Acronyme	**Campus AAR**
Titre du projet	**Campus « Archives Audiovisuelles de la Recherche ».** *Le démonstrateur d'un environnement numérique pour la production, description/indexation et publication d'archives audiovisuelles.* *Domaine d'application : les humanités numériques.*
Proposal title	**ARA Campus.** The Audiovisual Research Archive Campus – a demonstrator for the production, description and publishing of digital audiovisual archives. Domain of expertise: Digital humanities

Tâche 2.1 : Terminologies et modèle de données v1

Partenaire(s) concerné(s)	FMSH, INA
Référence convention/décision	**ANR-13-CORD-0016-01**
Rédacteur(s) du rapport	Steffen Lalande, Peter Stockinger
Téléphone de contact	0149832136
Adresse électronique (de contact)	slalande@ina.fr
Date :	30/03/2015

ID document	Campus AAR – Rapport
Distribution	☒ Public ☐ Restreint ☐ Interne
Description	Document de présentation du modèle de données et des ontologies

Résumé général du rapport

Ce rapport décrit succinctement le modèle de données adopté dans le cadre du projet Campus-AAR, l'articulation entre ressources terminologiques ainsi que la reprise d'antériorité des ressources terminologiques du modèle ASA-SHS.

Ce modèle a été élaboré avec le souci premier d'englober les différents langages de description utilisés par les différents partenaires dans leurs systèmes actuels, cela sans perte d'expressivité, mais avec un objectif d'unification de certains concepts communs, d'amélioration des capacités d'extension et des capacités de requêtage.

Ce modèle prend la forme d'un ensemble de fichiers OWL liés par la notion d'import inhérente à OWL. L'intégralité du modèle peut être visualisée par l'intermédiaire du logiciel *Protégé* de l'université de Stanford. Les descriptions reprises sont, elles, exprimées dans le langage NQUAD, un format alternatif à RDF/XML permettant de représenter la notion de graphe.

Une description plus détaillée du modèle de données de Campus-AAR, et notamment de l'utilisation des possibilités logiques du langage OWL pour élaborer de manière déclarative des modèles de description, sera présentée dans le rapport de recherche.

Table des matières

1. Les objets statiques du modèle de description

L'objectif du projet en termes de description de contenu média est ambitieux. Il consiste à définir un *langage de description* s'appuyant sur les langages du Web sémantique et dont le cahier des charges répond aux caractéristiques suivantes :

1. Offrir des capacités de représentation compatibles avec les caractéristiques des systèmes existants chez les partenaires du projet: Le modèle ASA-SHS de l'ESCoM-AAR de la FMSH constitue en ce sens un modèle propriétaire dont la structure et les mécanismes de mise en œuvre sont complexes.

2. Offrir la possibilité de décrire un média selon différents points de vue : ce qui signifie l'utilisation de vocabulaires différents et des structures d'annotation dont la complexité peut être variable (du simple « tag » à une analyse sémiotique très structurée).

3. Offrir la possibilité d'étendre les ressources terminologiques et de définir de manière déclarative de nouveaux modèles de description.

4. Permettre une articulation optimale entre base de description et base de connaissances : Distinguer les connaissances factuelles réutilisables et les informations qui n'ont un sens que dans le contexte du média en exploitant la notion d'individu et celle de graphe d'assertions.

5. Assurer un accès rapide et générique à l'ensemble des informations, indépendamment de la complexité de la description : le système doit pouvoir répondre à des requêtes articulant différentes axes de description, notamment pour la partie republication de contenu.

Le modèle est construit sur les langages RDFS/OWL du W3C et la notion de graphe issue des triple-store et du langage SPARQL. La structure de base du modèle est définie au sein de l'ontologie racine « core.owl ». Ce fichier définit l'ensemble des objets statiques du modèle et les relations qui les unissent. Ces objets sont en nombre restreint afin de permettre une applicabilité du modèle à des usages très divers.

Media

La classe de média représente les individus média abstrait. Chaque média peut posséder plusieurs formats différents (Instances) ainsi que des adresses physiques (url) auxquelles ces instances sont accessibles via différents protocoles (Streaming, Download, FTP, …).

Les métadonnées de base (d'ingestion dans Campus–AAR) sont définies sur une strate particulière et un segment particulier par souci d'homogénéité du modèle. Le format OAI a été retenu comme ontologie de description de ces métadonnées de base.

Strate

Une strate constitue un axe de description d'un média selon une ontologie de domaine particulière. Dans les cas de vidéos, la classe « Strate temporelle » est utilisée ; elle permet de référencer un ensemble de segments temporels, ces derniers pouvant potentiellement se recouvrer. La classe « Strate média » est un type de strate particulière dont chaque média possède une instance et dont l'ontologie de domaine est OAI.

Segment

Un segment temporel constitue l'élément central de la description. Il définit un contexte particulier pour les assertions (annotations) qui sont produites par l'annotateur. Ces annotations prennent la forme générale d'un graphe de connaissances constitué d'individus appartenant aux classes de l'ontologie de domaine utilisé par la strate possédant le segment.

Analyse

Une analyse est un regroupement de strates. Ce regroupement peut être contraint à différents types de strates en définissant des sous-classes de la classe « Analyse » dans une extension du modèle de base. Il est ainsi possible de définir différents schémas d'analyse.

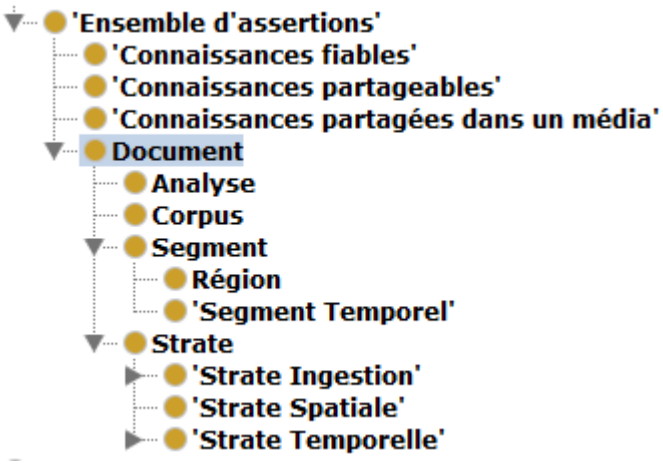
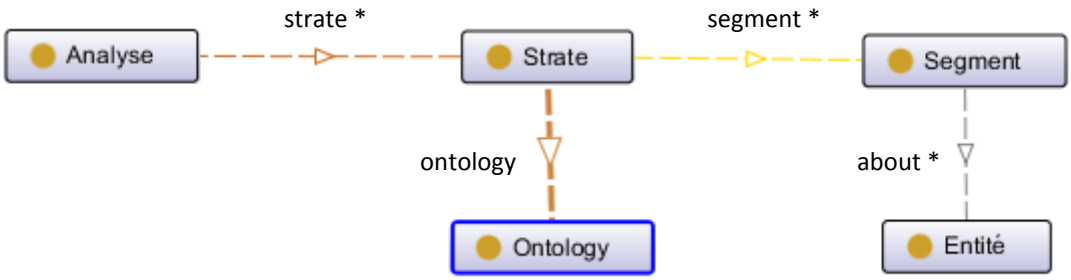


Figure 1 hiérarchie des classes d'objets de description



Contexte et graphe

La notion de contexte joue un rôle central dans le modèle de description proposé. Cette notion est réifiée dans le modèle par la classe « Ensemble d'assertions » et opérationnalisée dans le système par l'utilisation de graphes au sens SPARQL. Ces graphes contiennent un ensemble d'assertions (triplets Sujet, Prédicat, Objet) dont la validité n'est affirmée que dans le contexte de description. Chaque segment est ainsi à la fois un objet du modèle et un graphe dans lequel sont insérés l'ensemble des triplets composant la structure de description issue de ce segment.

L'appartenance de tout triplet à un « graphe de contexte » particulier permet la création de graphes connexes maximisant la réutilisation des objets communs tout en garantissant une séparation claire des assertions faites sur ces objets.

2. Les objets des domaines de description

L'ontologie cœur « core.owl » héberge également la définition des grandes classes d'« entités nommées » (Personnes, Organisations, Lieux, Monuments, etc.) dans le but de rendre leurs instances partageables entre les domaines.

Un ensemble de propriétés consensuelles est en cours de définition sur chacune de ses classes ; les mécanismes de dérivation ontologique OWL permettront également de définir des propriétés spécifiques à un domaine précis sur ces mêmes classes. Les individus de ces classes auront ainsi des propriétés visibles en permanence et d'autres visibles et éditables en fonction du domaine d'usage.

Le modèle est agnostique quant aux ontologies de domaines construites dans les extensions du « core ». A noter que ces domaines peuvent mixer des hiérarchies RDFS/OWL et des ressources SKOS (graphes d'instances reliés par des relations de généralisation/spécialisation).

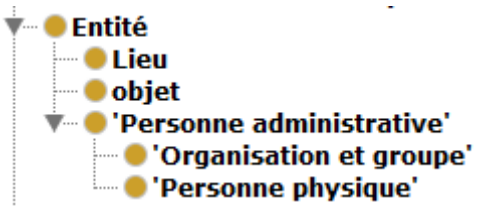


Figure 2 Hiérarchie courante des classes d'entités partagées

3. Extensibilité du modèle et structuration des ressources

La construction d'une ressource de description d'un domaine particulier se caractérise par la création d'une ou plusieurs ontologies « important » l'ontologie cœur. L'exemple ci-après de reprise d'antériorité des différents domaines ASA-SHS montre l'utilité d'une modularité dans la définition des extensions : des ontologies intermédiaires permettent de définir des concepts partagés tels que le modèle d'annotation par sujet adopté par l'ensemble des domaines ASA-SHS.

Les classes définies dans les extensions peuvent être définies comme des extensions de la hiérarchie des entités partagées du cœur. De plus, le multi-héritage de RDFS/OWL permet de garder pour chaque domaine sa propre hiérarchie tout en se greffant sur l'arbre des entités partagées.

Les différents mécanismes de restrictions d'OWL portant sur les valeurs de propriétés permettent de définir sous formes logiques des graphes de description valides pour un domaine particulier et de construire ainsi des modèles permettant à la fois de guider les utilisateurs, d'alléger la sélection de ressources de description et de spécifier une politique éditoriale.

4. Reprise d'antériorité ASA-SHS de l'ESCoM-AAR

La reprise d'antériorité des contenus de trois domaines AGORA (patrimoine audiovisuel SHS), ARC (diversité culturelle) et AHM (histoire des mathématiques) des archives ASA-SHS constitue une véritable évaluation des capacités de représentation du modèle construit au vue de la complexité des ontologies et des modèles de description définis dans le langage propriétaire de l'ESCoM-AAR.

L'enjeu est à la fois de pouvoir disposer d'une version OWL de ces différentes ontologies pour la production de nouvelles analyses via l'application d'analyse de Campus AAR mais aussi de pouvoir traduire dans le nouveau format l'ensemble des analyses existantes afin de disposer rapidement d'un corpus conséquent et exploitable

4.1) Reprise du modèle de description et des ontologies de domaines

4.1.1) Présentation synthétique des ressources métalinguistiques ASA-SHS

La figure 3 montre les différentes parties qui composent le système des ressources métalinguistiques ASA. Nous y distinguons d'abord la partie *Bibliothèque de modèles de description*. Une bibliothèque de modèles de description est composée au minimum d'une, mais en règle générale de plusieurs *modèles de description* qui représentent la *vision* de l'*univers du discours d'une archive audiovisuelle donnée*. La *vision* de l'univers du discours d'une archive peut évoluer en fonction des intérêts, des objectifs ou tout simplement en fonction de son objet audiovisuel. Ainsi une bibliothèque de modèles peut subir des changements. Mais les

conséquences éventuelles d'un tel changement sur les analyses déjà réalisées doivent être évaluées avec précaution.

Une bibliothèque de modèles de description est composée de modèles spécialisés dans l'analyse d'une partie ou d'un aspect spécifique de l'objet textuel tel qu'il est appréhendé dans une perspective *sémiotique*. Ainsi, distinguons-nous, comme déjà expliqué dans la première partie de ce livre, entre des modèles qui nous servent à réaliser :

- la description d'une analyse elle-même (i.e. la tâche de la *méta-description*) ;
- la *description paratextuelle* (i.e. de l'identité formelle du texte audiovisuel dans une perspective tout à fait comparable à celle que nous dresse le standard Dublin Core) ;

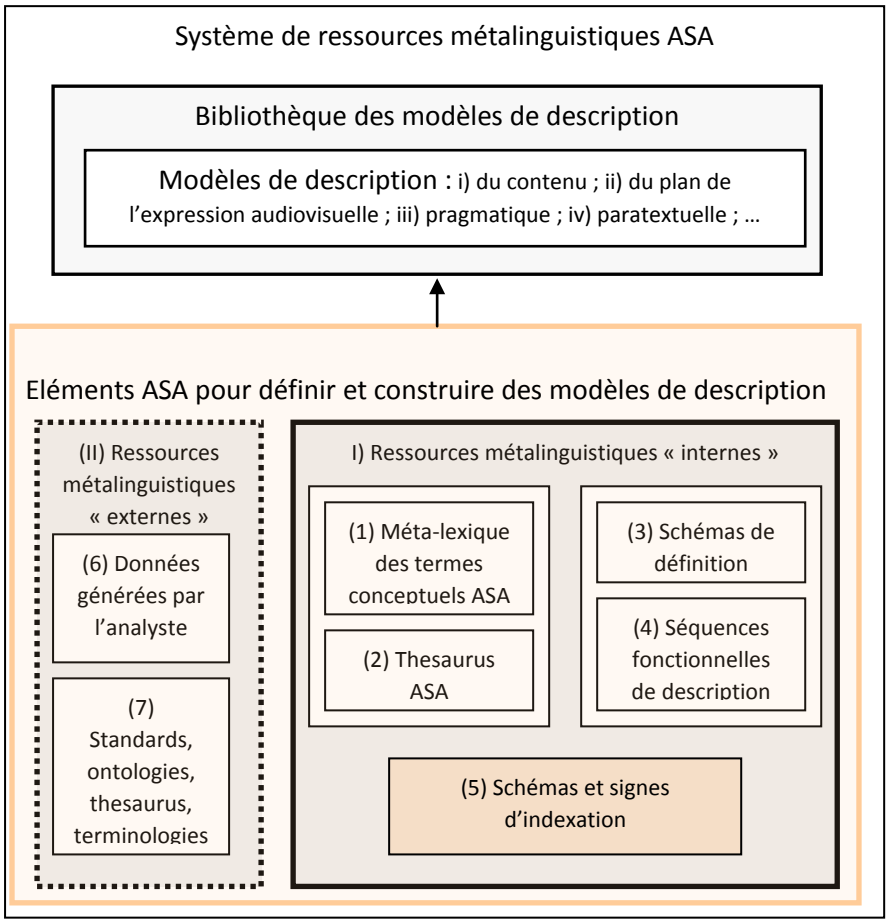


Figure 3 : **Vue d'ensemble des ressources métalinguistiques ASA**

- la *description audiovisuelle* stricto sensu (i.e. des plans visuels, acoustiques, et/ou audiovisuels du texte audiovisuel) ;
- la *description thématique* (i.e. des sujets développés dans le texte audiovisuel) ;
- la *description pragmatique* (i.e. l'ouverture de l'identité, de la spécificité du texte audiovisuel à la culture et aux attentes d'un public ainsi qu'aux exigences propres au contexte dans lequel il doit être utilisé),
- l'adaptation-traduction à proprement parler que nous rangeons également sous l'étiquette de *description pragmatique* et qui vise de surmonter l'obstacle de la langue dans la réception et l'appropriation de l'objet audiovisuel.

Ces différents types de modèles de description organisent l'interface de travail du logiciel d'analyse de médias audiovisuels développé dans le cadre du projet Campus AAR.

La moitié inférieure de la figure 3 montre les différentes parties du système des ressources métalinguistiques ASA qui permettent la construction d'un modèle de description spécifique ou d'une bibliothèque de tels modèles. Nous y distinguons deux ensembles complémentaires :

1. l'ensemble des ressources métalinguistiques qui font partie du système ASA,
2. et l'ensemble des ressources qui lui sont externes mais qui sont *mis en relation avec ce dernier*.

L'ensemble des *ressources métalinguistiques propres au système ASA* connaît trois classes de ressources plus spécifiques et fonctionnellement différentes :

1. une classe de ressources lexicales constituées d'une part d'un méta-lexique hiérarchique de *termes conceptuels génériques* et d'autre part d'un vocabulaire contrôlé qui est le *thesaurus du système ASA* ;
2. une classe de ressources *structurales* ou *configurationnelles* qui sélectionnent et positionnent les termes génériques et/ou du thesaurus les uns par rapport aux autres selon les spécificités d'un domaine d'expertise donné et selon les besoins de l'analyse ;
3. une classe de ressources – appelée *schémas d'indexation* - qui permettent de réaliser une description, de l'exécuter

La figure 3 identifie également un ensemble de ressources métalinguistiques qui sont externes au système ASA. On y trouve d'abord les *données générées par les analystes eux-mêmes* qui se servent des modèles de description ASA pour traiter (décrire, indexer, annoter, …) leurs corpus audiovisuels. Ces données générées par les analystes constituent une base d'expressions sémiolinguistiques (verbales mais aussi iconiques, acoustiques, etc.) qui peut servir de référentiel à des nouvelles analyses du même texte audiovisuel.

Une deuxième catégorie de ressources métalinguistiques externes au système ASA est composée de *standards*, *ontologies*, *thesaurus* et autres *terminologies* avec lesquels ont été ou peuvent être créés des correspondances, des *ponts*. Ces correspondances ou ponts servent à rendre – dans la mesure du possible – *interopérables* les résultats des analyses concrets réalisés à l'aide du Studio ASA avec ceux réalisés en référence à ou à l'aide d'autres ressources métalinguistiques (ontologie, thesaurus, etc.).

Le système ASA sera présenté et discuté plus en détail dans le rapport de recherche qui constitue le livrable R1 *(à fournir pour mi-mai 2015)*.

4.1.2) Simplification des concepts et compatibilité avec la philosophie RDFS/OWL :

Le modèle ASA-SHS décrit ci-dessus combine une grande diversité de typologies de ressources pour réaliser la description des contenus audiovisuels. Cette typologie propriétaire et le processus de description en découlant demandent à être adaptés afin de répondre à la philosophie de RDFS/OWL.

Plusieurs axes de modification ont été ainsi identifiés et traduits sous formes de mécanismes de transformation implémentés dans un outil de conversion ASA-SHS -> CAMPUS-AAR (OWL) utilisant le langage SWI-Prolog et sa bibliothèque SemWeb (Web sémantique).

Des parties fixes

Les principaux axes de transformation concernent :

1. **La définition de propriétés** permettant de lier sémantiquement les entités utilisées en annotation. La structure d'annotation thématique évolue ainsi d'un ensemble d'entités apparaissant aux feuilles d'un arbre de description (formulaire hiérarchique) à de véritables graphes d'annotation.
2. **La réduction du nombre de hiérarchies** covariantes : La hiérarchie des objets d'analyses, celles des listes d'expressions, du Micro-thésaurus et des facettes forment autant de hiérarchies de structures voire de libellés quasi-identiques, séparant classes (termes conceptuels) et individus dans des hiérarchies séparées. L'objectif est de regrouper sous une seule hiérarchie « objets d'analyses » (figure 4) l'ensemble des objets du domaine et d'y accrocher les individus ASA-SHS (Termes définis) sous la forme d'individus OWL ainsi que les facettes terminales (regroupements contextuels d'individus) du modèle ASA-SHS sous la forme de classes.
3. **La réduction du nombre de types de concepts ASA-SHS liés à la définition de modèles** et à la réutilisabilité : Le modèle ASA-SHS utilise un nombre important de notions permettant de définir des modèles et d'optimiser la réutilisabilité des ressources à de nombreuses étapes. Les notions ASA-SHS de Sujet, Séquence, Topique, Schéma, Annotation, Module permettent ainsi de définir un contexte très précis de l'annotation qui est faite par un utilisateur mais elles ne se prêtent pas facilement à une interrogation facile et performante du modèle. L'objectif poursuivi était de transformer certains éléments en propriétés (figure 5), de ventiler dans la partie axiome de OWL certaines parties de ce modèle relatif au guidage de l'annotation (figure 6) et finalement de centrer la description sur la notion de sujet (Figure 7), d'objet d'analyse et d'annotations (Figure 8) sur ces objets d'analyse, en supprimant ainsi les notions de séquences, de schémas et de modules.

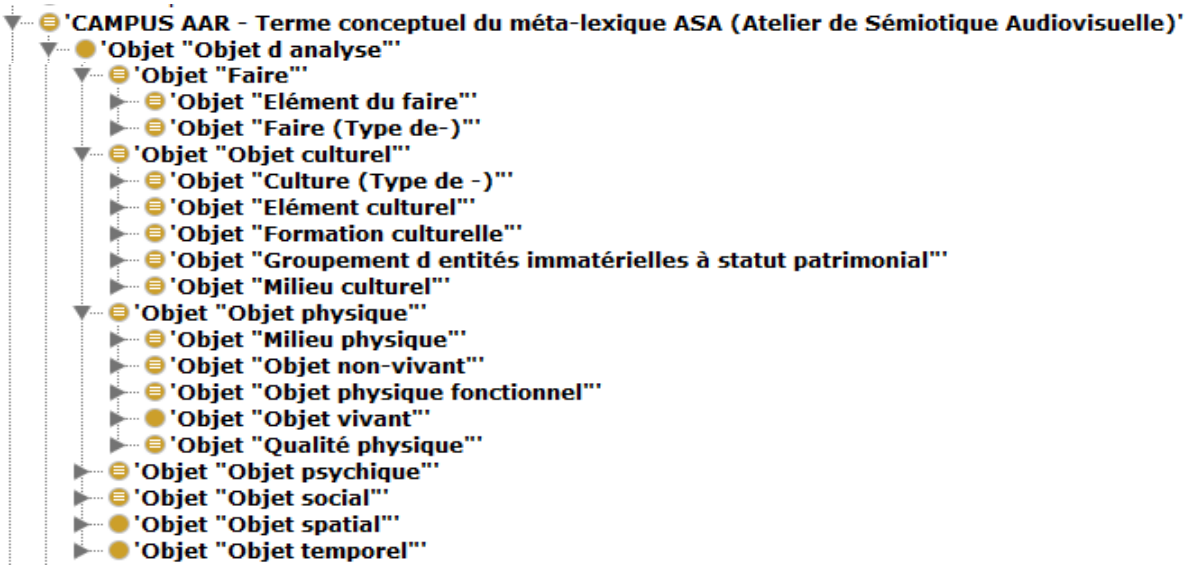


Figure 4 : La hiérarchie des objets d'analyse

Le modèle de données résultant permet ainsi de définir des graphes de description simplifiés tout en respectant, grâce à l'utilisation de contraintes OWL, les contraintes exprimées dans les modèles ASA-SHS sources.

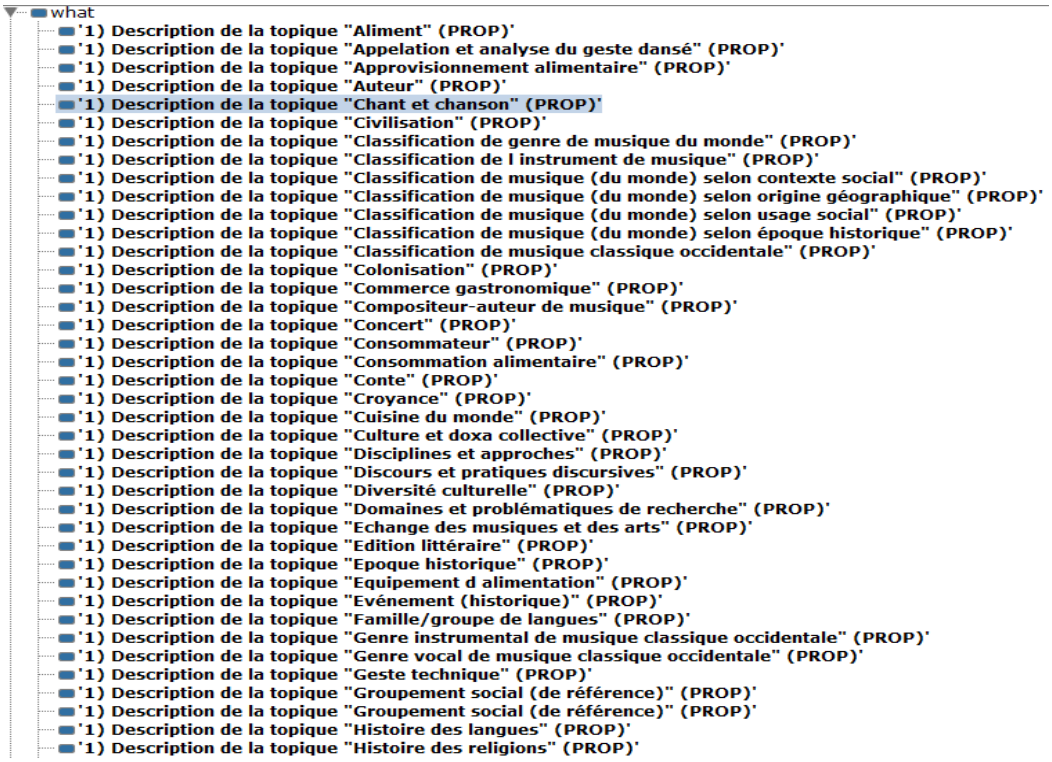


Figure 5 : Exemples de propriétés « Topiques » générés automatiquement à partir des ressources ASA-SHS

La transformation de ces ressources a été guidée par des annotations faites sur le modèle source ASA-SHS. Ces annotations ont notamment permis de spécifier les éléments à transformer en propriétés ainsi que les connexions entre facettes et « objets d'analyse ».

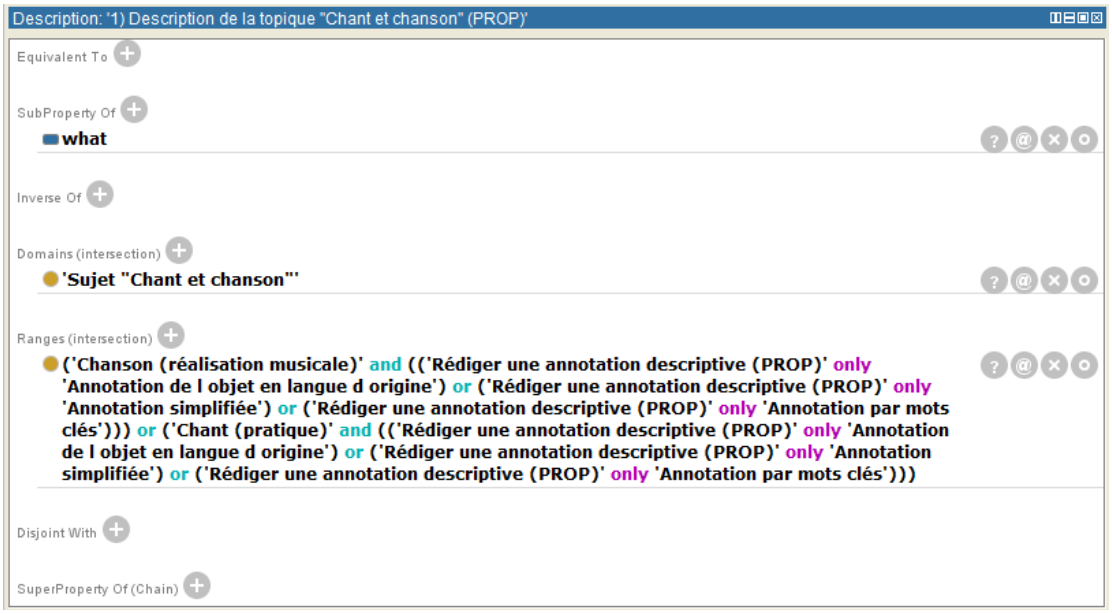


Figure 6 : Exemple de propriété « Topique » et de transformation automatique des schémas ASA-SHS associés en restrictions OWL sur les types de valeurs admises sur cette propriété

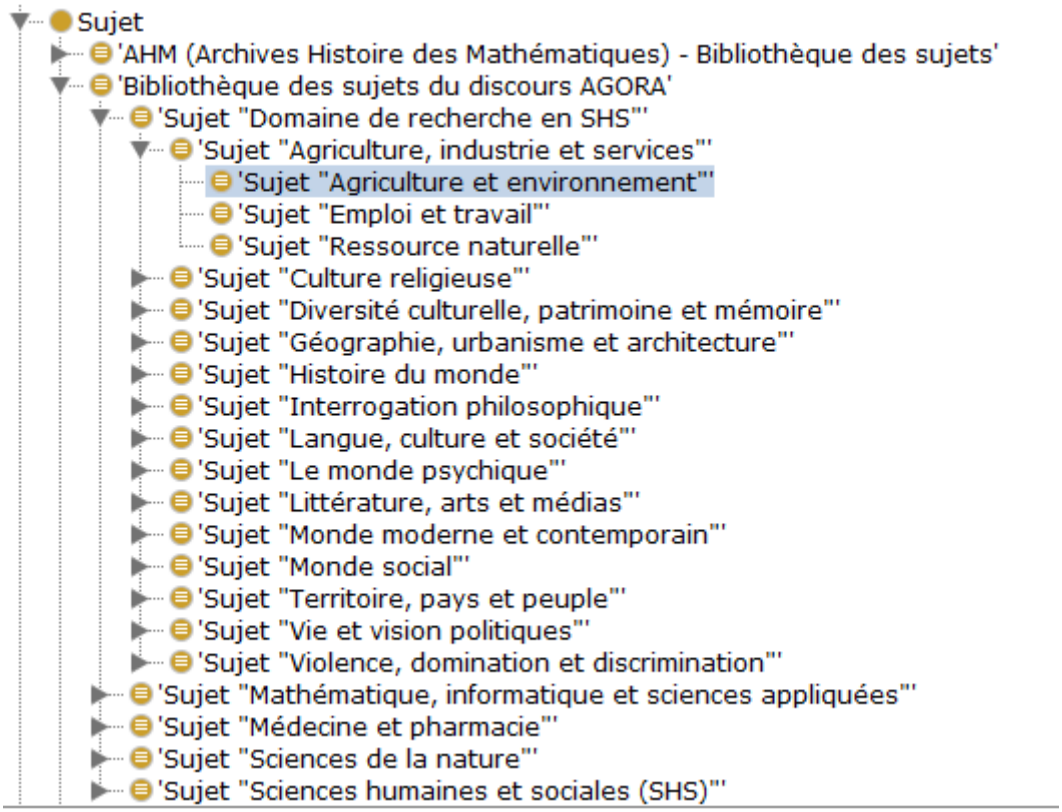


Figure 7 : Extrait de la hiérarchie OWL des Sujets

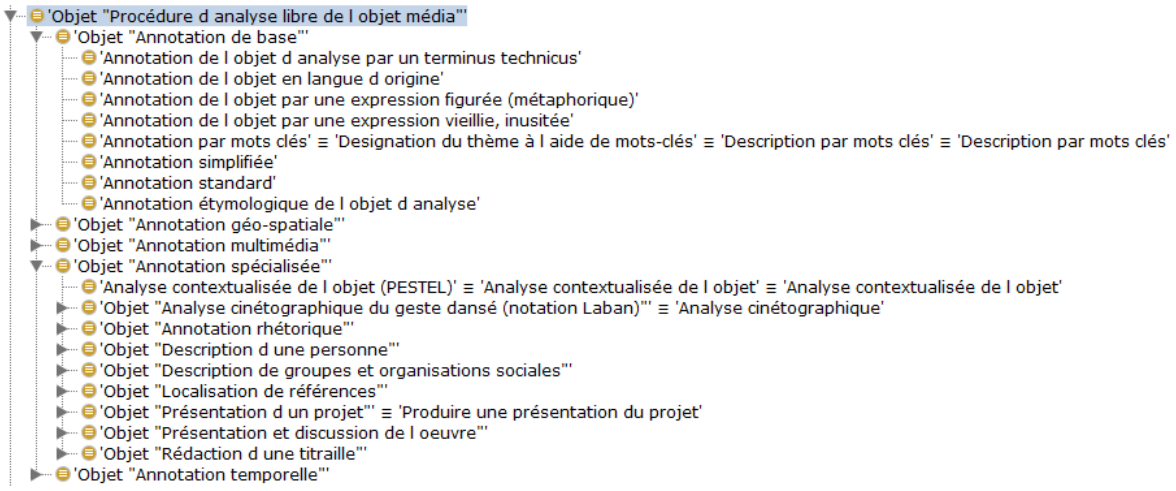


Figure 8 : La hiérarchie OWL des Annotations

4.1.3) Ressources OWL de Campus AAR

L'outil de conversion des ontologies ASA-SHS dans le format OWL de Campus-AAR génère un ensemble de fichiers OWL dont la liste exhaustive est donnée en figure 9.

Ce processus de transformations s'exécutant en plusieurs étapes, il génère des structures temporaires destinées à disparaitre de l'ontologie finale. Ces structures sont conservées afin de guider le processus de reprise d'antériorité des descriptions ASA-SHS elles-mêmes.

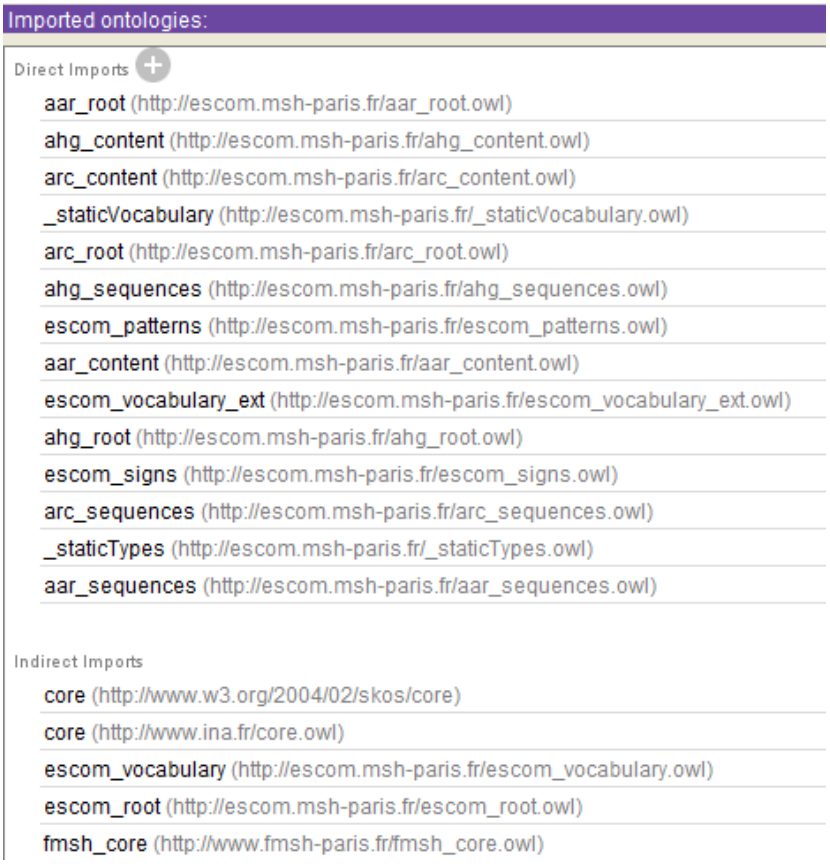


Figure 9 : Liste des fichiers OWL constituant le modèle cœur de Campus-AAR et la reprise du modèle ASA-SHS pour trois domaines AHG, AGORA et ARC

4.2) Reprise des descriptions

Un outil de reprise d'antériorité, également développé en Prolog et utilisant l'ontologie OWL générée permet de convertir les analyses des trois domaines AHG, AGORA et ARC de leur format propriétaire ASA-SHS (XML) en analyses « Campus-AAR » s'exprimant en RDF/NQUAD.
Cette transformation est exhaustive : elle concerne l'intégralité des plans de description du modèle d'origine et l'ensemble des métadonnées sur chaque plan.

La figure 10 ci-dessous montre un zoom d'un graphe de description centré sur un des sujets décrit dans un segment particulier de vidéo.

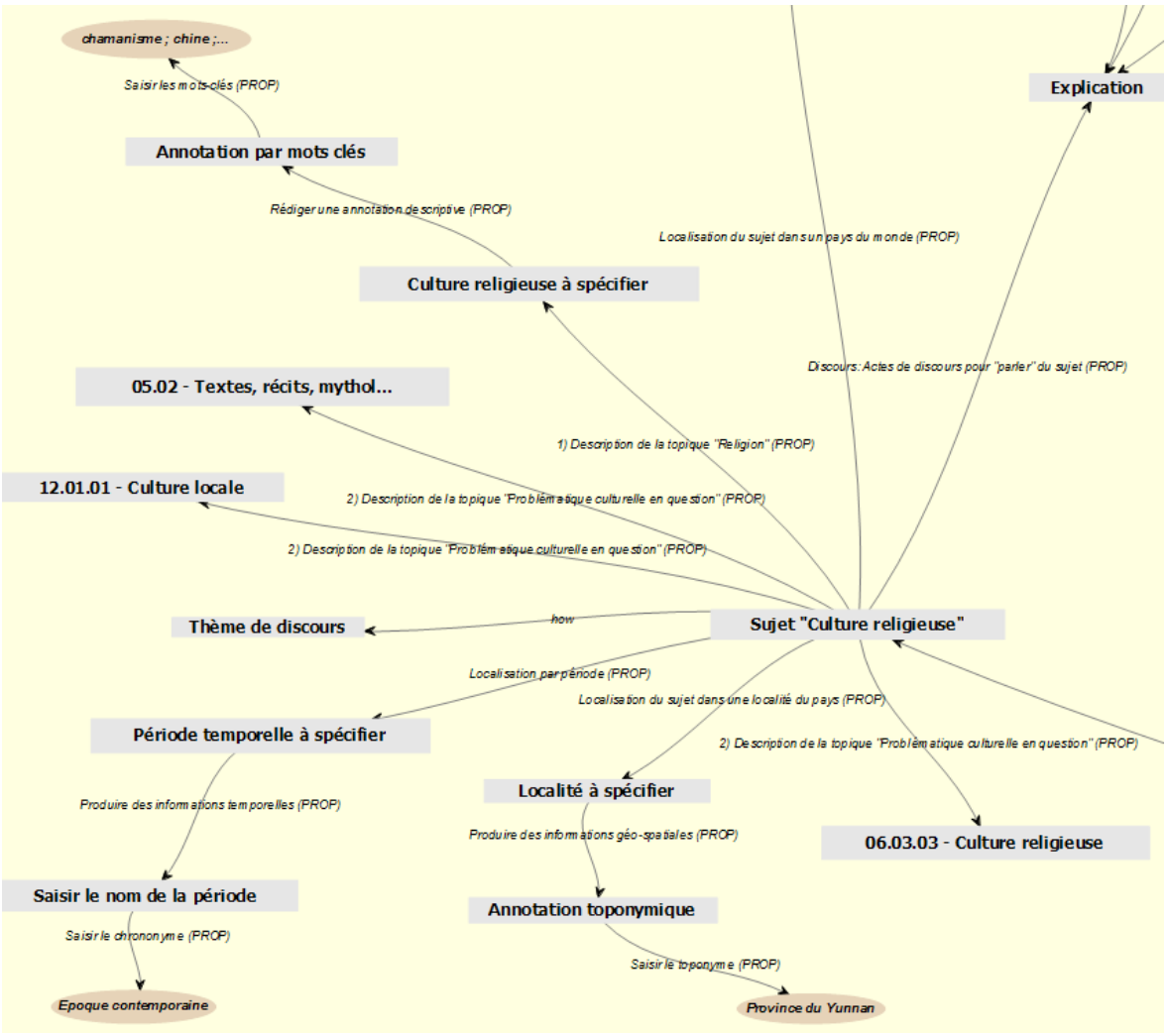


Figure 10 : Exemple d'extrait de graphes de description générée par l'outil de reprise d'antériorité

La reprise des plans de descriptions thématiques, visuels et sonores du modèle ASA-SHS donne lieu à la génération automatique de strates de description spécifiques ainsi que de segments sur lesquels viennent s'ancrer les différentes parties pertinentes de la description.

Ce découplage temporel entre les différents plans de description est l'un des apports de Campus-AAR par rapport au modèle ASA-SHS (figure 11).

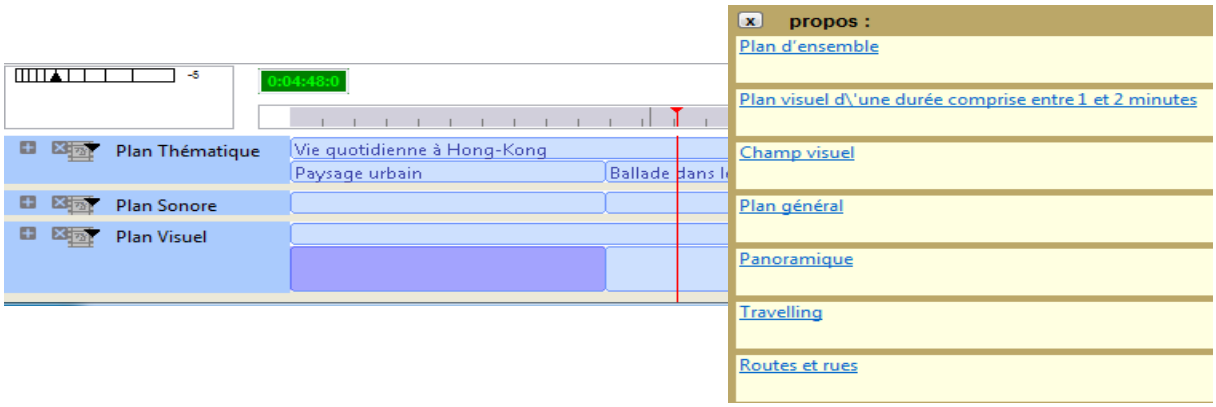


Figure 11 : Exemple de segment plan visuel généré et les métadonnées qui y sont associées

Volumétrie :

352 analyses portant sur les trois domaines AGORA, AHG et ARC ont été converties à ce jour par l'outil.

L'ensemble des ressources ontologiques et des descriptions converties correspondent à un peu moins de 500.000 triplets RDF.