de données d'inventaire: Patrimoine artistique*



La valeur de *dcterms:type* est une ou plusieurs URI issues de la catégorie *Techniques et domaines* de l’inventaire général du patrimoine culturel[[3]](#footnote-4). Les termes sont ceux donnés dans l’inventaire

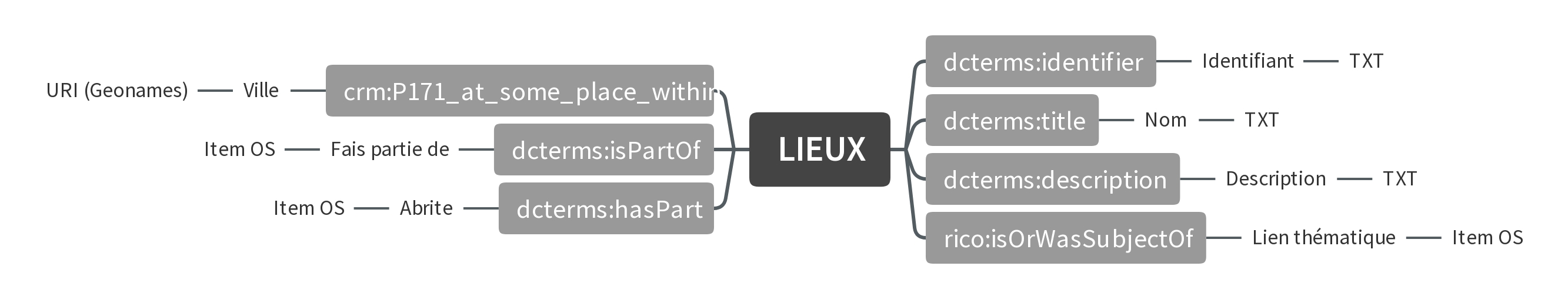
Les valeurs de *rico:isOrWasIncludedIn* et *rico:includesOrIncluded* sont des liens vers les ensembles qui ont été déterminés par Elizabeth Dandel dans ses rapports en PDF, et qui ont servi de structure pour ordonner la collection (par exemples, *Peintures d’Augustin Seguin, Matériel liturgique,* etc).

Notons qu’il y a deux valeurs de dimensions (*rico:measure* et *crmP43\_has\_dimension*), comme dans le fichier d’inventaire Excel. Nous avons utilisé *rico:measure* comme propriété pour les dimensions principales : les dimensions d’autres ressources (archives, instruments) sont renseignées avec cette propriété.

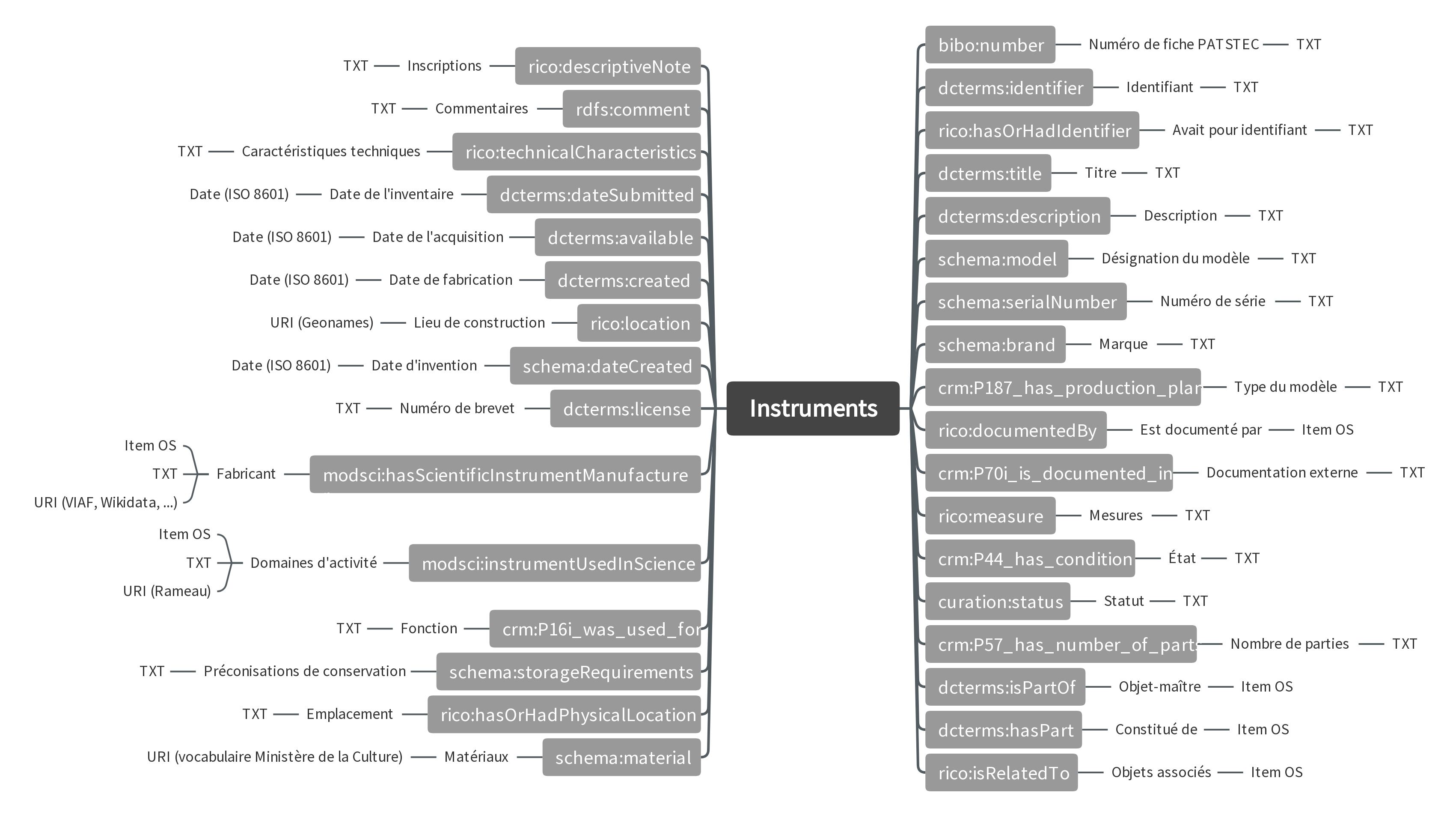
La propriété *crm:P94i\_was\_created\_by* est différente de la propriété de créateur retenue pour les archives (*rico:hasCreator*), pour distinguer l’acte de création artistique et l’acte de production d’un document. Soulignons aussi qu’il peut arriver que la paternité d’une œuvre soit incertaine. Dans ce cas, une annotation sur la propriété devra préciser cela (les annotations ne peuvent être ajoutées que manuellement).



## *5.5) Modélisation de données d’indexation: Lieux*

 Rien à signaler sur ce template.

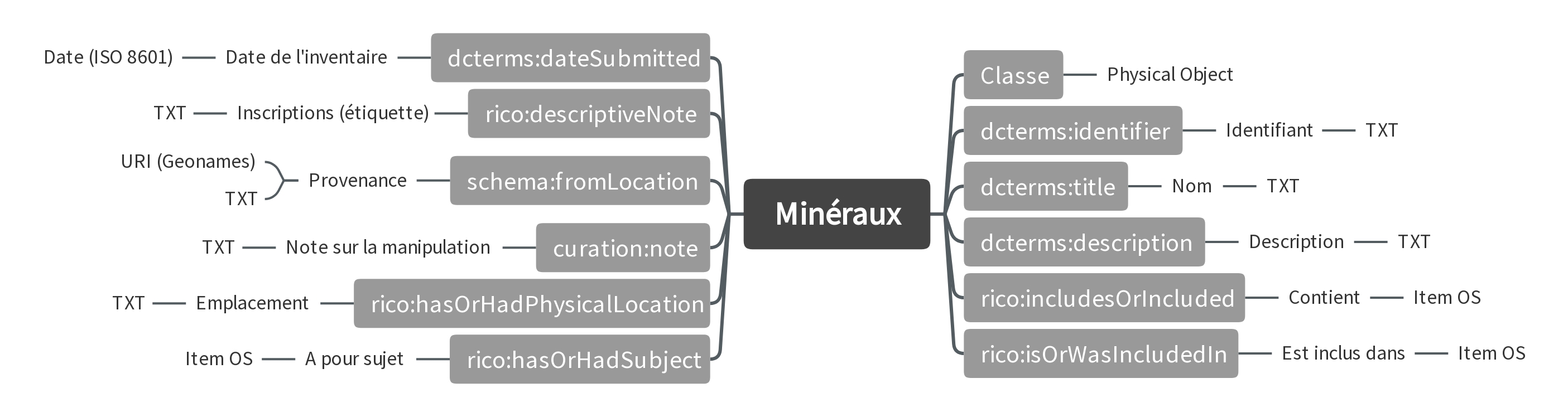
## *5.6) Modélisation de données d'inventaire: Instruments*

 Les propriétés rico:documentedBy et crmP70i\_is\_documented\_in contiennent deux types d’informations différentes : le premier est un lien vers une pièce d’archives dans lequel l’instrument est mentionné, tandis que le second est le contenu concaténé des éléments <lien\_information\_internet> et <documentation> du fichier XML de PATSTEC.

Les propriétés *crm:P44\_has\_condition* et *curation:status* sont quelque peu différentes : la première donne un constat d’état, la seconde précise si l’intégrité de l’objet est garantie et/ou s’il est fonctionnel.

Le fichier PATSTEC contenait trois numéros par instrument : le numéro de fiche de leur propre inventaire, le numéro d’inventaire normalisé en SEGUIN-0XXX, et le troisième est le numéro précédent. Le *dcterms:identifier* (rappelons-le, obligatoire) retenu comme numéro d’inventaire principal est la forme SEGUIN-0XXX. Les autres ont reçu des propriétés similaires, mais utilisées uniquement dans ce *template*.

## *5.7) Modélisation de données d'inventaire: Minéralogie*

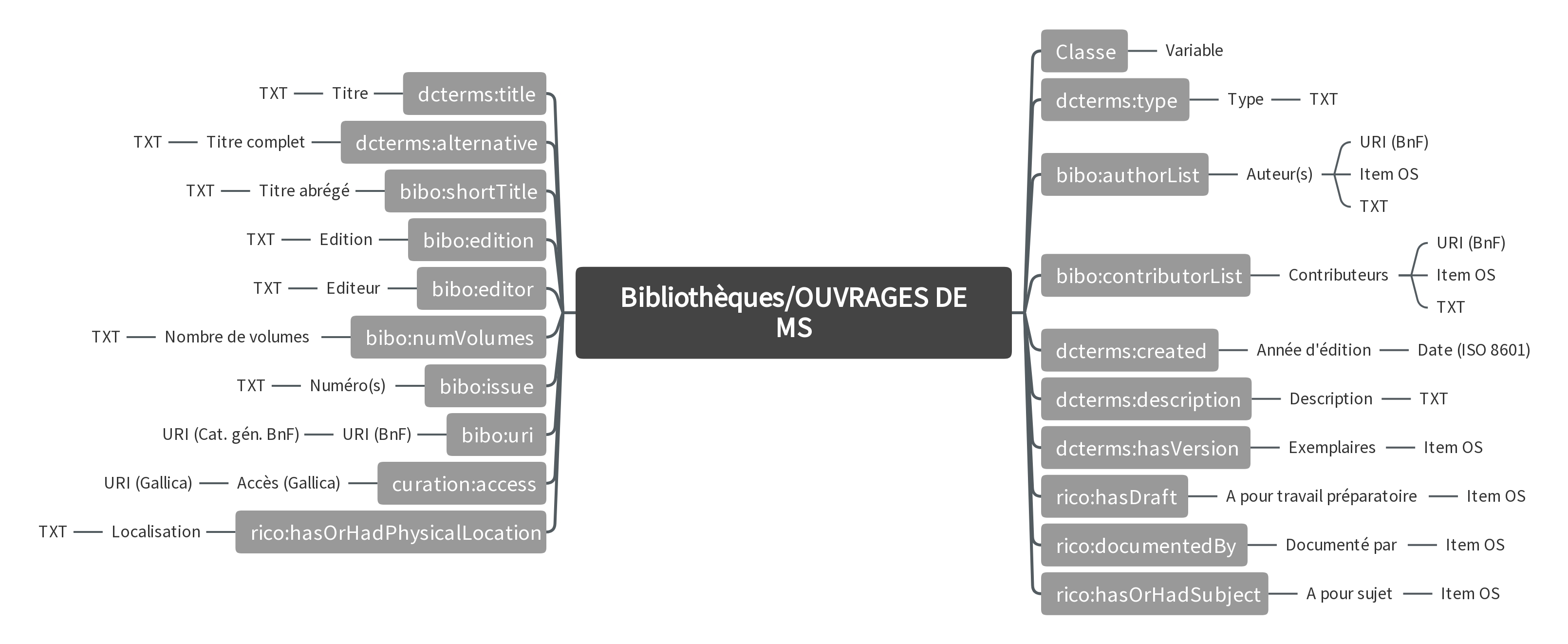


Les propriétés *rico:includesOrIncluded* et *rico:isOrWasIncludedIn* ne sont présentes dans le *template* uniquement pour le référencement des cinq collections identifiées par Taya Flaherty.

On notera que *schema:fromLocation* est une propriété différente de celle du lieu de production des archives, pour distinguer le lieu de collecte d’un lieu de production.

La propriété *rico:descriptiveNote*, qui a été utilisée pour les inscriptions sur les instruments notamment, est ici réutilisée pour préciser la mention contenue sur l’étiquette. Nous avons ici considéré qu’une étiquette était une forme d’inscription sur la pièce.

## *5.8) Modélisation de données d’inventaire/d’indexation: Livres*

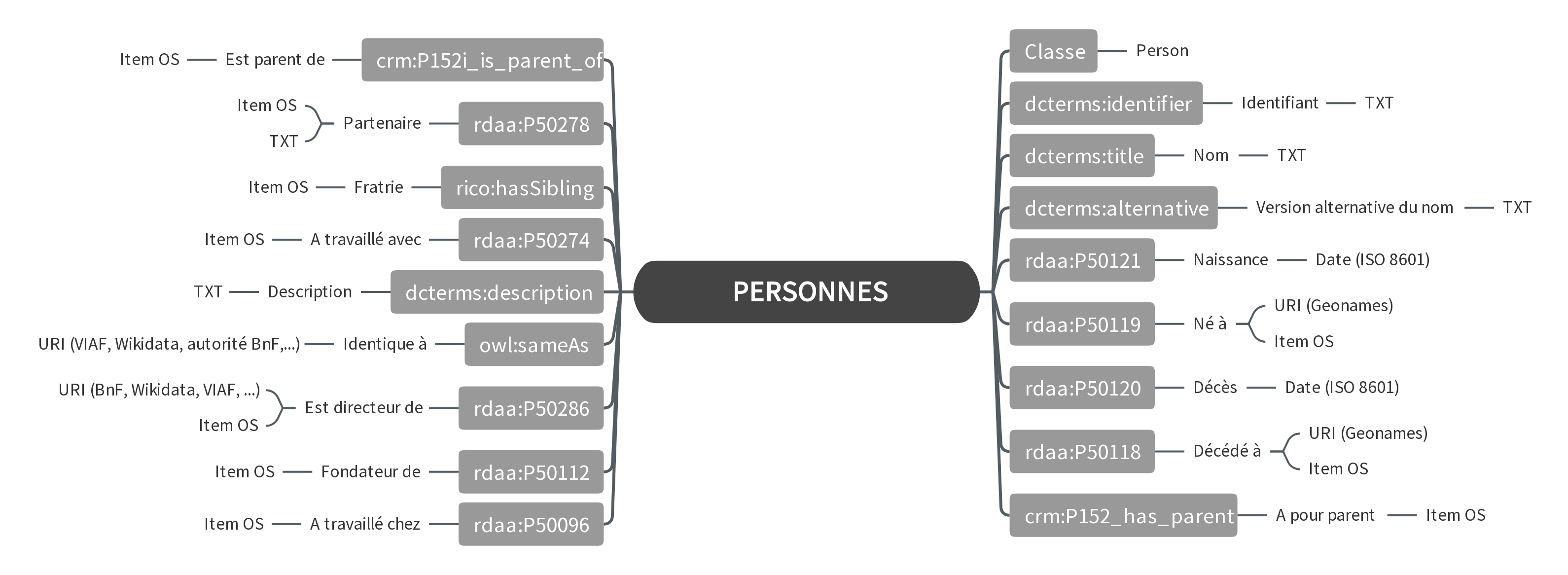
 Ce *template* a servi a la fois pour décrire les ouvrages contenus dans la bibliothèque (données d’inventaire, dans l’*item set* « Bibliothèque de Varagnes ») et pour les œuvres écrites de Marc Seguin (données d’indexation, dans l’*item set* «Ouvrages, contributions et travaux de Marc Seguin »). Le choix de l’utilisation peut avoir une incidence sur les valeurs que certaines propriétés peuvent recevoir.

Ainsi, un ouvrage donné ne peut avoir qu’une seule année d’édition, un seul URI de la BnF, et (normalement[[4]](#footnote-5)) un seul URI de Gallica. En revanche, dans le cas où ce *template* est utilisé pour décrire une œuvre, une d’année d’édition, d’URI BnF et de Gallica peut être renseignée autant de fois qu’il y a d’édition de l’œuvre. Pour les distinguer les unes des autres, nous pouvons annoter chaque sujet pour préciser l’édition. Malheureusement, les annotations ne peuvent pas être réalisées en masse par CSV Import (voir plus loin) et doivent être réalisées à la main.



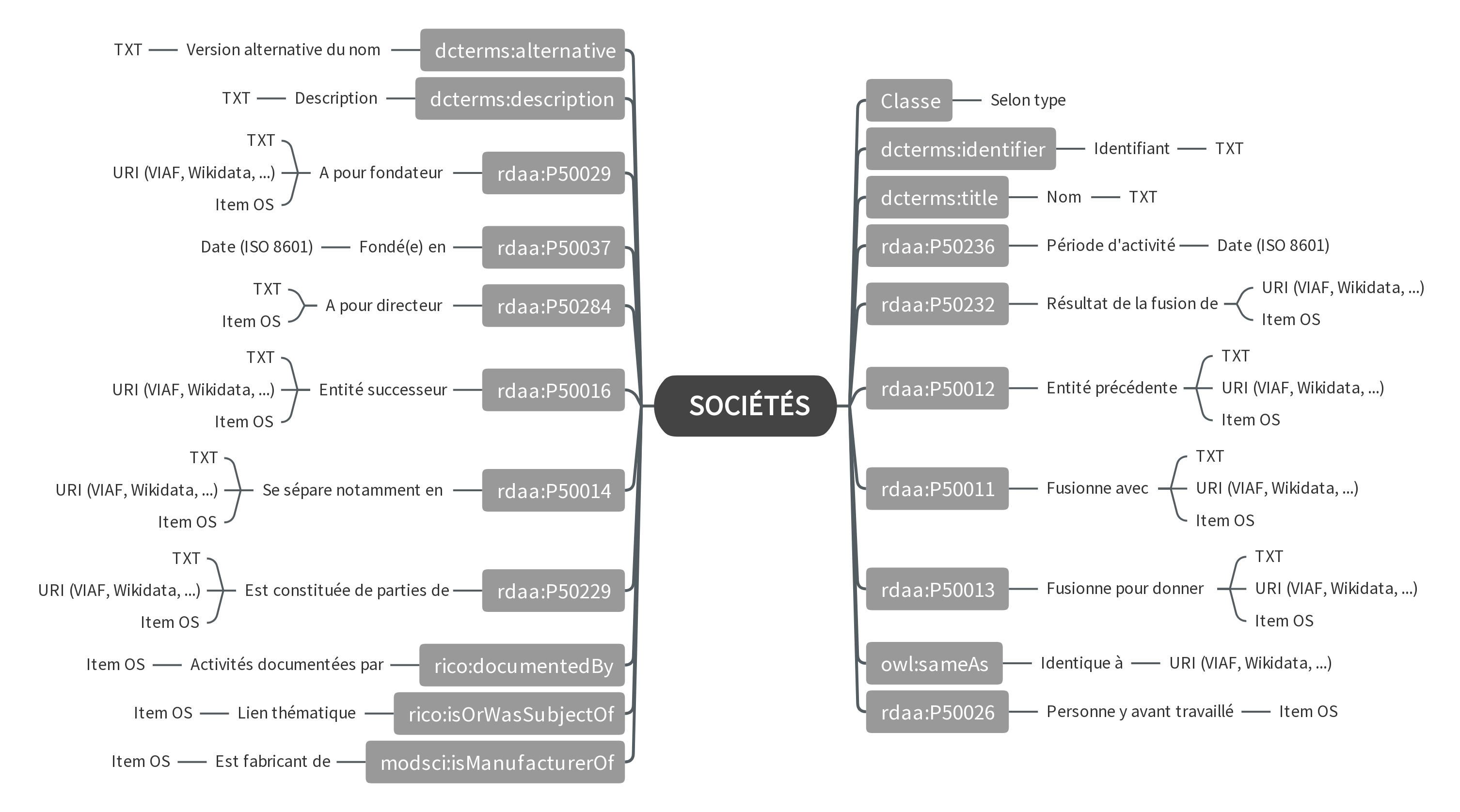
Enfin, notons que même si *bibo:authorList* et *bibo:contributorList* sont des listes, elles servent de propriété d’auteur pour le *template*, et doivent être remplis mêmes si elles ne comportent qu’une seule valeur.

## *5.9) Modélisation de données d’indexation: Personnes*

 Les propriétés *rdaa:P50119* et *rdaa:P50118* doivent comporter un URI issu de Geonames. La seule exception est lorsque la personne est née ou est décédée à Varagnes (qui est un *item* issu des Lieux).

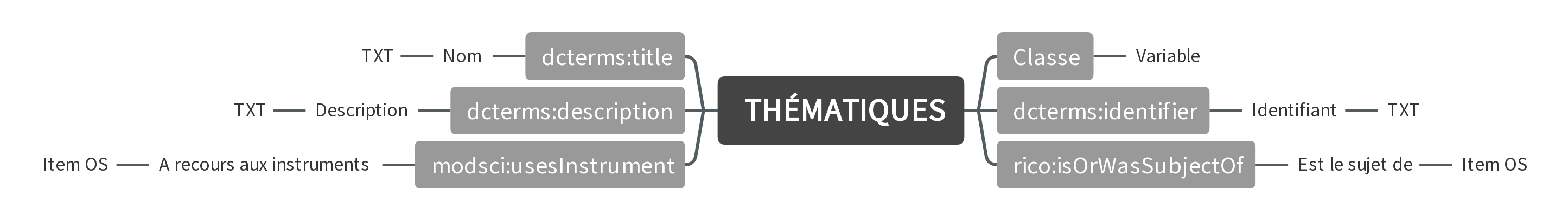
De manière générale, l’onglet « Metadata » de l’*item* (qui n’affiche que les triplets dont l’item est le sujet) suivant ce *template* devrait garder uniquement des données biographiques. Ainsi, on évite d’avoir une page surchargée d’informations[[5]](#footnote-6), et nous pouvons aller dans l’onglet « Linked ressources » pour voir les *items* qui ont cette personne en tant qu’objet (voir plus loin, section « Visualiser les données »).

## *5.10) Modélisation de données d’indexation: Sociétés*

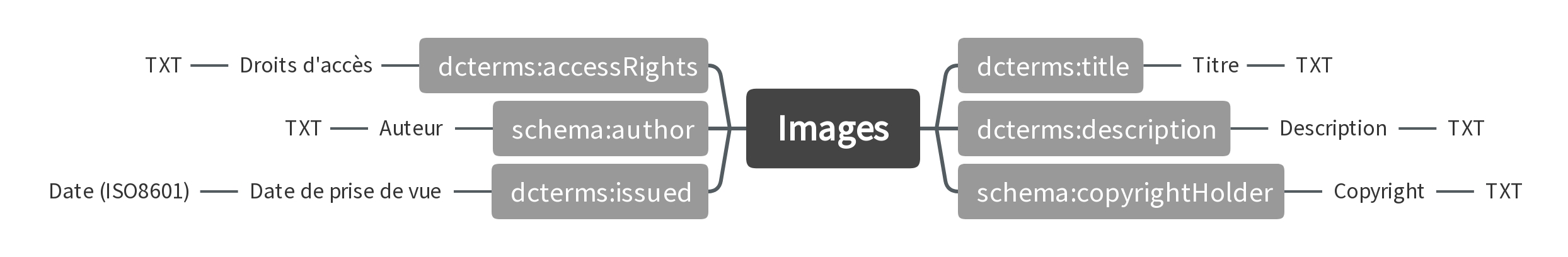


Les sociétés possèdent une propriété *rido:documentedBy* qui indique les archives qui documentent leurs activités. Dans le cas des sociétés relatives à la Buire et à Gnome et Rhône (les deux ayant subi des changements de statuts et de forme juridique), une thématique *Buire* et *Gnome et Rhône* reprend tous les documents qui sont normalement répartis entre les différentes versions de l’entreprise.

## ***5.11) Modélisation de données d’indexation: Thématiques***

 Rien à signaler sur ce *template*.

## *5.12) Modélisation de données de médias: Images*

 Ce *template* s’applique aux images (qui sont un type de *media*) lors de l’import d’images reliées à des *items*. A l’heure d’écrire ces lignes, les photos disponibles ne sont que des photos de pièces d’inventaire. Si la base de donnée s’enrichit un jour de photographies historiques, ce *template* sera à revoir en profondeur pour contenir des informations de nature historique (et non plus uniquement administrative).

# 6) Visualiser les données

## *6.1) Généralités: les fiches d’*item

Par défaut, cliquer sur une fiche *item* renvoie sur un onglet appelé « Metadata ». **Cet onglet affiche tous les triplets pour lequel l’*item* est le sujet**.



Dans l’onglet suivant, « Linked resources », nous pouvons voir **tous les triplets pour lesquels l’*item* actif est l’objet**. On notera également que, à la différence des données montrées dans « Metadata », nous pouvons ordonner les liens selon les propriétés utilisées.



## *6.2) Parcourir les collections*

### 6.2.1) Via les *item sets* de données d’indexation

C’est probablement la manière la plus intuitive de parcourir les collections. Cependant, l’exhaustivité de l’indexation est variable, selon les *item sets*.

- *item set* de personnes : l’indexation porte sur les champs « auteurs » (qu’il s’agisse d’une archive, d’un livre ou d’un instrument scientifique) et sur les champs « destinataire » d’une pièce d’archives. Cela a pu être fait durant la durée du stage, et le taux d’exhaustivité doit approcher les 100 %.

- *item set* des ouvrages et contributions de Marc Seguin : l’indexation porte sur les pièces d’archives qui sont des notes préparatoires aux ouvrages, qui sont des exemplaires de cet ouvrage, ou qui documentent de manière générale cet ouvrage. L’exhaustivité est estimée autour des 80 %.

- *item set* des sociétés : l’indexation a pour but d’identifier les documents d’archives témoignant des activités de la société. Le résultat est inégal selon les sous-fonds : si nous avons pu analyser les archives de Marc Seguin et de sa génération en profondeur, les archives d’Augustin Seguin et de sa génération n’ont été indexées que superficiellement, sur bases des notes prises par Lisa Lafontaine. Un dépouillement et une indexation approfondis de ces archives B est à prévoir.

- *item set* des lieux : la plupart des lieux servent aussi de thématiques, et nous formulons la même remarque que pour les sociétés. A noter que l’*item* « Varagnes » est également un lieu servant à l’indexation des pièces d’archives produites à cet endroit.

- *item set* des thématiques : Même remarque que pour les archives et les lieux.

### 6.2.2) Via les *item sets* d’inventaire (hors Archives)

Les *item sets* d’inventaire ne permettent pas tous une navigation satisfaisante, car les inventaires initiaux ont été créés de manière inégale et les données de la base n’en sont qu’un reflet. Certains proposaient un classement (c’est le cas de la minéralogie et de l’artistique), d’autres non (c’est le cas de la bibliothèque et des instruments scientifiques). Pour ces deux derniers, il est exclus de procéder à un classement par lieu de stockage qui, par définition, est susceptible de changer au fil des années.

- l’*item set* de minéralogie affiche directement les cinq collections qui ont été identifiées par Taya Flaherty. Chaque collection comporte une description, extraite des fichiers PDF qu’elle a remis.

- l’*item set* de l’artistique a lui aussi été structuré par Elizabeth Dandel, selon des catégories de type d’œuvres et, parfois, en sous-catégories selon leur créateur.

- l’*item set* des bibliothèques n’est pas structuré, et il tient davantage de la liste de références bibliographiques. De plus, les livres n’ont pas été indexés (faute de temps). En l’état, la seule manière de parcourir l’inventaire de manière pertinente est via le formulaire de recherche.

- l’*item set* des instruments est également dépourvu de classement, mais il est tout de même bien plus navigable que l’inventaire de la bibliothèque de par sa taille plus restreinte (271 pièces, contre dix fois plus pour la bibliothèque).

### 6.2.3) Le cas des archives

Les archives sont un cas spécifique, car le type d’inventaire est très différent de celui des autres pièces. Là où les autres inventaires possèdent peu ou pas de classement, celui des archives s’étale sur une grande profondeur (jusqu’à 7 niveaux d’arborescence) – ce qui en fait d’ailleurs sa richesse.

Or, cette profondeur s’accommode mal de la « mise à plat » que constitue la mise en base de données : qu’il s’agisse d’une série, d’un dossier ou d’une pièce, tous sont représentés par une fiche dont la position par rapport à d’autres niveaux de description est, au mieux, floue.

Techniquement, il est cependant possible de naviguer dans les archives depuis la page de l’*item set*. Celle-ci contient un lien vers chacun des sous-fonds, dont les pages contiennent elles-mêmes des liens vers les séries, etc. Il est cependant à prévoir que l’utilisateur se retrouvera vite perdu dans cette suite de fiches, au design tous similaires.

Nous avons donc doublé l’*item set* avec l’édition d’un inventaire en ligne, sous forme de fichier HTML. Le fichier Excel initial a été transformé en fichier XML-EAD[[6]](#footnote-7) (qui servira d’ailleurs pour les collaborations ultérieures entre Varagnes et les AD de Privas), et celui-ci a à son tour été transformé en HTML avec XSLT. Il a cet aspect :

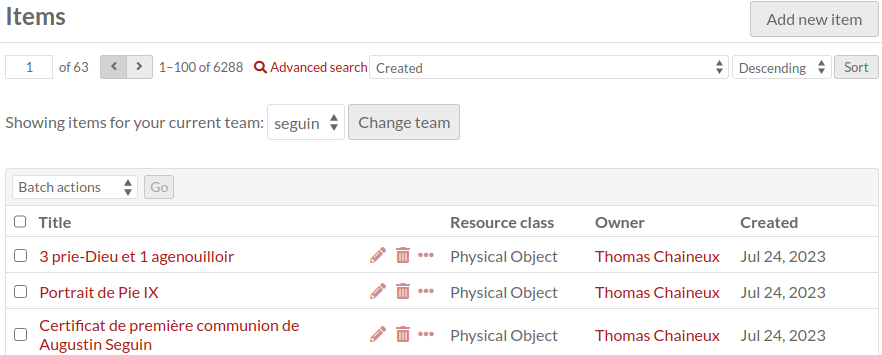


Chaque flèche pointant vers la droite indique que le niveau de description contient des éléments hiérarchiquement inférieurs. Tout fonctionne comme un menu, à dérouler successivement, pour arriver au niveau d’information voulu. La lien en fin de ligne (« Vers la Base ») renvoie vers la page de l’*item* d’Omeka-S correspondant, sur le site de l’École Centrale de Nantes.

Lorsqu’aucune flèche n’apparaît pas à gauche de l’élément, cela signifie qu’il n’y a pas de niveau hiérarchiquement inférieur. Cela se produit lorsque l’on atteint le niveau de la pièce, ou bien que le dossier d’archives n’est pas décrit plus en profondeur.

### 6.2.4) Formulaire de recherche

Nous avons aussi accès à un formulaire de recherche : depuis la liste des *items*, cliquer sur *Adanced Search* en haut de la page :



Cela nous donne accès à un formulaire de recherche standard, qui permet de croiser les critères de recherche, selon :

- une recherche plein texte dans les pages

- selon la propriété et l’objet, pour afficher le sujet. Il est possible de remplir plusieurs critères pour maximiser l’efficacité. **Il est nécessaire de soit bien connaître les propriétés utilisées pour chaque description, soit de se référer aux schémas des pages précédentes pour les identifier selon ce que vous cherchez**.

- selon la classe

- selon le *resource template* utilisé

- selon l’*item set* auquel appartient d’*item*

- selon que l’*item* possède un *media* ou non

- selon l’ID de l’*item*. Attention, nous parlons bien ici **de l’ID de l’*item* dans Omeka-S**, pas dans l’objet de la propriété *dcterms:identifier*. L’ID d’un *item* d’Omeka-S se trouve notamment dans l’URL de sa page :



- selon les différents champs de date. La sélection des propriétés se fait parmi les champs qui ont été renseignés comme des champs dates dans le *resource template* utilisé pour la description de l’*item*.

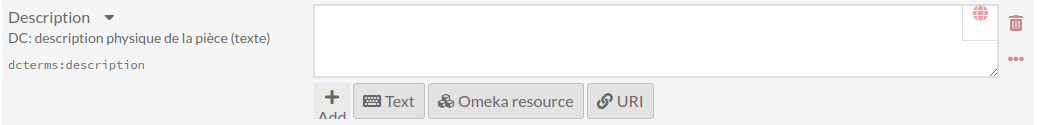
# 7) Mises à jour de la base

## *7.1) Mise à jour manuelle*

### 7.1.1) Modification, une *ressource* à la fois

La modification ne présente aucune difficulté majeure car, dans la mesure où la création d’un *item* impose l’utilisation d’un *resource template*, tous les champs disponibles pour ce dernier apparaîtront sur la fiche. Il n’y a plus qu’à remplir.

**Par défaut**, un champ va apparaître sous cette forme :

 Nous voyons que nous pouvons renseigner trois types de valeurs : du texte, une « Omeka resource », et un URI. Notez que vous pouvez inclure plusieurs valeurs différentes, en cliquant successivement sur l’un ou l’autre. Cliquer sur la case correspondante selon le besoin.

1) Le texte : rien à signaler : écrivez ! **Attention** tout de même : il est facile de remplir un champ avec tout un tas de notes diverses, pour mentionner des choses variées. **Cette pratique est à réduire au maximum** : comme dit en introduction, le RDF est un format hautement structuré, et l’information doit être aussi concise et représentative de la propriété que possible. Il y a cependant des propriétés naturellement dédiées à de l’écriture plus libre : *dcterms:description*, *rdfs:comment*, *rico:descriptiveNote…* Mais elles tiennent davantage de l’exception que de la règle. De même, **ne pas oublier que cette base de données est destinée à être ouverte au public** : des notes personnelles, des mentions trop administratives, etc. sont à proscrire… à moins d’éditer le *template* et d’y ajouter un champ consacré, qui sera rendu non-visible selon le profil utilisateur.

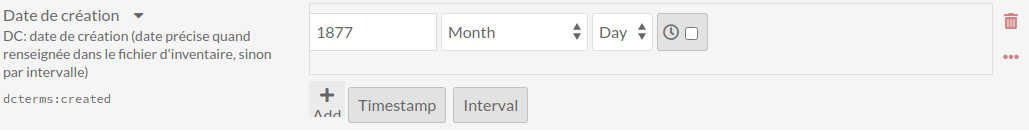
2) « Omeka resource » : il s’agit de créer un lien avec une *resource* déjà créé dans Omeka-S. Cliquer dessus vous permettra de choisir entre une *resource* de type *item, item set* ou *media.* Sélectionner : apparaît alors dans la barre latérale droite, un champ de recherche pour identifier la *resource* à lier.

Si on a sélectionné *item set*, la liste de ceux-ci apparaît.

Pour les *items*, la solution la plus simple et la moins génératrice d’erreur est de chercher par son identifiant (la valeur de la propriété *dcterms:identifier*). Par exemple, si vous cherchez « Marc Seguin », vous aurez à choisir parmi 1146 *items*, dont trois se nomment *Marc Seguin…* Si vous cherchez REFPERS0002 (l’identifiant de cette personne, en assumant que nous parlons bien du Marc Seguin « principal »), vous n’en aurez qu’un.

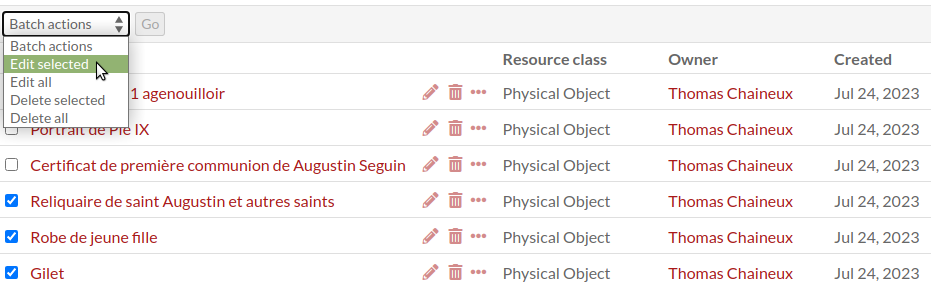
3) URI : cliquer sur l’URI vous fait apparaître deux champs : renseigner l’URI dans le premier, et le label optionnel dans le deuxième.

**Si le *resource template* a défini des champs sous format de date**, l’édition vous donnera le choix entre « Timestamp » ou « Interval ». Cliquer selon le besoin, et remplir. « Timestamp » est une date fixe, « Interval » est un… intervalle, avec une date minimale et une date maximale.



### 7.1.2) Modification par « batch »

A chaque fois que vous avez devant vous une liste d’*items* (que ce soit depuis la liste globale, en résultat d’une recherche, ou après une sélection par *item set*), vous pouvez les sélectionner en cochant la petite case à gauche de son titre. En en-tête de tableau, un menu déroulant (« Batch actions ») vous propose quelques manipulations à faire avec votre sélection.



Cela permet de régler des paramètres pour l’ensemble de la sélection, tels que la visibilité, l’application d’un *template*, l’appartenance à un *item set* ou à un site, ou encore d’effacer le contenu d’une propriété. Cela permet aussi d’ajouter une propriété et un objet, en cliquant sur « Add text value », « Add resource value » ou « Add URI value » en bas de page. Cela s’appliquera à la sélection. Attention cependant : le choix de la propriété se fait parmi l’ensemble des propriétés, pas parmi celles contenues dans le *template*! **Veillez à choisir la bonne**. Les modifications par « batch » ne sont **pas annulables** par la suite.

### 7.1.3) Création de nouveau *items*

Les remarques pour la création sont en tout point pareille que pour la modification. La première chose à faire est de choisir le *resource template* selon la nature de la chose à décrire ; les propriétés présélectionnées apparaissent alors. Remplir, en tenant compte des remarques précédentes, et du fait que certaines propriétés doivent **obligatoirement** être complétées.

Ne pas oublier de renseigner une classe.

## *7.2) Mise à jour de masse – CSV Import*

Le module utilisé pour l’import de données en masse est CSV Import. Une bonne documentation est accessible sur internet[[7]](#footnote-8) et est régulièrement mise à jour. Nous n’allons pas reprendre ici la documentation générale, mais plutôt apporter quelques précisions utiles qui résultent de l’usage du module.

**1) Une propriété avec des types de valeurs différentes.**

Il peut arriver qu’une même propriété ait pour objet des valeurs de types différents, comme par exemple du texte, un URI ou un lien vers un item d’Omeka-S. Or, il n’est pas possible de spécifier plusieurs types de données différents pour une même colonne lors de l’import par CSV.

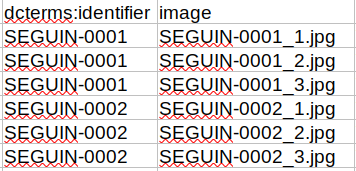
Cela implique que lors de la création du CSV de chargement, une colonne distincte doit être créée pour chaque type de données. Lors des Import Settings, chaque colonne doit recevoir son type de données différent. Si, comme dans l’exemple ci-dessous, le type de donnée a été inclus dans le nom de la colonne pour éviter toute confusion, il ne faut pas oublier de régler le mapping.

 Notez que cette capture d’écran montre également des types de valeurs différentes pour les dates. Lors de la création des données en masse, il faut bien veiller à avoir une colonne qui contient les dates fixes (1857-12-03, par exemple), et une colonne qui contient les intervalles (1841-07-14/1841-08-11). Les deux suivent la norme ISO 8601 mais le processus d’insertion dans la base diffère, d’où le besoin de créer deux colonnes.

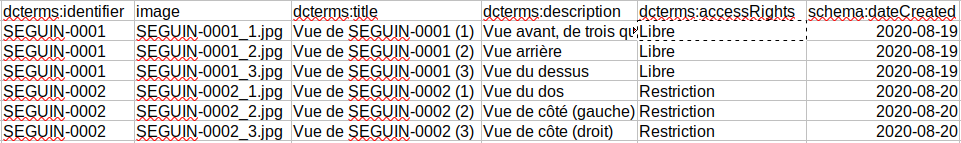
**2) Création d’images**

Le module CSV Import peut se cumuler avec le module File SideLoad[[8]](#footnote-9), comme spécifié dans la documentation de ce dernier. Un point doit cependant être spécifié : si les médias ne peuvent être créés qu’en étant liés à un item ou à un item set, ils doivent être considérés comme des créations de nouveaux éléments et non comme des mises à jour d’items/item sets.

Cela a une influence sur la manière de construire le CSV. En effet, chaque nouvelle image doit avoir sa propre ligne, et si un seul item doit recevoir plusieurs images, alors l’identifiant de cet item doit se répéter autant de fois qu’il y a d’images.



Dans l’exemple ci-dessus, la colonne *dcterms:identifier* ne sert finalement que d’ancre pour relier l’image à une pièce. Si d’autres propriétés sont ajoutées en colonnes, elles s’appliqueront bien à l’image et non à la pièce. Un *resource template* « Images » existe et doit être utilisé.



**3) Mises à jour d’éléments**

Cela est spécifié relativement discrètement dans la documentation, aussi est-il préférable de le mentionner ici : dans les *Advanced settings*, et à l’exception de l’action *Create a new ressource*, **toutes les autres actions ne sont pas réversibles**. Cela signifie qu’une action autre que *Create a new ressource* doit être particulièrement bien préparée et vérifiée avant d’être exécutée, sous peine de devoir refaire d’autres imports ultérieurs pour corriger des erreurs non-réversibles.

Ces actions nécessitent d’identifier la ressource à modifier ou supprimer. Bien qu’il soit techniquement possible d’identifier la ressource selon l’objet de n’importe quelle propriété, il convient de choisir des données selon des données désambiguïsées. La propriété *Dublin Core/ Identifier* est suffisante, mais l’*Internal ID* est encore mieux.

# 8) Exporter les données

Nous avons demandé l’installation du module *Bulk Export*, qui permet d’exporter des données en plusieurs formats : CSV, JSON-LD, JSON-Table, ODS, TSV et TXT.

Du fait que nous avons utilisé le module *CSV Import* pour charger les données, nous n’avons exporté qu’en CSV. Les notes suivantes ne vont donc concerner que ce format.

L’export en CSV peut se faire à bien des niveaux d’affichage : depuis une liste d’*items*, depuis une page de *resource…* L’export ne va a chaque fois concerner que la page active.

La quantité de lignes exportables est de 1.000 par CSV. Les paramètres de *Bulk Export* peuvent être modifiés ; cela est à voir avec Matthieu Quantin de l’École Centrale de Nantes.

L’export est le seul moyen (actuel) de récupérer les ID des *resources* en masse, ce qui est nécessaire pour construire d’autres CSV impliquant des liens entre elles.

# 9) Perspectives d’amélioration

La base actuelle est tout à fait fonctionnelle et met un système en place, mais elle peut-être améliorée.

## *9.1) Normaliser les données au maximum*

La courte durée du stage fait que des données ont parfois dû être insérées dans la base sous une forme qui, sans être mauvaise, est améliorable en terme de référencement par URI. C’est le cas des ouvrages contenu dans la bibliothèque, en particulier.

En effet, les données initiales sous Excel n’étaient pas suffisamment concises (nom d’auteur, titre, …) pour interroger les API de la BnF de manière optimale. Nous avons néanmoins pu effectuer une partie du travail, et ainsi récupérer des URIs d’auteurs lorsque la requête ne renvoyait qu’un unique résultat ; au-delà, la vérification humaine aurait pris trop de temps. Le fait qu’un auteur soit référencé en données textuelle dans la base ne signifie donc pas qu’il ne possède pas d’identifiant dans le catalogue général de la BnF, mais bien qu’il n’a pas pu être trouvé de manière certaine dans la requête Python. Un travail humain est à prévoir.

Une fois que les auteurs auront été identifiés par leurs URIs, il sera bien plus simple de trouver les URIs de leurs ouvrages (dans le catalogue général) et de leurs numérisations (sur Gallica) pour les insérer dans la base et l’enrichir considérablement.

Bien que cela soit infiniment plus restreint, quelques URI de personnes et de lieux pourraient probablement être trouvés (de manière humaine), afin de référencer des auteurs et destinataires d’archives, des fabricants d’instruments, ou des lieux de collectes de minéraux.

## 9.2) Poursuivre le travail d’indexation

Une des forces d’Omeka-S et du RDF est de pouvoir construire des bases qui mettent l’accent sur l’interaction entre ressources à travers des propriétés. Notre travail a essayé de poursuivre cet objectif de la meilleure manière possible, mais le temps imparti nous a contraint à devoir appliquer ce traitement de manière beaucoup plus lâche sur une partie des archives (celles d’Augustin et de sa génération, en particulier).

De plus, bon nombre de documents concernent des matières techniques très pointues, dont nous ne sommes pas en mesure de juger pleinement la portée. Quelqu’un avec des connaissances scientifiques et historiques plus pointues pourrait certainement enrichir les indexations déjà existantes.

Le système est néanmoins en place : le principe de créer des données d’indexation selon les sujets voulus, de lier des *resources* entre elles, … peut être étendu selon les besoins futurs.

1. Nous reviendrons plus loin sur ce fichier EAD. [↑](#footnote-ref-2)
2. <https://fr.wikipedia.org/wiki/ISO_8601> (consulté le 27 juillet 2023). [↑](#footnote-ref-3)
3. <http://data.culture.fr/thesaurus/page/ark:/67717/T262> (consulté le 27 juillet 2023) [↑](#footnote-ref-4)
4. Il arrive qu’un ouvrage soit numérisé une seconde fois, pour suivre l’évolution de la qualité des numérisations. Mais cela reste rare. [↑](#footnote-ref-5)
5. Pour donner un ordre de grandeur, Marc Seguin a écrit et 201 archives et est le destinataire de 330 autres. Afficher tout ceci sur sa page l’alourdirait beaucoup trop. [↑](#footnote-ref-6)
6. *Encapsuled Archival Description*. Format XML spécifique à la description archivistique. Notons que nous avons utilisé l’EAD 2002, malgré qu’il existe un EAD 3. Ce dernier peine à se faire une place, et reste encore largement sous-utilisé. [↑](#footnote-ref-7)
7. <https://omeka.org/s/docs/user-manual/modules/csvimport/> (consulté le 26 juillet 2023) [↑](#footnote-ref-8)
8. <https://omeka.org/s/docs/user-manual/modules/filesideload/> (consulté le 26 juillet 2023) [↑](#footnote-ref-9)