Tutoriel alignement de vocabulaire : Getty AAT

Auteur(s) : , Lizzie Scholtus,
Institution émettrice : Consortium MASA
contact mail du projet : olivier.marlet@univ-tours.fr

Avant-propos

Ce tutoriel a pour objectif d'expliquer comment effectuer un alignement de vocabulaire vers le référentiel Art & Architecture Thesaurus créé par le Getty Research Institute, à l'aide du logiciel libre OpenRefine[1]. Le but de cet alignement est de créer un document d'équivalences permettant ensuite l'appareillement des données concernées avec la plateforme ARIADNEplus. Il ne s'applique que pour des jeux de données déjà alignés avec le référentiel PACTOLS de Frantiq[2]. Ce document est établi à l'aide des données provenant de la base *Atlas des établissements ruraux de la Beauce antique* (AERBA) qui se présente au format xml.

Aspects techniques

Ce tutoriel est écrit en R Markdown à l'aide du package "Bookdown" créé par Yihui Xie (*bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.24., 2021).

I. Récupérer les informations nécessaires à l'alignement

Afin de réaliser l'alignement entre les données sources et le vocabulaire du Getty, il est préférable de réunir au préalable plusieurs informations qui faciliteront par la suite la réconciliation.

1. Les URI PACTOLS

Quel que soit le format du jeu de données (tableur, xml, etc.)[3], il est d'abord nécessaire de regrouper au sein d'une même colonne l'ensemble des URI PACTOLS contenues dans le document. Parce qu'il est fort probable qu'un même terme réapparaisse plusieurs fois au sein du jeu de données, il faut en supprimer les doublons. Pour ce faire, on peut trier la colonne les contenant et utiliser l'option **"Vider des valeurs répétées dans des cellules consécutives"** proposée dans la rubrique d'édition de cellule (figure 1.1).

Figure 1.1: Supprimer des valeurs en double

On supprime ensuite les lignes devenues vides à l'aide d'une facette par valeur vide (figure 1.2).

Figure 1.2: Supprimer des lignes vides

1. Les *preflabels* PACTOLS

L'alignement avec le référentiel du Getty AAT se fait d'après les termes eux-mêmes et non leur URI. Il est donc nécessaire de récupérer cette information, mais en anglais. Effectivement, si une partie du référentiel AAT est traduite en français, ce n'est pas le cas pour l'ensemble des termes qu'elle recense. Par ailleurs, le service de réconciliation, qui sera utilisé par la suite, s'appuie sur le *preflabel* de l'AAT qui est lui-même en anglais. Afin d'automatiser au maximum le processus d'alignement, nous avons choisi d'importer le *preflabel* anglais contenu dans les PACTOLS. Il faut toutefois signaler que cette information n'est pas systématiquement renseignée et qu'un certain nombre de termes seront donc à traduire manuellement, soit avant la réconciliation, soit après.

A. Moissonner une URL

La première étape consiste donc à récupérer l'ensemble des informations PACTOLS en reconstituant les URL de chaque terme tel qu'indiqué par le webservice d'Opentheso et d'en moissonner le contenu. OpenRefine permet de réaliser ces deux actions à l'aide de la fonction **"Ajouter une colonne en moissonnant des URL…"** qui offre la possibilité d'utiliser une expression GREL[4] pour reconstituer l'URL (figure 1.3).

Expression GREL[5] :

```
"https://pactols2.frantiq.fr/pactols2/api/" + value.replace(/.*\/(\d+.*)/, "$1")
+ ".ttl"
```

pactols

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix iso-thes: <http://purl.org/iso25964/skos-thes#> .
@prefix opentheso: <http://purl.org/umu/uneskos#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

<https://ark.frantiq.fr/ark:/26678/crt5AOuUmbGVy> a skos:Concept;
  skos:prefLabel "établissement rural"@fr;
  skos:altLabel "installation rurale"@fr;
  skos:related <https://ark.frantiq.fr/ark:/26678/pcrt9PJh9aTXv4>;
  skos:broader <https://ark.frantiq.fr/ark:/26678/pcrtn16fXQKqwM>;
  skos:inScheme <https://ark.frantiq.fr/ark:/TH_1>;
  opentheso:memberOf <https://hdl.handle.net/>;
  skos:exactMatch <https://www.wikidata.org/wiki/Q10354598>;
  dcterms:created "2018-12-13"^^<http://www.w3.org/2001/XMLSchema#date>;
  dcterms:modified "2020-07-30"^^<http://www.w3.org/2001/XMLSchema#date>;
  dcterms:identifier "246237";
  dcterms:description "246691##246346##246366##246457##246481##246237";
  skos:definition "Ensemble d'unités bâties se situant en milieu rural, à la campagne."@fr .
```

Figure 1.3: Moissoner des URL avec Openrefine

B. Récupérer le terme en anglais

Il est alors possible d'extraire le *preflabel* en anglais des données importées, en créant une nouvelle colonne **Preflabel(en)** , d'après les valeurs moissonnées et en y sélectionnant l'information désirée (figure 1.4). On remarquera, comme cela a déjà été signalé, que cette information n'existe pas forcément pour l'ensemble des termes, les cellules correspondantes restant alors vides.

Expression GREL :

```
value.find(/".*"@en/)[0].replace(/.*"(.*)"@en.*/, "$1")
```

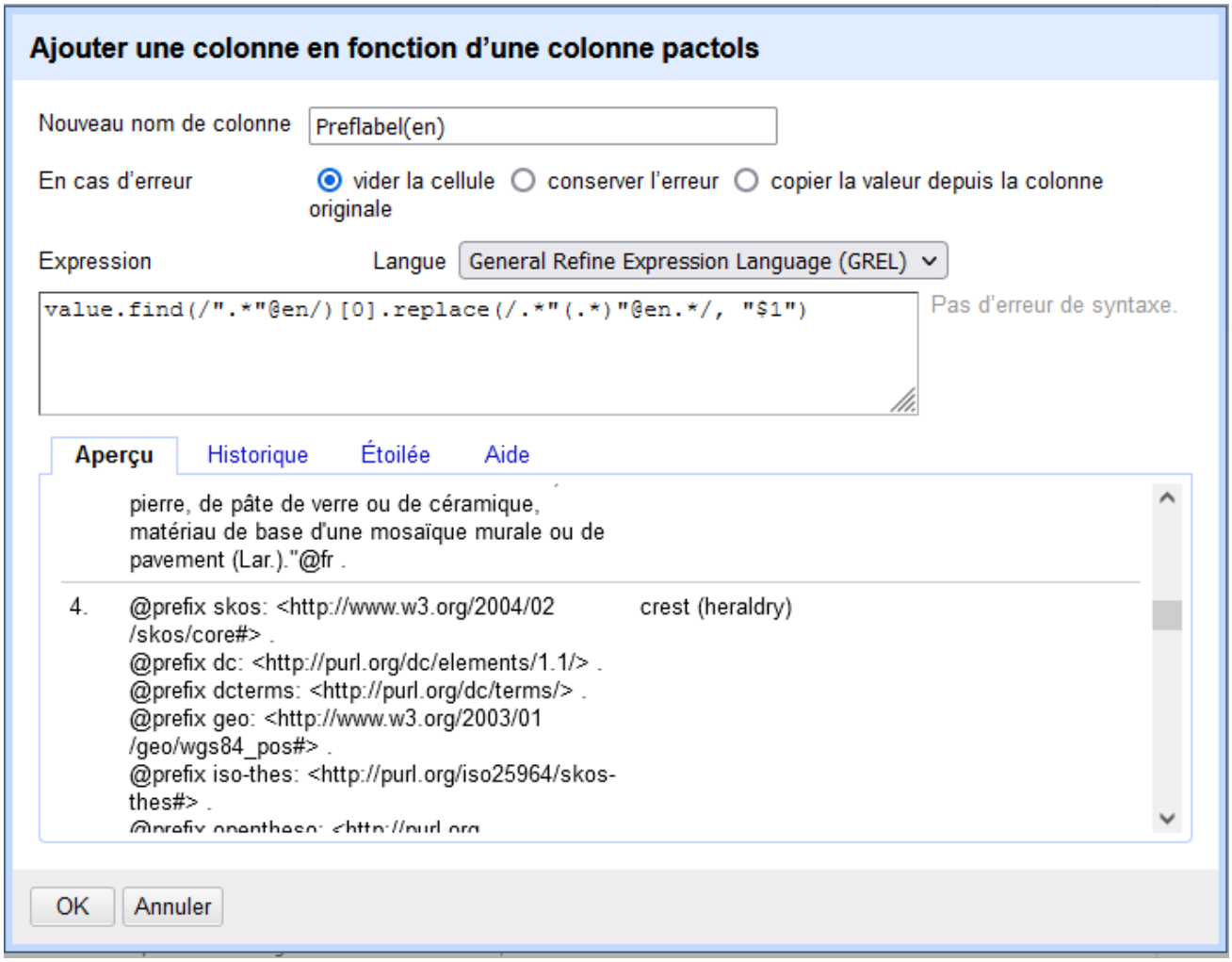


Figure 1.4: Extraire le preflabel en anglais

Dans les *labels* PACTOLS, il peut arriver qu'un terme soit suivi de précisions entre parenthèses qui risquent de nuire à la réconciliation avec le vocabulaire de l'AAT. Toutefois, ces précisions sont utiles pour la compréhension du concept et du champ auquel il se rattache. Aussi, plutôt que de les supprimer complètement, une nouvelle colonne **preflabel(en)2** est créée, en fonction des données de **preflabel(en)**, et ces précisions sont effacées à l'aide d'une nouvelle expression GREL.

Expression GREL :

```
value.replace(/(.*)\(.*\)/, "$1")
```

C. Récupérer le terme en français

En suivant le même processus d'ajout de colonne d'après les informations moissonnées précédemment au sein de la colonne **pactols**, il peut également être intéressant de récupérer le *preflabel* français. Cet élément nous permet de vérifier les traductions anglaises déjà importées, mais sera également utile dans la constitution du document final (voir partie 3).

Expression GREL :

```
value.find(/".*"@fr/)[0].replace(/.*"(.*)"@fr.*/, "$1")
```

2. Récupérer les *broaderlabels*

Le service de réconciliation de l'AAT – que l'on présentera dans une partie suivante (partie 2) – offre la possibilité de préciser le rapprochement de vocabulaire en comparant d'autres propriétés que les stricts caractères composants le terme en lui-même. Il est donc possible de comparer les définitions, les termes parents (*broaderlabels*) ou encore des synonymes. Parce que les définitions et les synonymes sont en français dans les PACTOLS, nous avons choisi d'utiliser les termes parents pour préciser la réconciliation.

A. Récupérer les URI

La récupération de cette information nécessite plusieurs étapes. En effet, ce n'est pas le terme parent qui est contenu dans les informations PACTOLS, mais son URI. Il faut donc d'abord extraire cette information en créant une nouvelle colonne **Broader** d'après les informations contenues dans la colonne **pactols** et en y séléctionnant l'information désirée à l'aide d'une expression GREL.

Expression GREL :

```
value.find(/broader.*/)[0].replace(/.*(http.*)>.*/,"$1")
```

B. Moissonner les URL

Il est alors possible de commencer une nouvelle opération de moissonnage à partir de cette nouvelle URI en réutilisant les mêmes fonctions que précédemment (voir partie 1.2.1 et figure 1.3), au sein d'une colonne appelée **broader_pactols**.

Expression GREL*(pour rappel)***:**

```
"https://pactols2.frantiq.fr/pactols2/api/" + value.replace(/.*\/(\d+.*)/, "$1")
+ ".ttl"
```

C. Récupérer le terme en anglais

Il faut donc maintenant récupérer ce terme en anglais en effectuant exactement les mêmes démarches que lors de la récupération du *preflabel* en anglais (voir partie 1.2.2). On crée donc une nouvelle colonne **borader_label(en)** d'après les valeurs de la colonne **broader_pactols** dans laquelle on va sélectionner le *preflabel* en anglais. Il s'agit bien ici du *preflabel* et non du *broaderlabel* puisqu'il s'agit de données importées depuis l'URI de ce terme et non depuis celles du terme réellement recherché.

Expression GREL*(pour rappel)***:**

```
value.find(/".*"@en/)[0].replace(/.*"(.*)"@en.*/, "$1")
```

De la même manière que pour le *preflabel* précédemment, une nouvelle colonne **broader_label2** est créée, dans laquelle les possibles précisions entre parenthèses sont supprimées.

Expression GREL*(pour rappel)***:**

```
value.replace(/(.*)\(.*\)/, "$1")
```

3. Récupérer les alignements effectués par PACTOLS

Dans la nouvelle version des PACTOLS (à paraître[6]), un certain nombre de termes a déjà été aligné aux vocabulaires de l'AAT. Il est possible de récupérer cet alignement en suivant la même méthode que pour les informations précédentes. On crée donc une colonne **pactols-AAT** d'après les informations contenues dans la colonne **AAT 1** sur lesquelles on applique une nouvelle expression GREL.

Expression GREL :

```
value.find(/.*exactMatch.*<http:\/\/vocab.*>/)[0].replace(/.*exactMatch.*<(http:\/\/voca
"$1"))
```

Les colonnes correspondant aux différents moissonnages de données (**pactols** et **broader_pactols**) ne nous seront plus utiles par la suite et peuvent donc être supprimées. Les données sont désormais prêtes pour l'étape de réconciliation (voir figure 1.5).

Toutes	URI	Preflabel(en)	Preflabel(en)2	preflabel(fr)	broader_label(e	broader_label2	AAT 1
1.	https://ark.frantiq.fr/ark:/26678/crt5AOuUmbGVy			établissement rural			
2.	https://ark.frantiq.fr/ark:/26678/crt9ey83pmtqt			objet en terre cuite			
3.	https://ark.frantiq.fr/ark:/26678/crtEd5UW8qzVu			tesselle			http://vocab.getty.edu/aat/300010730
4.	https://ark.frantiq.fr/ark:/26678/crtP1RhPYYLc1	crest (heraldry)	crest	cimier	helmet	helmet	
5.	https://ark.frantiq.fr/ark:/26678/crtP7JUFfkSYS			vantail	opening	opening	
6.	https://ark.frantiq.fr/ark:/26678/crtq6ld0rakNf			découverte fortuite			
7.	https://ark.frantiq.fr/ark:/26678/crtSrWQs2w2KV			fouille programmée			
8.	https://ark.frantiq.fr/ark:/26678/crtVqcSVS0rm7			prospection pédestre	survey	survey	
9.	https://ark.frantiq.fr/ark:/26678/crtWAS0j7kGTX			objet en verre			
10.	https://ark.frantiq.fr/ark:/26678/pcrt019SSqbR7W	tile	tile	carreau			http://vocab.getty.edu/aat/300010676
11.	https://ark.frantiq.fr/ark:/26678/pcrt02KdlM0gwJ	basin (structure)	basin	bassin (structure)	built unit by form	built unit by form	
12.	https://ark.frantiq.fr/ark:/26678/pcrt0zPNjcnqaK	roman box flue tile	roman box flue tile	tubulure	part of building	part of building	
13.	https://ark.frantiq.fr/ark:/26678/pcrt1DMOWvDF4j	pottery	pottery	céramique (matériau)	terracotta	terracotta	http://vocab.getty.edu/aat/300235507
14.	https://ark.frantiq.fr/ark:/26678/pcrt1dndv82zn9	pipe	pipe	tuyau			
15.	https://ark.frantiq.fr/ark:/26678/pcrt1eKrSANesO	way	way	voie terrestre	public road network	public road network	
16.	https://ark.frantiq.fr/ark:/26678/pcrt1Gqf0UzaZ3	millstone	millstone	meule	grinding tools	grinding tools	
17.	https://ark.frantiq.fr/ark:/26678/pcrt1tSmMf0Tw8	fibula	fibula	fibule	garment fastener	garment fastener	http://vocab.getty.edu/aat/300209303

Figure 1.5: Extrait du jeu de données obtenu après ces différentes manipulations

II. Faire la réconciliation[7]

La réconciliation des données permet de rapprocher chaque valeur du jeu de données sources vers de possibles candidats d'un vocabulaire contrôlé. Elle permet par la suite d'enrichir les données sources en important de nouvelles informations depuis ce vocabulaire de référence, telles que son identifiant, sa définition, son *preflabel* ou tout autre élément jugé utile[8]. Il s'agit donc également d'un alignement de vocabulaire, mais au lieu d'utiliser une API comme c'était le cas pour les PACTOLS, celui-ci repose sur un service spécialement conçu pour OpenRefine[9].

1. Le service de réconciliation du Getty AAT

A. Installer le service

La première étape consiste à installer le service de réconciliation du Getty AAT dans OpenRefine. Pour ce faire, on sélectionne la colonne à aligner – en l'occurrence **preflabel(en)2** – et l'option **"Démarrer la réconciliation"** dans **"Réconcilier"** (figure 2.1). Lors d'une première utilisation, seul le service de réconciliation de Wikidata est installé. Il faut donc ajouter le service du Getty AAT en cliquant sur le bouton **"Ajouter un service standard"** (figure 2.2) puis indiquer l'adresse du service de réconciliation de l'AAT : `http://services.getty.edu/vocab/reconcile/` (figure 2.3).

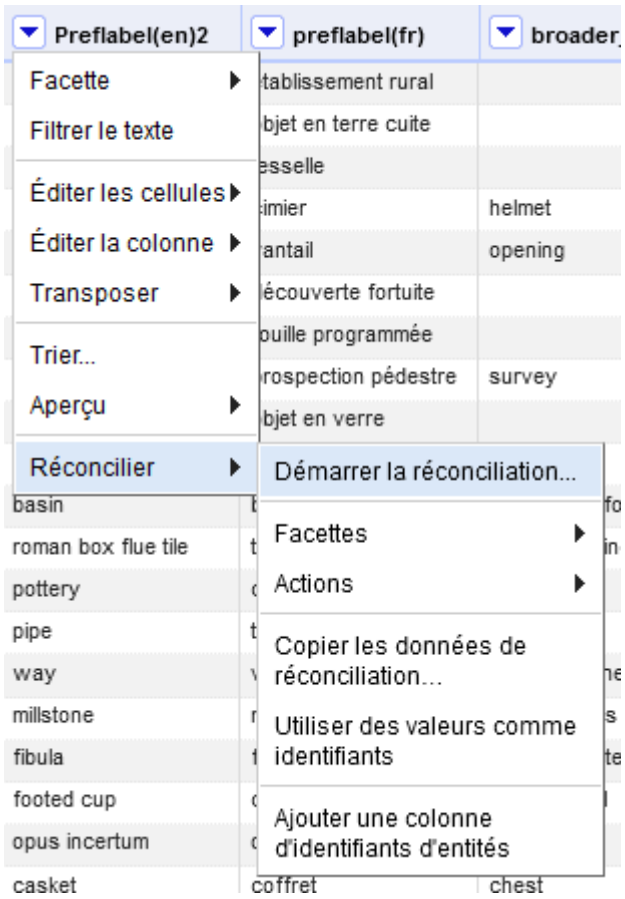


Figure 2.1: Ouvrir le service de réconciliation

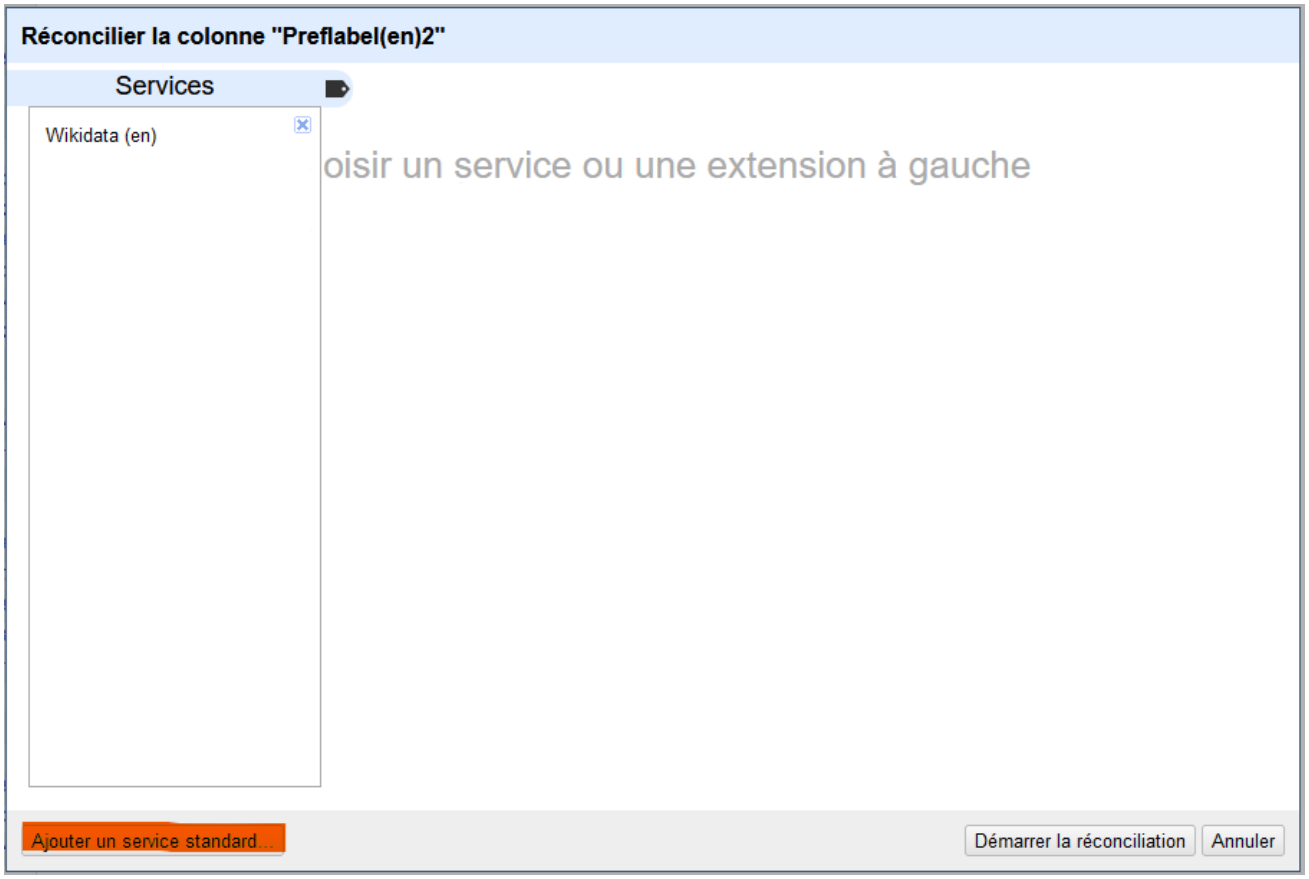


Figure 2.2: Ajouter un service de réconciliation

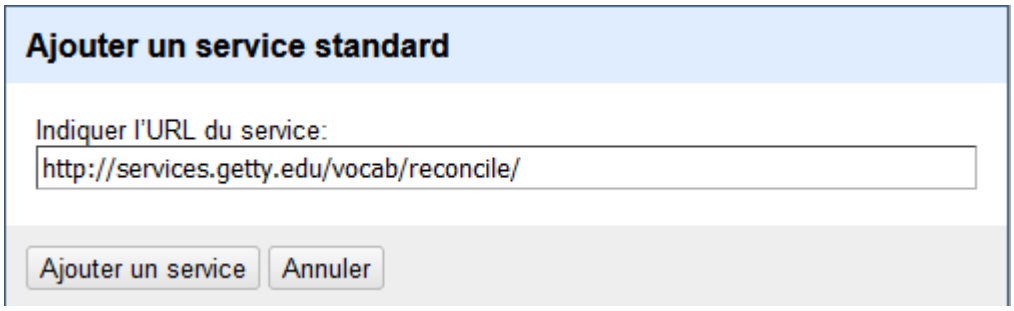


Figure 2.3: Indiquer l'adresse du service

Le *"Getty Vocabulary Reconciliation Service"* est désormais installé dans l'ordinateur et peut être sélectionné dans l'espace à gauche de la fenêtre (figure 2.4).

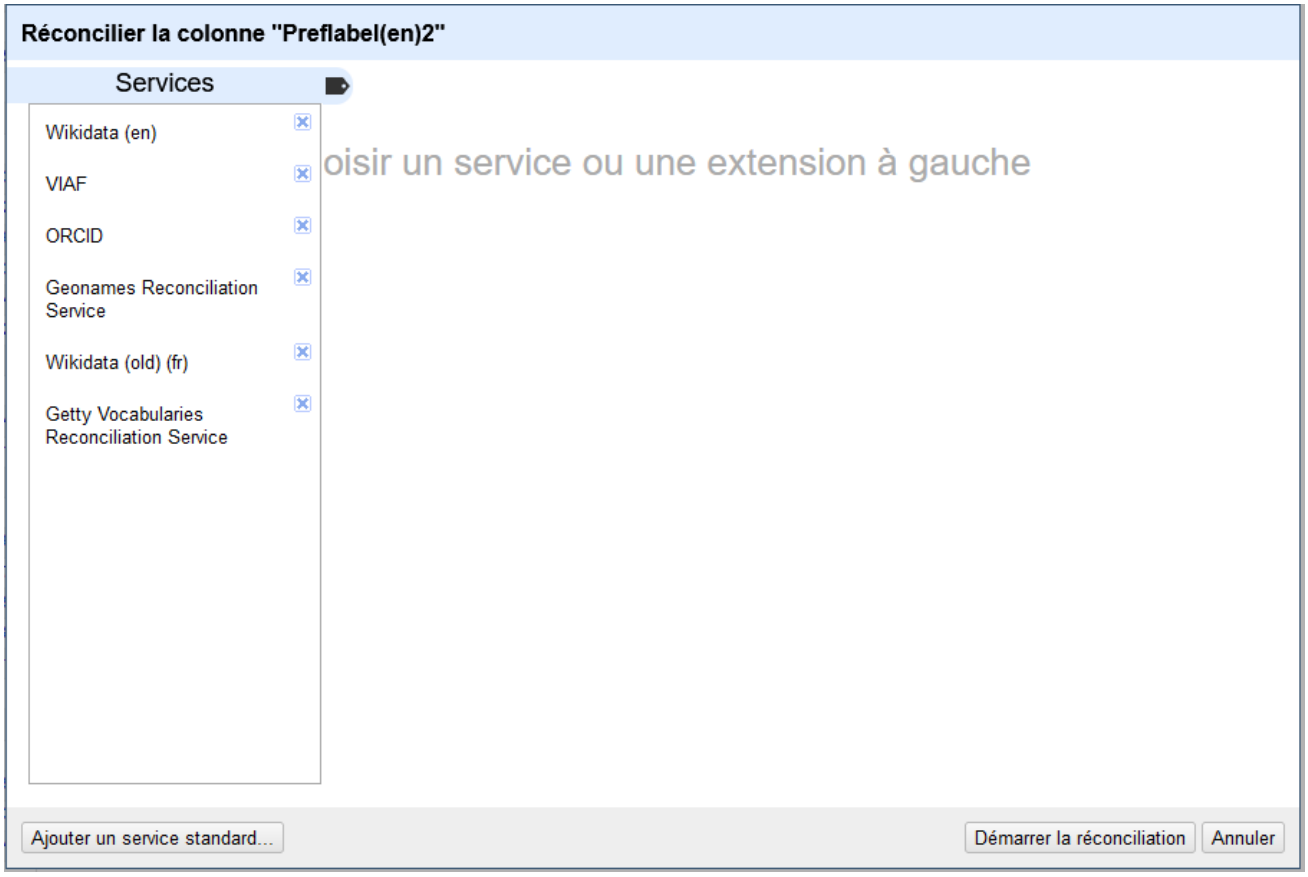


Figure 2.4: Le service de réconciliation de l'AAT est désormais disponible dans OpenRefine

B. Lancer la réconciliation

Le Getty propose la réconciliation avec plusieurs vocabulaires contrôlés qui apparaissent dans la partie gauche de la fenêtre (figure 2.5). Dans le cas qui nous intéresse, il faut cocher "AAT".

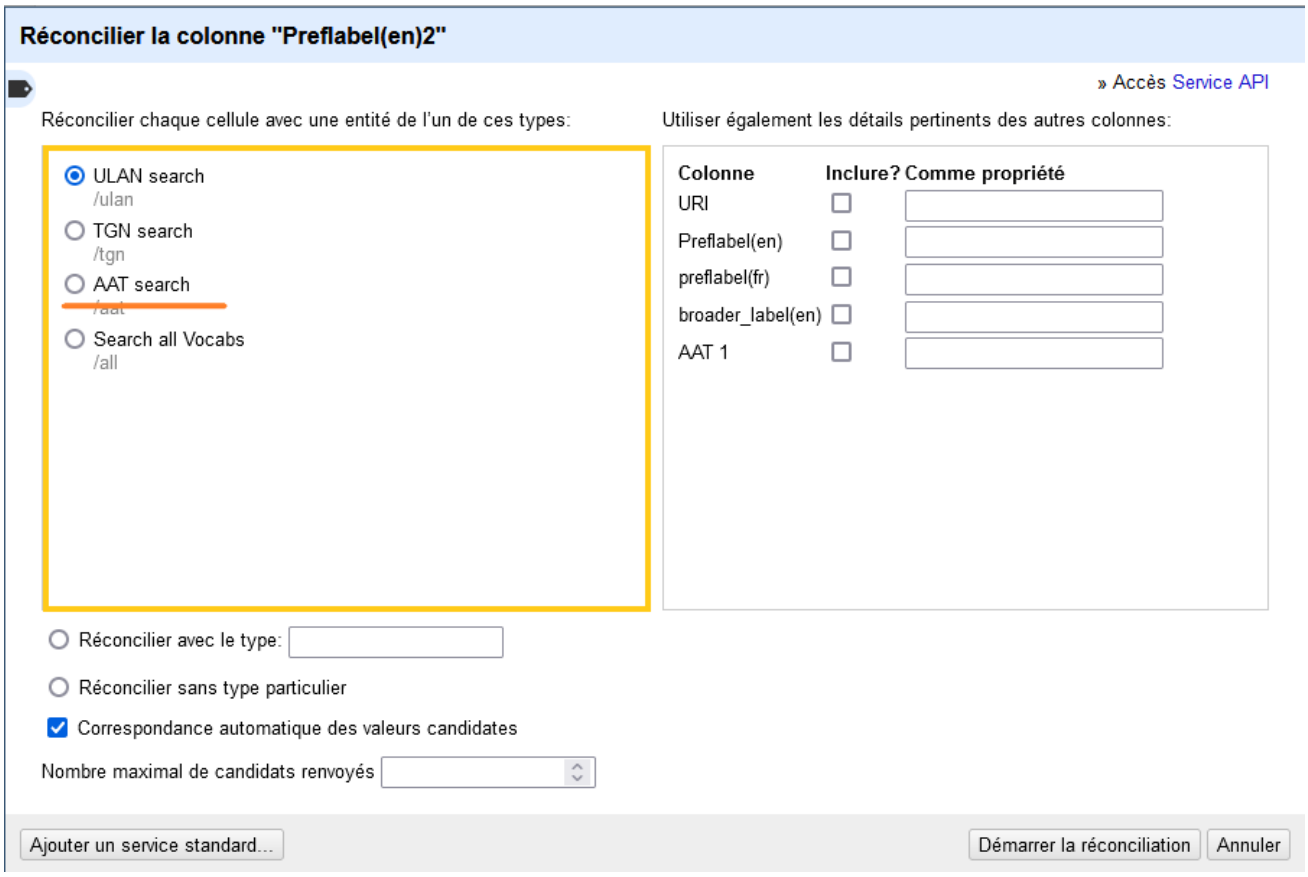


Figure 2.5: Choisir le bon vocabulaire

Comme cela a déjà été mentionné (partie 1.3), la réconciliation peut être améliorée en croisant des informations relatives au terme recherché. Cette possibilité est particulièrement utile pour différencier des homonymes ; du mortier de construction n'aura, par exemple, pas le même terme parent que le récipient appelé mortier. Cet élément peut être ajouté dans la colonne de droite en cochant la colonne du jeu de données sources où se situe l'information (ici **broader_label(en)2**) et en indiquant à quelle propriété du vocabulaire cible elle se rapporte. Il faut inscrire ici **"broaderExt"** (figure 2.6)[10].

Openrefine propose aussi d'aligner automatiquement les termes pour lesquels il détecte un haut pourcentage de correspondance en cochant l'option **"Correspondance automatique des valeurs candidates"** (figure 2.6). Cette fonctionnalité est à utiliser avec précaution et donne des résultats plus ou moins bons selon les référentiels recherchés. Dans la plupart des cas il vaut mieux vérifier manuellement les résultats obtenus pour en certifier la concordance.

Une fois ces éléments indiqués, il est possible de démarrer la réconciliation. Cette opération peut prendre un certain temps, c'est pourquoi il était important de supprimer les éventuels doublons au début de ce tutoriel (partie 1.1). De la même manière, dans le cas de jeux de données conséquents, il est préférable de les diviser.

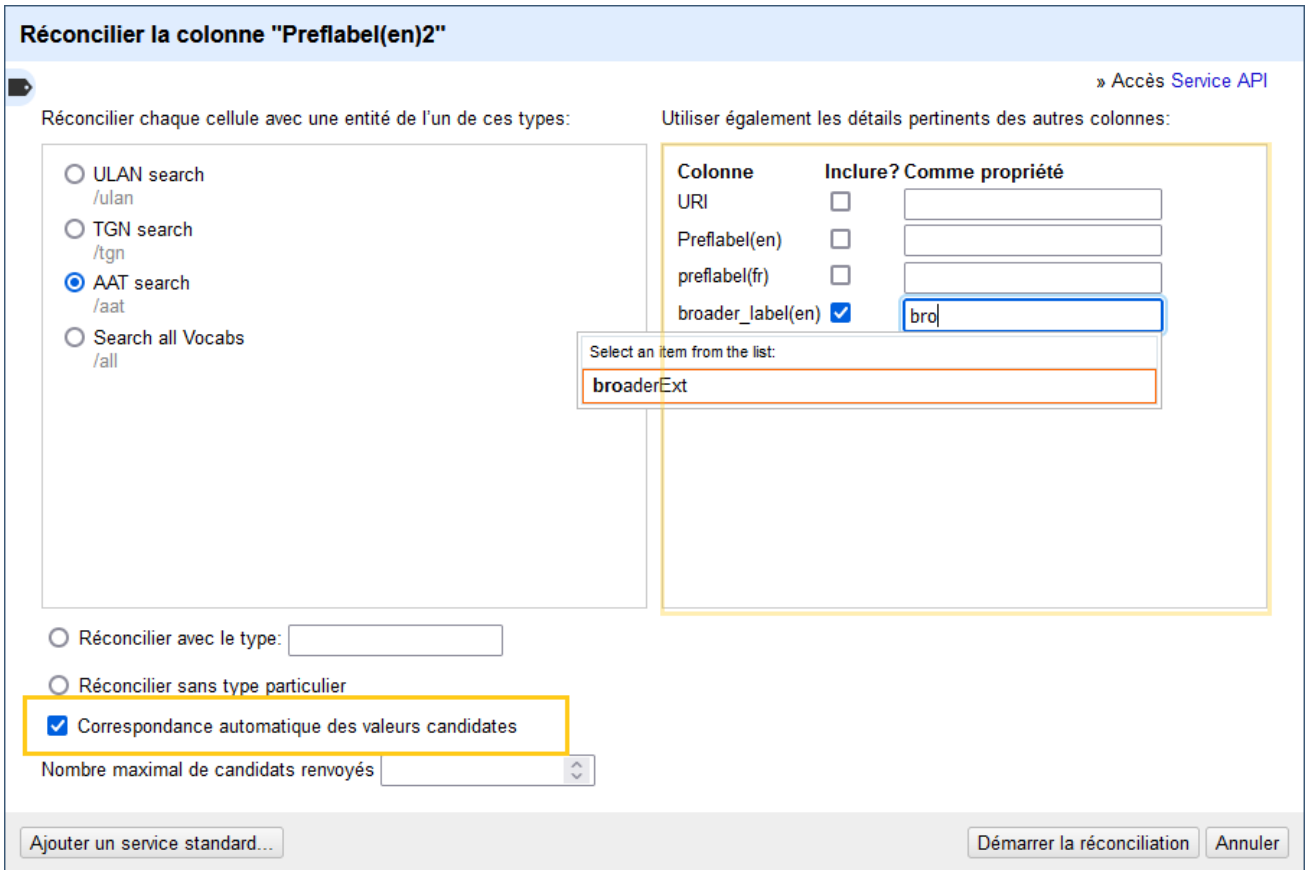


Figure 2.6: Choisir le bon vocabulaire

C. Valider la réconciliation

Le résultat de la réconciliation consiste en une liste de possibles correspondances pour chaque terme (figure 2.7). Bien entendu, il arrive qu'aucun alignement ne puisse être établi avec le référentiel cible (ces éléments seront traités dans la partie suivante 2.4). On remarque que le tableur en lui-même n'a pas été modifié et qu'il n'existe toujours qu'une cellule pour chaque terme. Par ailleurs, le contenu des cellules n'est pas réellement modifié non plus, elles ne contiennent toujours que le terme.

Toutes		URI	Preflabel(en)	Preflabel(en)2	preflabel(fr)	broader_label(e	broader_label2	AAT 1
	1.	https://ark.frantiq.fr/ark:/26678/crt5AOuUmbGVy			établissement rural			
	2.	https://ark.frantiq.fr/ark:/26678/crt9ey83pmtqt			objet en terre cuite			
	3.	https://ark.frantiq.fr/ark:/26678/crtEd5UW8qzVu			tesselle			http://vocab.getty.edu/aat/300010730
	4.	https://ark.frantiq.fr/ark:/26678/crtP1RhPYYLc1	crest (heraldry)	crest crest rails (16) crest tile (14) crests (motifs) (14) Créer un nouveau sujet	cimier	helmet	helmet	
	5.	https://ark.frantiq.fr/ark:/26678/crtP7JUFfkSYS			vantail	opening	opening	
	6.	https://ark.frantiq.fr/ark:/26678/crtq6ld0rakNf			découverte fortuite			
	7.	https://ark.frantiq.fr/ark:/26678/crtSrWQs2w2KV			fouille programmée			
	8.	https://ark.frantiq.fr/ark:/26678/crtVqcSVS0rm7			prospection pédestre	survey	survey	
	9.	https://ark.frantiq.fr/ark:/26678/crtWAS0j7kGTX			objet en verre			
	10.	https://ark.frantiq.fr/ark:/26678/pcrt019SSqbR7W	tile	tile tiles (visual works) (25) tiles (object genre) (24) tile (material) (23) tile makers (12) tile painters (12) floor tile (12) paving tile (12) ridge tile (12) mathematical tile (12) clay tile (12) asbestos tile (12) ventilating tile (11) tile work (visual works) (11) cover tile (11) Spanish tile (11) encaustic tile (11) barrel tile (11) metal tile (11) heavy tile (11) large tile (11) rounded tile (11) concrete tile (11) copper tile (11) flat tile (11) roofing tile (11) Créer un nouveau sujet	carreau			http://vocab.getty.edu/aat/300010676
	11.	https://ark.frantiq.fr/ark:/26678/pcrt02KdlM0gwJ	basin (structure)	basin enclosed basin stands (13) basins (landforms) (12) Great Basin Native American styles (11) basin stands (11) medical basin (11) corner enclosed basin stands (10) flushing basin (10) laboratory basin (10) serving basin (10) rinsing basin (10) Pre-Columbian Great Basin styles (10)	bassin (structure)	built unit by form	built unit by form	

Figure 2.7: Résultat de la réconciliation

Cette liste de possibles concordances se compose en réalité d'hyperliens vers le référentiel. Passer la souris sur l'un des termes ouvre un encadré dans lequel on retrouve la définition et en-dessous l'arborescence du terme choisi, c'est-à-dire les éléments permettant de valider ou non la concordance entre les deux termes. Cliquer dessus ouvre la page du Getty AAT lui correspondant et permet de voir les différentes traductions du terme, les synonymes et éventuellement de naviguer dans l'arborescence pour trouver un meilleur alignement. Si le terme analysé correspond, il suffit de cliquer sur le bouton **"Apparier cette cellule"** pour valider la réconciliation du terme (figure 2.8). Si le même terme, porteur de la même signification, peut se répéter dans le jeu de données, il est aussi possible de cliquer sur le bouton **"Apparier toutes les cellules identiques"** pour qu'il soit aligné dans l'ensemble du document. Cette opération est à refaire pour chaque terme[11].

Research

Research Home ▸ Tools ▸ Art & Architecture Thesaurus ▸ Full Record Display

Art & Architecture Thesaurus® Online
Full Record Display

New Search Previous Page Help

Click the icon to view the hierarchy.

Semantic View (JSON, RDF, N3/Turtle, N-Triples) Representative Images: 1 2 3 4

ID: 300410344 **Record Type:** concept
Page Link: http://vocab.getty.edu/page/aat/300410344

tiles (visual works) (<visual works by form>, visual works (works), ... Visual and Verbal Communication (hierarchy name))

Note: Visual works for which tile is the support, such as painting, or for which tile is the medium into which a relief is pressed or carved, or encaustic design is created. For visual works comprising multiple tiles, whether or not individual tiles are decorated, use "tile works (visual works)." For tile as a material in roofing, etc., use "tile (material)" or a more specific narrower context.

Terms:
tiles (visual works) (**preferred**,C,U,English-P,D,U,PN)
tile (visual work) (C,U,English,AD,U,SN)
art tiles (visual works) (C,U,English,UF,U,U)
art tile (visual works) (C,U,English,UF,U,U)
tiles, art (visual works) (C,U,English,UF,U,U)
siertegels (beeldende werken) (C,U,Dutch-P,D,U,U)
siertegel (beeldend werk) (C,U,Dutch,AD,U,U)

Facet/Hierarchy Code: V.VC

Hierarchical Position:
Objects Facet
.... Visual and Verbal Communication (hierarchy name) (G)
..... Visual Works (hierarchy name) (G)
..... visual works (works) (G)
..... <visual works by form> (G)
..... tiles (visual works) (G)

Figure 2.8: Détail des choix possibles

2. Enrichir les données après la réconciliation

La réconciliation des données avec un référentiel externe permet aussi de récupérer des informations présentes dans ce référentiel et de les ajouter aux données sources. Cette opération s'appelle un enrichissement des données.

Puisque la réconciliation effectuée ici a pour objectif d'aligner des vocabulaires sur le référentiel du Getty, nous avons besoin de récupérer les URI de l'AAT. Pour ce faire, il faut créer une nouvelle colonne (**AAT URI**) d'après la colonne **Preflabel(en)2** qui contient les données réconciliées. L'utilisation d'une nouvelle expression GREL permet d'extraire de ces données l'identifiant unique de l'AAT (figure 2.9). On peut aussi y ajouter les éléments nécessaires à la reconstitution de cette URI.

Expression GREL :

```
"http://vocab.getty.edu/" + cell.recon.match.id
```

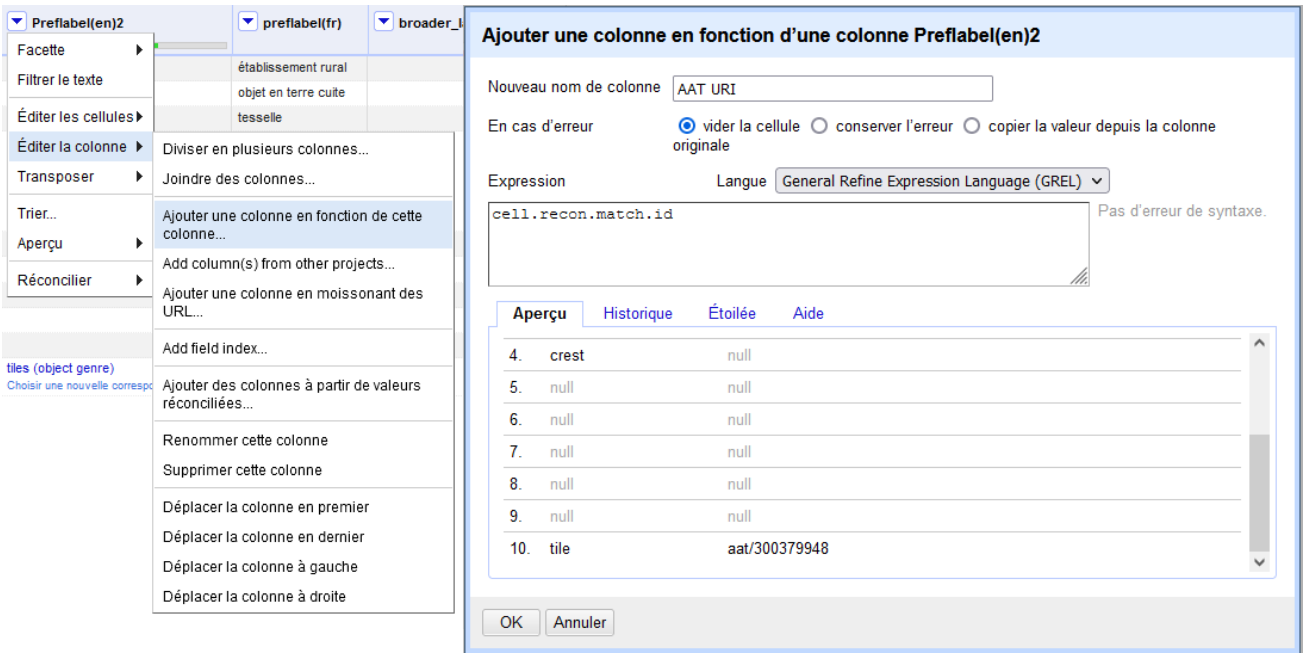
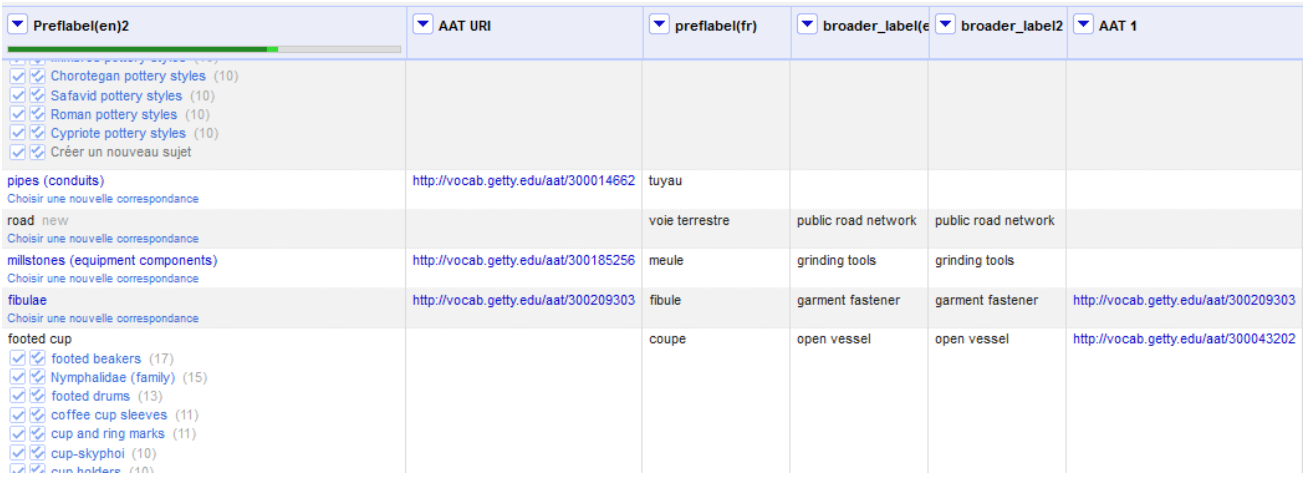


Figure 2.9: Récupérer les identifiants uniques de l'AAT

3. Comparer les URI

Le jeu de données contient désormais deux séries d'URI du Getty à AAT, l'une provenant de l'alignement PACTOLS (voir partie 1.4) et l'autre de l'enrichissement des données après réconciliation (voir partie 2.2). Pour certains termes, un alignement n'est proposé que dans une de ces deux colonnes, pour d'autre, une URI – identique ou différente – est proposée dans chacune et pour certains aucune équivalence n'a pu être établie à ce stade (figure 2.10).



Preflabel(en)2	AAT URI	preflabel(fr)	broader_label(e	broader_label2	AAT 1
Chorotegan pottery styles (10) Safavid pottery styles (10) Roman pottery styles (10) Cypriote pottery styles (10) Créer un nouveau sujet					
pipes (conduits) Choisir une nouvelle correspondance	http://vocab.getty.edu/aat/300014662	tuyau			
road new Choisir une nouvelle correspondance		voie terrestre	public road network	public road network	
millstones (equipment components) Choisir une nouvelle correspondance	http://vocab.getty.edu/aat/300185256	meule	grinding tools	grinding tools	
fibulae Choisir une nouvelle correspondance	http://vocab.getty.edu/aat/300209303	fibule	garment fastener	garment fastener	http://vocab.getty.edu/aat/300209303
footed cup footed beakers (17) Nymphalidae (family) (15) footed drums (13) coffee cup sleeves (11) cup and ring marks (11) cup-skyphoi (10)		coupe	open vessel	open vessel	http://vocab.getty.edu/aat/300043202

Figure 2.10: Extrait du jeu de données après réconciliation et enrichissement

Selon la taille du jeu de données, il peut être intéressant de diviser ces différents éléments en fonction de cet état. Il s'agit de créer une nouvelle colonne **comparaison** pour comparer l'alignement PACTOLS et la réconciliation, qui pourra aussi être utilisée pour trier les termes selon leur statut d'alignement à l'aide d'une facette.

Cette nouvelle colonne est créée d'après les données de l'une ou l'autre des deux colonnes contenant des URI provenant de l'AAT. Une expression GREL de condition, introduite par "if", nous permet ensuite de comparer ces deux informations (figure 2.11). Ainsi, si la réconciliation n'a pas permis de relier le terme source au référentiel, l'information sera "Manquante". À l'inverse, si les PACTOLS ne proposent pas d'alignement avec ce terme, la comparaison est inutile et "Aucun" alignement n'est effectué. Si les URI présentes dans les deux colonnes sont similaires, la comparaison est "Identique", dans le cas contraire elle est "Différente".

Expression GREL :[12]

```
if(cells["AAT URI"].value == null, "Manquante", if(cells["AAT 1"].value == null, "Aucun", if(cells["AAT URI"].value == cells["AAT 1"].value, "Identique", "Différent")))
```

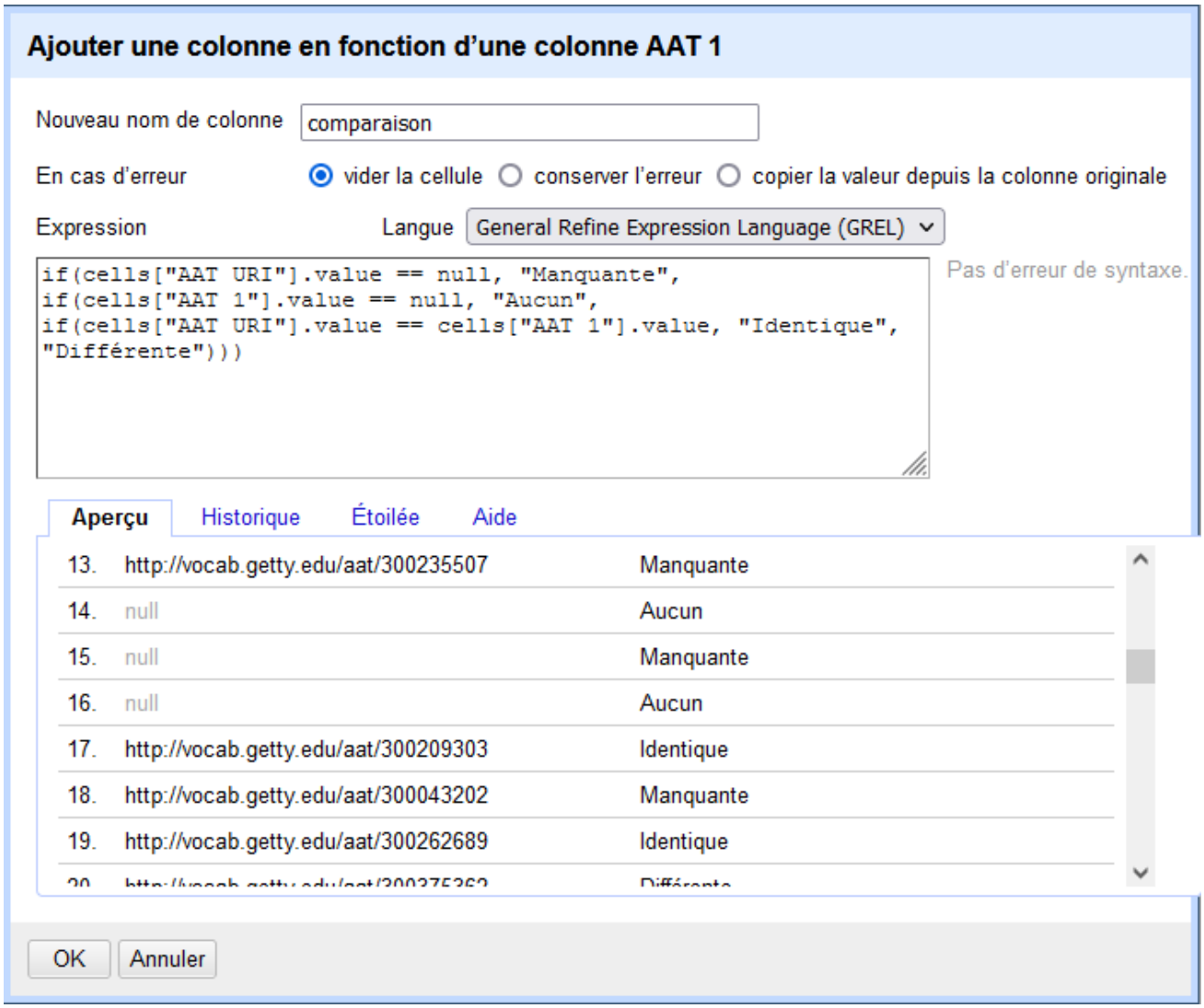


Figure 2.11: Création de la colonne ***comparaison***

Il peut aussi être utile de distinguer les termes pour lesquels la réconciliation n'a pas été effectuée, mais qui possédaient déjà un alignement dans les PACTOLS.

Pour faire cette distinction, il faut d'abord réaliser une facette textuelle sur la colonne **comparaison**, qui permet de filtrer les données en sélectionnant uniquement les lignes pour lesquelles l'information "Manquante" est indiquée (figure 2.12). Ces éléments sont ensuite transformés – en utilisant l'option "**Transformer**" dans l'onglet "**Éditer les cellules**" – à l'aide d'une nouvelle expression GREL (figure 2.13). Ainsi, si la ligne contient une URI provenant de l'alignement PACTOLS, la comparaison indiquera "Pactols", sinon sa valeur sera toujours "Manquante".

Expression GREL :

```
if(cells["AAT 1"].value == null, value, "Pactols")
```

Figure 2.12: Réaliser une facette textuelle à partir de la colonne ***comparaison***

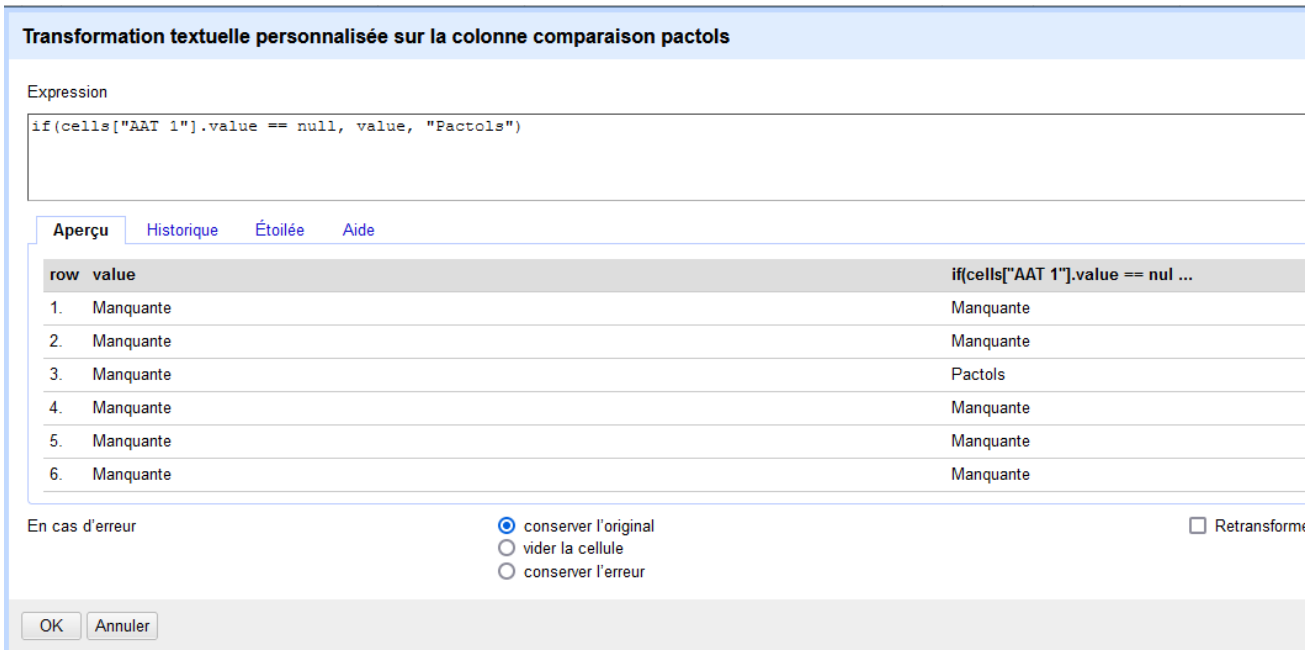


Figure 2.13: Transformer les cellules "Manquante" dans la colonne ***comparaison***

Enfin, avant de passer à la recherche de concordances pour les termes manquants (partie 2.4), une autre information peut être ajoutée pour faciliter la création du fichier pour ARIADNEplus que nous développerons dans une prochaine partie (partie 3). En effet, dans ce document il est demandé de préciser la nature de l'alignement d'après le standard SKOS[13] (voir partie 3.1). Puisque l'ensemble des URI provenant de l'alignement PACTOLS correspond à un *exact match* SKOS, cette information (`http://www.w3.org/2004/02/skos/core#exactMatch`) peut déjà être indiquée de manière automatique.

On peut donc créer une nouvelle colonne **skos match** d'après les informations contenues dans la colonne **comparaison** et y modifier les valeurs à l'aide d'une autre expression GREL de condition précisant que si les données sont "Identiques" ou "Pactols" alors le SKOS *match* sera exact.

Expression GREL :

```
if(cells["comparaison"].value == "Identique",
"http://www.w3.org/2004/02/skos/core#exactMatch", if(cells["comparaison"].value
```

```
== "Pactols", "http://www.w3.org/2004/02/skos/core#exactMatch", null))
```

4. Rechercher les concordances manquantes

Au terme de ces différentes manipulations, il reste encore un certain nombre de termes pour lesquels aucune concordance n'a pu être déterminée, en plus de ceux pour lesquels l'alignement proposé par PACTOLS et celui établi par la réconciliation sont différents. Pour pallier ces manques et problèmes, il est nécessaire de parcourir et rechercher plus en détail le thesaurus de l'AAT.

Le logiciel OpenRefine n'étant pas prévu pour être un outil de saisie, mais bien de manipulation de données, il peut être d'abord préférable d'exporter le nouveau jeu de données (figure 2.14) au format csv (*comma separated values*) et travailler directement dans un logiciel de tableur tel que Calc ou Excel.

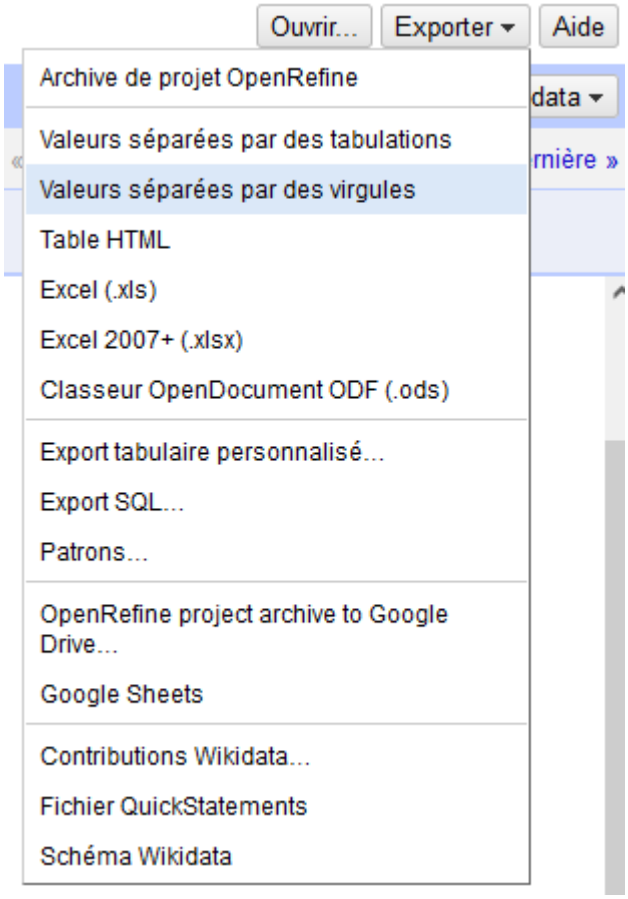


Figure 2.14: Exporter les données au format csv

Une dernière manipulation de données peut être effectuée pour faciliter la recherche manuelle de correspondance. Les URI récupérées pour l'AAT (par exemple : http://vocab.getty.edu/aat/300010676) mènent vers une page difficilement lisible et sur laquelle l'arborescence des termes n'est pas visible. L'ajout d'un élément dans cette URI permet d'accéder au site du Getty AAT et à l'ensemble de ces informations (http://vocab.getty.edu/page/aat/300010676). Cette manipulation peut être réalisée en une seule opération grâce à l'utilisation d'une expression régulière[14] dans l'outil "**Rechercher et Remplacer**".

Expression régulière :

Rechercher :

`(http.*edu\/)(.*)$`

Remplacer par :

`$1page/$2`

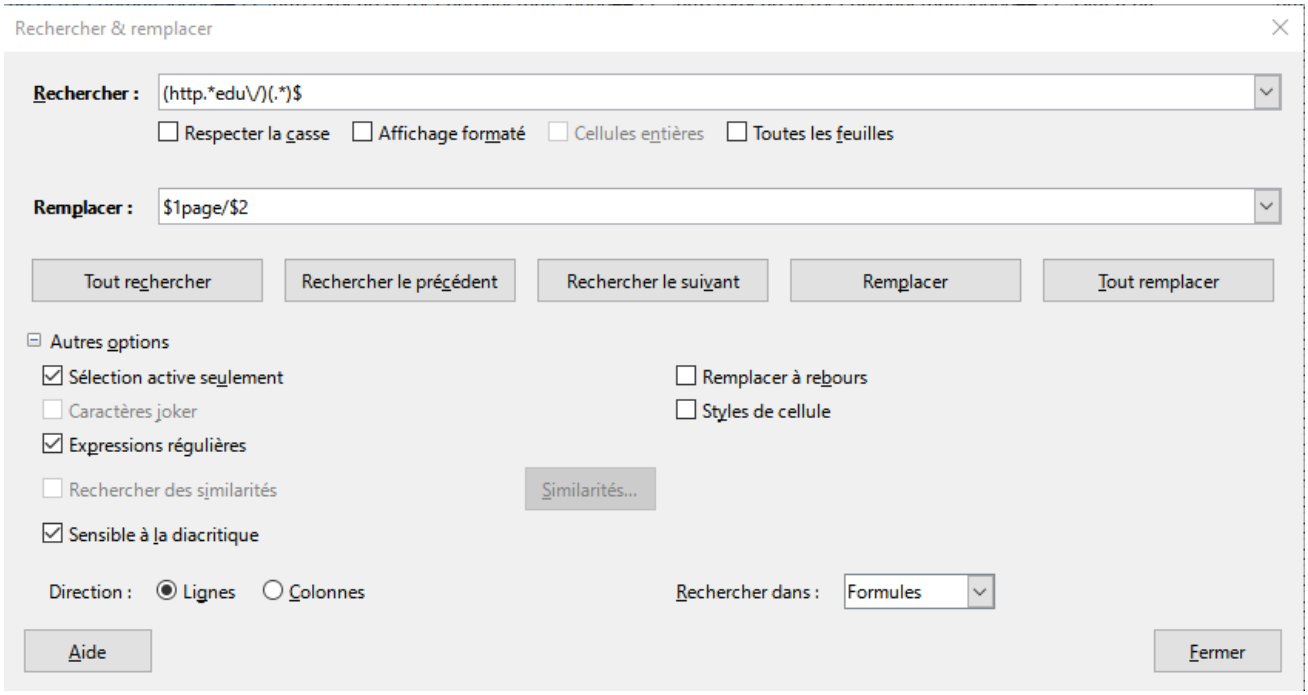


Figure 2.15: Utiliser une expression régulière pour modifier les URI dans Calc [15]

Il est maintenant possible de trier les données en fonction de la colonne **comparaison** pour exclure celles pour lesquelles l'alignement correspond à "Identique" ou "Pactols" – et est donc déjà validée – et se concentrer sur celles où il est "Différent". Pour ces cas, il faut consulter les deux termes proposés et sélectionner le plus approprié au jeu de données sources en s'appuyant sur les définitions proposées par l'AAT et par PACTOLS, ainsi que sur les arborescences et en particulier les termes parents.

Pour les termes dont la comparaison est absente (mentionné "Aucune") ou manquante, une nouvelle recherche au sein de l'AAT peut être bénéfique. Dans ce cas, il faut chercher une possible correspondance à l'aide des termes synonymes, d'une autre traduction en anglais, du terme en français ou éventuellement dans une autre langue. Lorsque qu'une concordance est trouvée, son URI peut être récupérée à côté de la mention "*Page Link*" en haut de la page web de l'AAT (figure 2.16).

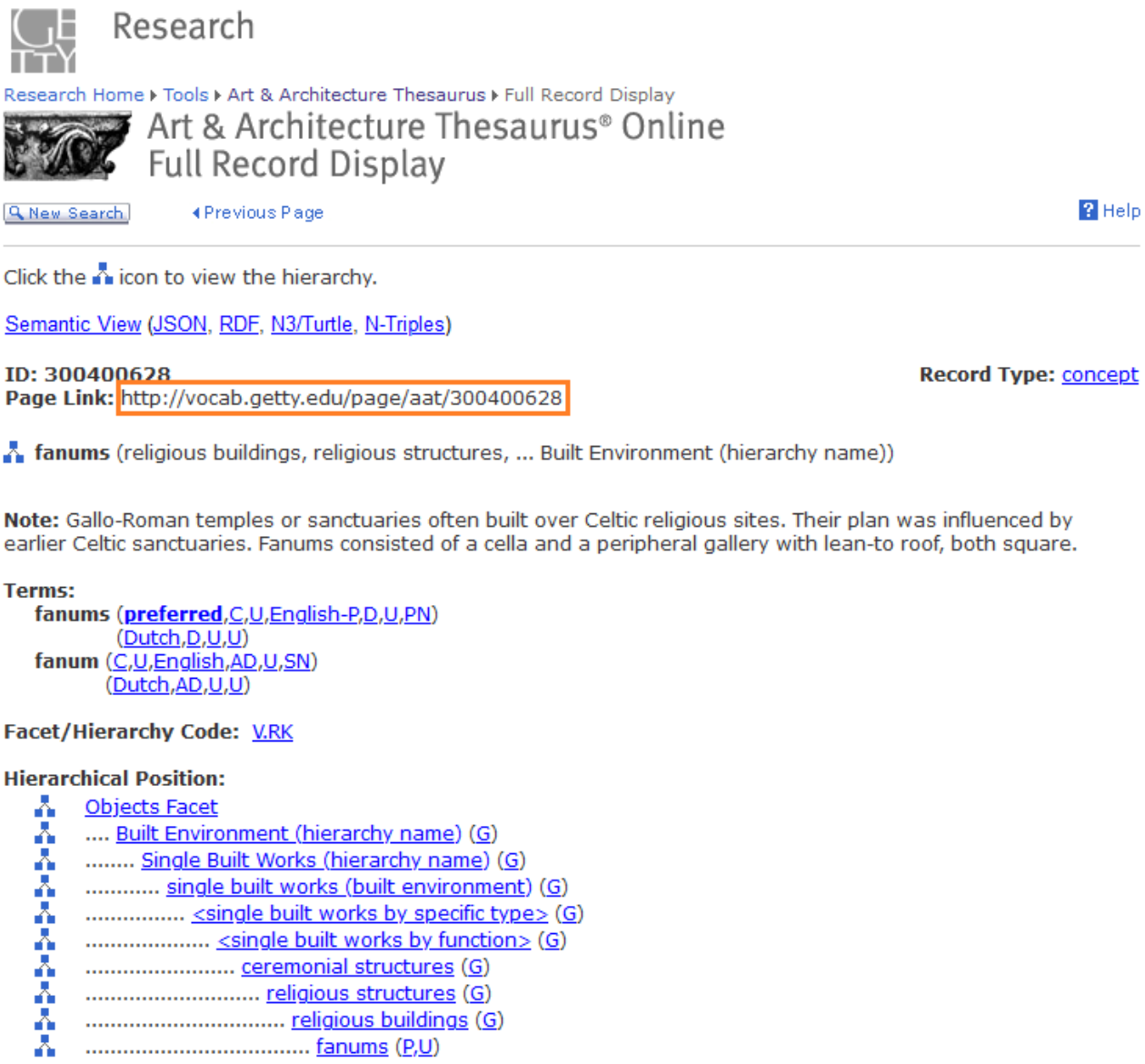


Figure 2.16: Récupérer l'URI sur l'AAT

Selon les cas, il est possible de ne pas pouvoir trouver de correspondance exacte : les définitions ne correspondent pas, les arborescences sont différentes ou encore le terme source n'existe simplement pas dans le vocabulaire de référence. Toutefois, dans ces cas, il est possible de rapprocher le terme source d'un autre terme dans le vocabulaire de référence qui peut être son parent, c'est-à-dire qui possède un sens plus large, ou qui peut être proche en se situant dans une arborescence de concepts différents par exemple. Il est conseillé, lorsque de tels choix sont faits, de tout de suite le préciser dans la colonne **SKOS match** par *broad match* (`http://www.w3.org/2004/02/skos/core#broadMatch`) pour un terme parent ou *close match* (`http://www.w3.org/2004/02/skos/core#closeMatch`) pour un terme proche.

III. Préparer le fichier pour ARIADNEplus

Pour pouvoir déposer un jeu de données dans la plateforme ARIADNEplus, il est nécessaire de créer un tableau d'équivalence entre les vocabulaires sources et ceux du Getty AAT. Ce document doit donc contenir les identifiants pérennes des deux référentiels – en l'occurrence ceux de PACTOLS et de l'AAT – ainsi que leur *preflabel*, la langue du vocabulaire source, le SKOS *match*, déjà présenté plus haut (partie 2.4), et la date de création. Il peut se présenter au format JSON ou csv qui sera celui utilisé dans l'exemple suivant.

Parce que pour obtenir certaines informations – en particulier les *labels* de l'AAT – il est nécessaire de moissonner une URL, le document csv généré précédemment est à nouveau importé dans OpenRefine. Par ailleurs, un certain nombre de colonne peut être supprimé de ce nouveau document pour ne conserver que celles nécessaires au fichier pour

ARIADNEplus, c'est-à-dire les colonnes contenant : les URI PACTOLS que l'on renomme **sourceURI**, les URI de l'AAT que l'on renomme **targetURI**, les SKOS *match* que l'on renomme **matchURI** et les *preflabels* PACTOLS en français que l'on renomme **sourceLabel**. De plus, pour que le lien vers le vocabulaire de l'AAT corresponde à l'URI et non plus au lien vers l'URL, il faut enlever l'élément `page/` que l'on y avait ajouté précédemment. Pour ce faire, il est possible d'utiliser la fonction "**Remplacer**" de OpenRefine en y indiquant une nouvelle expression régulière. On peut aussi utiliser directement la fonction "**Transformer**" dans l'onglet "**Cellule**" en y indiquant la même expression régulière sous la forme d'une fonction GREL.

Expression régulière :

Rechercher :

```
(http.*)\/page(.*)$
```

Remplacer par :

```
$1$2
```

Expression GREL :

```
value.replace(/(htt.*)\/page(.*$)/, "$1$2")
```

1. Les SKOS *match*

Les SKOS (*Simple Knowledge Organization System*)* sont un modèle de données partagé pour échanger et des systèmes d'organisation de connaissances sur le web. Ils sont utilisés pour expliciter les ressemblances entre des référentiels partagés et ainsi faciliter l'échange de données entre les applications[16]. Pour rappel, ils sont utilisés dans ARIADNEplus pour préciser la nature de l'alignement entre le référentiel source et celui de l'AAT.

Normalement, aucune manipulation n'est nécessaire ici. En effet, cet élément a déjà été précisé lors de plusieurs étapes de la réconciliation dans la partie précédente (parties 2.3 et 2.4).

2. Les *labels*

Dans ce fichier d'équivalence, le preflabel de l'AAT doit aussi apparaître. Cet élément n'apparaît pour l'instant pas dans le jeu de données. Il faut donc procéder à un moissonnage d'URL comme cela avait été fait pour les PACTOLS plus haut (partie 1.2.1). L'expression GREL suivante permet de créer l'URL à moissonner en récupérant l'identifiant unique contenu dans l'URI de l'AAT.

Expression GREL :

```
"http://vocabsservices.getty.edu/AATService.asmx/AATGetSubjectTerms?subjectID="
+ value.find(/\d+$/)[0]
```

Comme précédemment, cette opération permet de récupérer l'ensemble des éléments et des métadonnées de la page (figure 3.1). Une nouvelle colonne **targetLabel** est créée d'après les données moissonnées, desquelles le *preflabel* est sélectionné à l'aide d'une expression GREL (figure 3.2). Les données moissonnées peuvent alors être supprimées.

Expression GREL :

```
value.parseXml().select('Vocabulary')[0].select('Subject')[0].select('Terms')
[0].select('Preferred_Term')[0].select('Term_Text')[0].ownText()
```

AAT


```xml
<?xml version="1.0" encoding="utf-8"?>
<Vocabulary xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="http://vocabsservices.getty.edu/Schemas/AAT/AATGetSubjectTerms.xsd">
 <Subject Subject_ID="300014662">
  <Terms>
   <Preferred_Term>
    <Term_Text>pipes (conduits)</Term_Text>
    <Term_ID>1000014662</Term_ID>
    <Term_Languages>
     <Term_Language>
      <Language>English</Language>
      <Preferred>Preferred</Preferred>
      <Qualifier>conduits</Qualifier>
      <Term_Type>Descriptor</Term_Type>
      <Part_of_Speech>Plural Noun</Part_of_Speech>
      <Lang_Stat>Undetermined</Lang_Stat>
     </Term_Language>
    </Term_Languages>
    <Term_Contributors>
     <Term_Contributor>
      <Contributor_id>10000000/VP</Contributor_id>
      <Preferred>Preferred</Preferred>
     </Term_Contributor>
     <Term_Contributor>
      <Contributor_id>10000088/GCI</Contributor_id>
      <Preferred>Preferred</Preferred>
     </Term_Contributor>
    </Term_Contributors>
    <Term_Sources>
     <Term_Source>
      <Source>
       <Source_ID>2000008771/Roberts, Construction Industry Thesaurus (1976)</Source_ID>
      </Source>
      <Page>B tubes, no piping; pipes</Page>
      <Preferred>Preferred</Preferred>
     </Term_Source>
     <Term_Source>
      <Source>
       <Source_ID>2000035571/ROOT Thesaurus (1981)</Source_ID>
      </Source>
      <Page>pipes</Page>
      <Preferred>Preferred</Preferred>
     </Term_Source>
     <Term_Source>
      <Source>
       <Source_ID>2000051089/AATA database (2002-) </Source_ID>
      </Source>
      <Page>127398 checked 26 January 2012</Page>
      <Preferred>Preferred</Preferred>
     </Term_Source>
     <Term_Source>
      <Source>
       <Source_ID>2000094828/Getty Vocabulary Program rules</Source_ID>
      </Source>
      <Page>
      </Page>
      <Preferred>Alternate Preferred</Preferred>
     </Term_Source>
    </Term_Sources>
   </Preferred_Term>
   <Non-Preferred_Term>
    <Term_Text>pipe (conduit)</Term_Text>
    <Term_ID>1000285982</Term_ID>
    <Term_Languages>
     <Term_Language>
      <Language>English</Language>
      <Preferred>Non Preferred</Preferred>
      <Qualifier>conduit</Qualifier>
      <Term_Type>Alternate Descriptor</Term_Type>
      <Part_of_Speech>Singular Noun</Part_of_Speech>
      <Lang_Stat>Undetermined</Lang_Stat>
     </Term_Language>
    </Term_Languages>
    <Term_Contributors>
     <Term_Contributor>
      <Contributor_id>10000000/VP</Contributor_id>
      <Preferred>Non Preferred</Preferred>
     </Term_Contributor>
     <Term_Contributor>
      <Contributor_id>10000159/CHIN / RCIP</Contributor_id>
      <Preferred>Non Preferred</Preferred>
     </Term_Contributor>
     <Term_Contributor>
      <Contributor_id>10000299/AASLH</Contributor_id>
      <Preferred>Non Preferred</Preferred>
     </Term_Contributor>
    </Term_Contributors>
    <Term_Sources>
     <Term_Source>
      <Source>
       <Source_ID>2000094828/Getty Vocabulary Program rules</Source_ID>
```


Figure 3.1: Illustration des informations récupérées par moissonnage

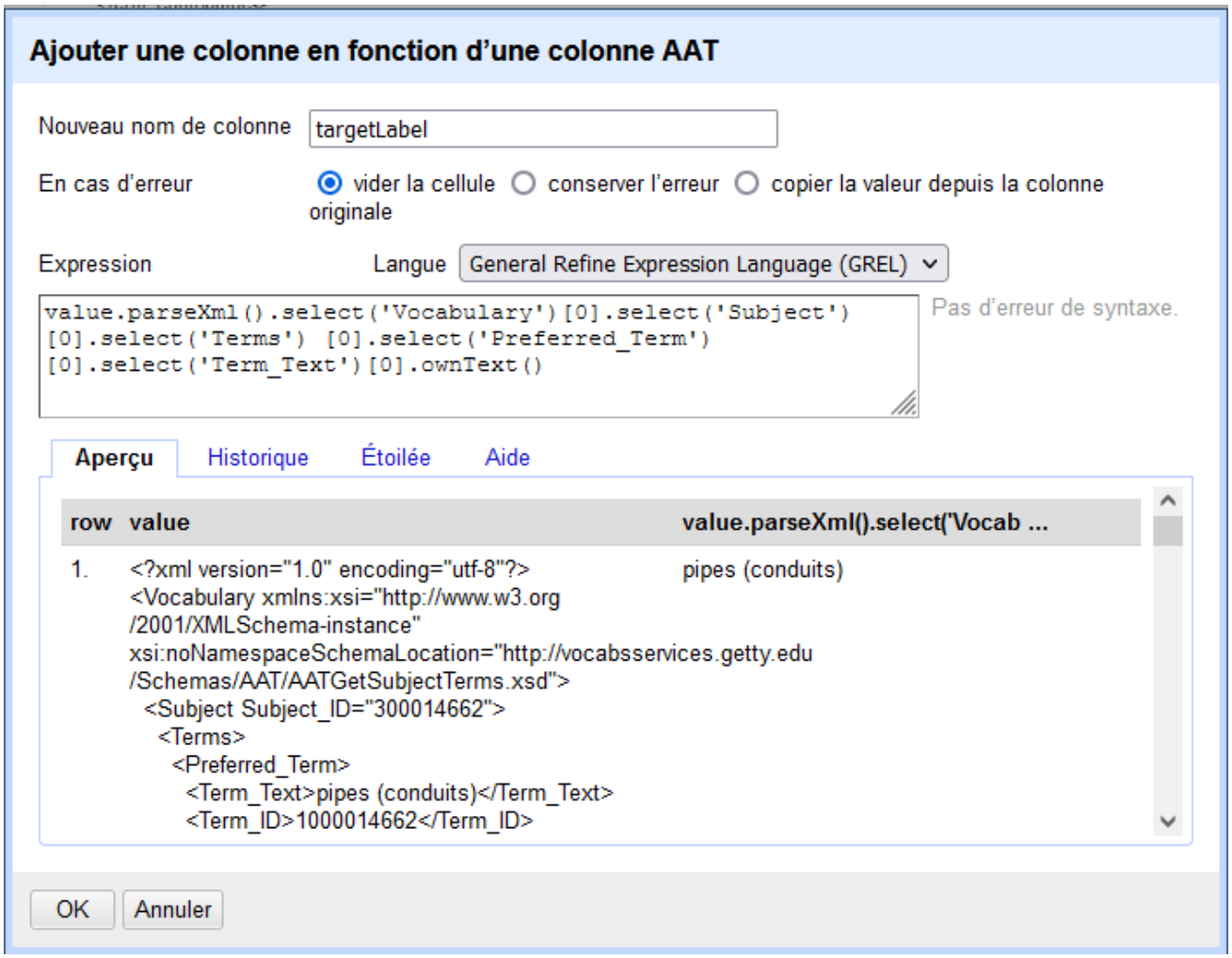


Figure 3.2: Récupération du preflabel

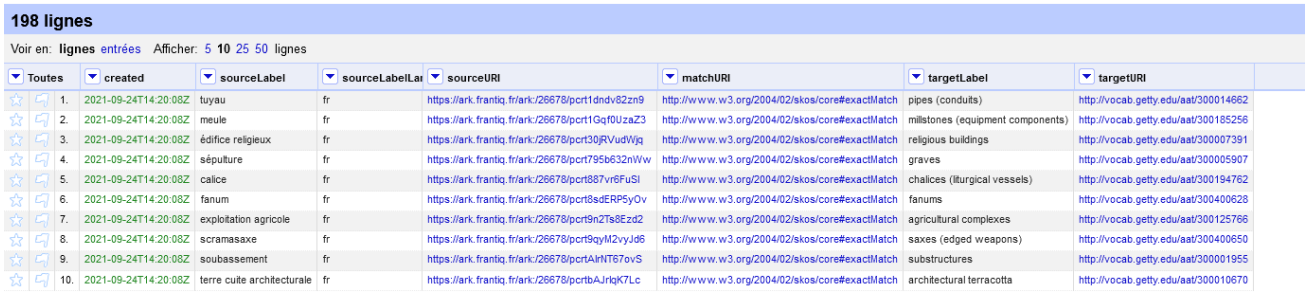
3. La date et la langue

Il faut encore rajouter une colonne indiquant la langue de chaque terme source. Dans le cas développé ici, il s'agit du français pour l'ensemble des termes. Une nouvelle colonne **sourceLabelLanguage** est alors créée en dupliquant n'importe laquelle des colonnes existante et, dans le cadre de transformation, la mention `value` est remplacée par `"fr"`.

Enfin, il faut encore préciser la date de création de chaque ligne du tableur au sein d'une nouvelle colonne dupliquée **created**. Openrefine possède une fonction pour indiquer automatiquement la date exacte sur l'ensemble des cellules. Le document est désormais prêt pour le dépôt dans ARIADNEplus (figure 3.3).

Expression GREL :

`now()`



198 lignes

Voir en: lignes entrées Afficher: 5 10 25 50 lignes

Toutes	created	sourceLabel	sourceLabelLa	sourceURI	matchURI	targetLabel	targetURI
1.	2021-09-24T14:20:08Z	tuyau	fr	https://ark.frantiq.fr/ark:/26678/pcrt1dndv82zn9	http://www.w3.org/2004/02/skos/core#exactMatch	pipes (conduits)	http://vocab.getty.edu/aat/300014662
2.	2021-09-24T14:20:08Z	meule	fr	https://ark.frantiq.fr/ark:/26678/pcrt1Gqf0UzaZ3	http://www.w3.org/2004/02/skos/core#exactMatch	millstones (equipment components)	http://vocab.getty.edu/aat/300185256
3.	2021-09-24T14:20:08Z	édifice religieux	fr	https://ark.frantiq.fr/ark:/26678/pcrt30jRVudWjq	http://www.w3.org/2004/02/skos/core#exactMatch	religious buildings	http://vocab.getty.edu/aat/300007391
4.	2021-09-24T14:20:08Z	sépulture	fr	https://ark.frantiq.fr/ark:/26678/pcrt795b632nWw	http://www.w3.org/2004/02/skos/core#exactMatch	graves	http://vocab.getty.edu/aat/300005907
5.	2021-09-24T14:20:08Z	calice	fr	https://ark.frantiq.fr/ark:/26678/pcrt887vr6FuSl	http://www.w3.org/2004/02/skos/core#exactMatch	chalices (liturgical vessels)	http://vocab.getty.edu/aat/300194762
6.	2021-09-24T14:20:08Z	fanum	fr	https://ark.frantiq.fr/ark:/26678/pcrt8sdERP5yOv	http://www.w3.org/2004/02/skos/core#exactMatch	fanums	http://vocab.getty.edu/aat/300400628
7.	2021-09-24T14:20:08Z	exploitation agricole	fr	https://ark.frantiq.fr/ark:/26678/pcrt9n2Ts8Ezd2	http://www.w3.org/2004/02/skos/core#exactMatch	agricultural complexes	http://vocab.getty.edu/aat/300125766
8.	2021-09-24T14:20:08Z	scramasaxe	fr	https://ark.frantiq.fr/ark:/26678/pcrt9qyM2vyJd6	http://www.w3.org/2004/02/skos/core#exactMatch	saxes (edged weapons)	http://vocab.getty.edu/aat/300400650
9.	2021-09-24T14:20:08Z	soubassement	fr	https://ark.frantiq.fr/ark:/26678/pcrtAlrNT67ovS	http://www.w3.org/2004/02/skos/core#exactMatch	substructures	http://vocab.getty.edu/aat/300001955
10.	2021-09-24T14:20:08Z	terre cuite architecturale	fr	https://ark.frantiq.fr/ark:/26678/pcrtbAJrlqK7Lc	http://www.w3.org/2004/02/skos/core#exactMatch	architectural terracotta	http://vocab.getty.edu/aat/300010670

Figure 3.3: Document final pour l'import dans ARIADNEplus

IV. Ressources et références

- Blaney, J. (2017). "Introduction to the Principles of Linked Open Data". *Programming Historian*. https://programminghistorian.org/en/lessons/intro-to-linked-data
- Frantiq. *Pactols*. Consulté 8 octobre 2021, à l'adresse https://pactols.frantiq.fr/opentheso/
- Frantiq. *Pactols2*. Consulté 8 octobre 2021, à l'adresse https://pactols2.frantiq.fr/pactols2/index.xhtml#
- Hooland, S. van, Verborgh, R., & Wilde, M. D. (2013). "Nettoyer ses données avec OpenRefine". *Programming Historian*. https://programminghistorian.org/fr/lecons/nettoyer-ses-donnees-avec-openrefine
- *OpenRefine user manual | OpenRefine*. Consulté 8 octobre 2021, à l'adresse https://docs.openrefine.org//
- Roussed, M. Pactols?: "Webservices" [Billet]. *Opentheso*. Consulté 8 octobre 2021, à l'adresse https://opentheso.hypotheses.org/50
- Saby, M. *Tutoriel OpenRefine 3.4?: nettoyer, préparer et transformer des données - 06/11/2020*. Consulté 8 octobre 2021, à l'adresse https://msaby.gitlab.io/tutoriel-openrefine
- *Vocabularies Download Center (Getty Research Institute)*. Consulté 8 octobre 2021, à l'adresse https://www.getty.edu/research/tools/vocabularies/obtain/download.html
- Williamson, E. P. (2017). "Fetching and Parsing Data from the Web with OpenRefine". *Programming Historian*. https://programminghistorian.org/en/lessons/fetch-and-parse-data-with-openrefine

- 1 Pour voir comment installer et utiliser OpenRefine lire ce tutoriel
- 2 Pour faire l'alignement avec PACTOLS voir *Tutoriel alignement de vocabulaire : PACTOLS*
- 3 Dans le cadre d'un document sous forme de tableur, le logiciel OpenRefine fonctionne mieux avec le format csv.
- 4 L'acronyme GREL correspond à *General Refine Expression Language*, un langage d'expression développé d'après JavaScript pour OpenRefine. Pour plus d'informations voir le manuel d'OpenRefine
- 5 On note que la première partie de cette expression ne correspond pas aux informations que l'on peut trouver sur l'API d'Opentheso. L'URL utilisée ici mène à la version 2 des PACTOLS qui n'est pas encore ouverte et sur laquelle on trouve de nombreuses autres informations en plus d'une organisation des concepts différente. Cette URL sera donc à modifier lorsque cette v2 sera ouverte. Pour moissonner l'actuelle version des PACTOLS, utiliser l'URL : `"https://pactols.frantiq.fr/opentheso/api/"`
- 6 Pour plus d'informations et pour accéder à cette version voir ce site
- 7 L'ensemble de ce chapitre est construit à l'aide du tutoriel rédigé par le Getty (en anglais)
- 8 Voir manuel Openrefine (en anglais)
- 9Saby 2020
- 10 Pour plus d'informations sur les indications d'éléments pertinents avec l'AAT voir le document établi par le Getty à la page 13.
- 11 Puisqu'elle nécessite de bien connaître les données sources et leur signification, mais aussi le vocabulaire métier, pour pouvoir les comparer aux définitions du référentiel cible, il est préférable que cette opération soit réalisée par quelqu'un qui maîtrise ce vocabulaire ou les données analysées.
- 12 Attention, OpenRefine est sensible à la casse. Il faut donc être vigilant dans l'écriture des noms de colonnes aux éventuelles majuscules et espaces.
- 13https://www.w3.org/2009/08/skos-reference/skos.html
- 14 L'emploi d'expressions régulières avec le logiciel Excel n'est pas des plus facile et il est plus pratique de privilégier Calc de la suite LibreOffice. Il est aussi possible d'effectuer cette modification avant l'export du csv, dans OpenRefine, en transformant les cellules ou en utilisant la fonction "**Remplacer**".
- 15 Quel que soit le logiciel employé, ne pas oublier de développer les options et de cocher l'option "Expression régulières".
- 16https://www.w3.org/TR/skos-reference/