La plateforme Okapi pour l'analyse et la publication de corpus de données audiovisuelles
Le cas du portail Inalco–MediaScape

Steffen LALANDE(1), Abdelkrim BELOUED(2), Odile FARGE(3)

(1) (2) INA Recherche
(3) Inalco - Plidam

Résumé : Nous présentons dans cet article la plateforme web Okapi, un outil destiné aux communautés de chercheurs en SHS qui se positionne à la croisée des systèmes de gestion de contenus multimédias et des bases de connaissances. Okapi se distingue d'autres outils par sa très grande généricité, sa richesse fonctionnelle et son interopérabilité avec les détenteurs institutionnels d'archives multimédia et les bases de connaissances du *Linked Open Data*. Les différentes fonctionnalités de l'outil sont détaillées et illustrées à travers leurs usages au sein de différents projets de recherche et plus particulièrement dans le cadre de la mise en place d'un portail institutionnel innovant à l'Institut National des Langues et Civilisations Orientales (INALCO).

Mots-clés : plateforme Web, généricité, annotation de corpus multimédia, *Linked Open Data*, valorisation d'archives multimédias.

***Abstract**: We present in this paper Okapi, a web-based application for research communities in social sciences and humanities offering the combined functionalities of Multimedia Content Management Systems and Knowledge Bases. Okapi differentiates from other systems by its extreme genericity, its functional richness and its interoperability with both online institutional archives and the Linked Open Data (LOD). The functionalities of the tool are presented and illustrated by their utilization within different research projects and more especially for the deployment of an institutional portal at the french national institute of oriental languages and civilizations (INALCO).*

***Keywords**: Web application, genericity, multimedia content annotation, Linked Open Data, multimedia archive publishing.*

1 Introduction

La plateforme Okapi (*Open and Knowledge-based Annotation and Publishing Interface*) est une plateforme logicielle client-serveur développée au sein du département Recherche et Innovation de l'Institut national de l'audiovisuel (INA). Elle offre un

ensemble intégré d'outils et d'interfaces Web dédiés à la valorisation de contenus multimédias. Son objectif est de favoriser le partage, la capitalisation et la médiation de connaissances s'articulant autour de corpus de médias, en proposant notamment des fonctionnalités de description collaborative de contenus, de création de corpus d'extraits ainsi que de l'aide à la création de portails et parcours hypermédias. Okapi se distingue sur ce point des outils de « Content Management System (CMS) » existants, par la liberté laissée aux communautés ou institutions dans la conception de modèles de description adaptés à leur domaine et à leurs contenus. Cet aspect générique de la plateforme Okapi est central ; il irrigue et contraint la réalisation des différentes interfaces et en retour la rend potentiellement pertinente pour une multitude d'usages croisant base de connaissances et médias. Ces usages incluent par exemple la documentation de collections, l'analyse critique de corpus, l'analyse filmique, l'analyse de discours, l'éducation aux médias ou encore la cartographie de compétences.

L'outil est notamment utilisé dans différents projets de recherche pluri-disciplinaires : dans le cadre du projet Mémo-Mines pour la valorisation du patrimoine minier des Hauts-de-France, dans le cadre du projet ANTRACT s'intéressant à la valorisation du fonds des Actualités Françaises à des fins historiques, ainsi que dans le cadre du projet OOPAIP portant sur l'analyse du discours politique en interview.

Nous décrivons dans cet article la plateforme Okapi et présentons ses principaux outils de description et de valorisation d'archives multimédias ainsi que leur utilisation dans le contexte des projets de recherche Mémo-Mines et Inalco-Mediascape. La première section est consacrée à la présentation d'un principe fondateur d'Okapi qui est la généricité de ses procédés et mécanismes. Nous présentons dans la deuxième section, la méthodologie pour la constitution des ensembles formalisés de connaissances potentiellement réutilisables autour de contenus multimédias. La troisième section décrit la méthodologie de description de contenus en utilisant des ressources terminologiques : découpage, annotation, analyse, construction de corpus, et valorisation d'archives en produisant différents produits de médiation. Enfin, la dernière section présente l'utilisation de la plateforme Okapi pour la médiation des travaux de recherche menés à l'Inalco à travers le portail MediaScape.

2 La généricité au cœur d'Okapi

L'ambition de la plateforme Okapi est de servir de support technique à l'élaboration et à la valorisation de corpus de connaissances dans des domaines et pour des objectifs très divers. Il s'agit tout particulièrement d'éviter l'écueil d'un cadre de modélisation trop contraint résultant, par essence, de compromis et n'ayant de pertinence que dans un ensemble très fermé de cas, ou, à l'inverse, l'écueil d'un cadre très lâche, informel, dans lequel la sémantique n'est présente que de manière implicite dans des textes bruts.

À l'origine de la conception d'Okapi réside en effet le credo que toute démarche scientifique propre aux domaines des humanités numériques peut tirer bénéfice - en supplément du travail central, personnel ou collectif, d'identification et d'organisation des connaissances - d'un effort de formalisation de ces connaissances. Cette étape de formalisation comporte en effet de nombreux avantages par les possibilités qu'elle

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

offre tout d'abord en matière de cadre et d'aide conceptuelle à l'organisation même des connaissances mais aussi et surtout par la variété et la qualité des accès aux connaissances et aux contenus offerts par une telle étape de sémantisation/désambiguïsation, ainsi que par les possibilités complémentaires d'y asseoir des études quantitatives portant sur des données explicitées et fiables.

Parce que chaque chercheur, chaque communauté savante, chaque projet collaboratif, chaque institution a à cœur, d'utiliser et de donner à voir ses propres objets de classification, de description, d'analyse de ses ressources, l'approche proposée par Okapi est celle d'une « généricité instrumentée » destinée à laisser pleinement ouvert ce cadre d'expression tout en favorisant, par le respect de standards, la réutilisation ou l'extension de terminologies existantes.

3 Ontologies de domaines et graphes sémantiques

La plateforme Okapi utilise comme support de cette formalisation, les langages de représentation du Web sémantique promus par le W3C : RDF (*Resource Description Framework*), RDFS (*Resource Description Framework Schema*), OWL (*Ontology Web Language*) et SKOS (*Simple Knowledge Organisation System*). Ces langages offrent de puissants outils conceptuels qui permettent de construire des thésaurus ou des ontologies de domaines complexes. La formalisation de connaissances prend dans ce cadre la forme duale, d'une part, de taxonomies, de typologies de classes (ou types) d'objets, organisées en hiérarchies et, d'autre part, d'objets de ces classes (ou instances) possédant leurs propres attributs et liés entre eux par un ensemble de relations porteuses de sens formant un réseau sémantique, i.e. un graphe sémantique.

Construit sur ce socle conceptuel, l'apport d'Okapi en matière de modélisation et d'interfaces de saisie d'information consiste en deux contributions distinctes :

- la première est la définition d'un méta-modèle de description des connaissances et de ressources multimédias prenant la forme d'une « *upper ontology* ». Cette dernière ne comporte qu'un nombre très limité de concepts organisant les branches hautes des objets du domaine modélisé et les objets techniques servant à la description ;
- la deuxième est l'élaboration d'un mécanisme générique de suggestion contextuelle de propriétés et de plages de valeurs pertinentes destiné à faciliter la tâche de construction manuelle du graphe de connaissances et la description de contenus multimédias.

3.1 Méta-modèle de domaine Okapi

Le modèle de domaine d'Okapi s'inspire fortement des différentes « *upper ontologies* » connues et notamment de DOLCE-lite en reprenant la classification dichotomique de haut niveau en « Perdurant » et « Endurant » ainsi que la définition de grands types d'objets du domaine : personnes physiques, personnes morales/organisation, lieux, événements, concepts abstraits, etc. Cette taxonomie très succincte permet d'adjoindre un nombre limité de propriétés à chacun de ces grands types, propriétés qui seront réutilisables par les extensions de chaque domaine.

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

Chaque modélisation de domaine vient étendre le modèle Okapi en utilisant les mécanismes d'extensions propres aux langages RDFS/OWL. Ces derniers permettent également à un modélisateur de définir sa propre « *upper ontology* » et de déclarer des points d'adhérence uniquement entre certaines classes de son modèle et celles d'Okapi. La Figure 1 montre la simplicité du modèle générique et illustre les possibilités de classements offertes en terme d'organisation des connaissances.

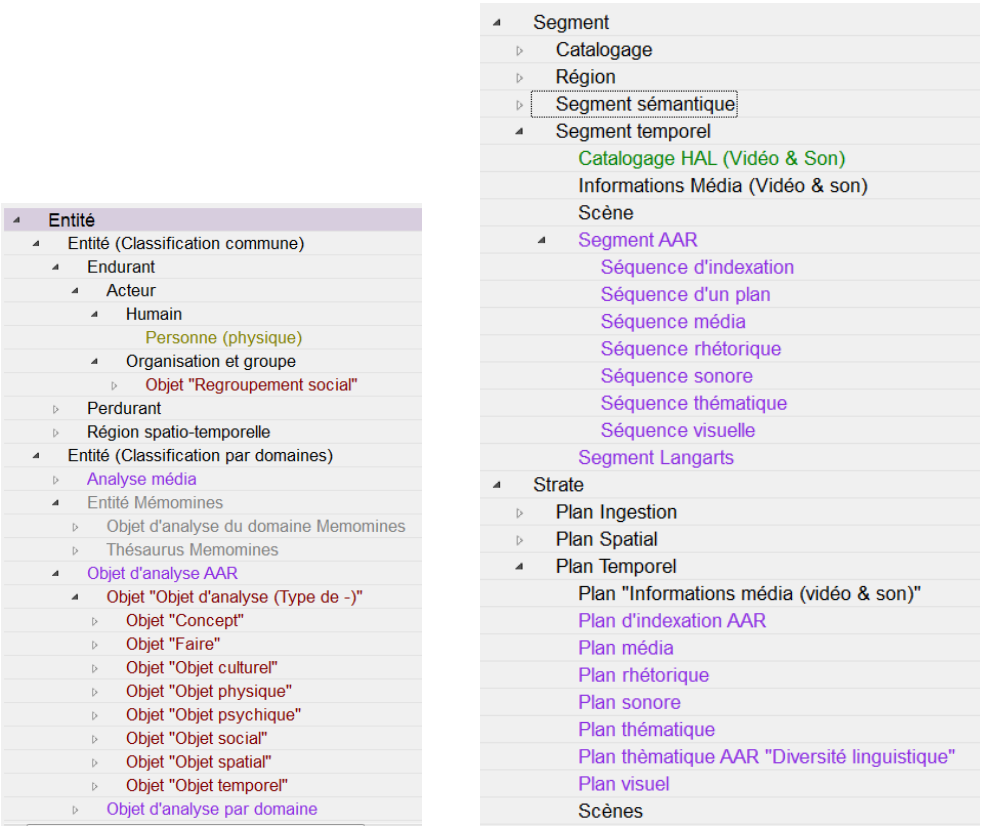


FIG. 1 : Hiérarchie des entités de domaines (en noir les entités du modèle Okapi, en couleur les extensions issues du modèle Mémo-Mines) et hiérarchie des segments et plans/strates de description

3.2 Méta-modèle de description multimédia

De manière similaire, le modèle de description multimédia d'Okapi regroupe un nombre limité d'objets structurant la description : le média (image fixe, vidéos, sons), la description (ou analyse de média), la strate (ou plan/point de vue de description), l'extrait (ou segment/fragment/région) de média, le corpus d'extraits et la publication hypermédia. À l'instar du modèle de domaine, l'ensemble de ces éléments est regroupé au sein d'une ontologie de description qu'il convient d'étendre afin de construire un modèle de description multimédia propre à son domaine. Le modèle proposé s'inspire notamment des notions de descriptions, de segments et de décomposition que l'on

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

retrouve dans des standards de description comme Mpeg7 et s'inscrit plus spécifiquement dans l'esprit des travaux de (Hunter, 2002) portant sur un portage de ce langage de XML Schema vers OWL.

Par le biais des mécanismes d'extension de OWL, le modèle permet de définir de manière déclarative les strates de descriptions que l'on souhaite utiliser pour un type de description particulier et d'associer à chaque type de segment porté par ces strates, une branche particulière du domaine. Le domaine ASA-SHS définit par exemple des plans thématiques, visuels, sonores et rhétoriques ; les extraits portés par le plan thématique sont associés à des objets « sujets », les extraits visuels portent, eux, sur une typologie d'objets visuels et de techniques de réalisation (Figure 2).

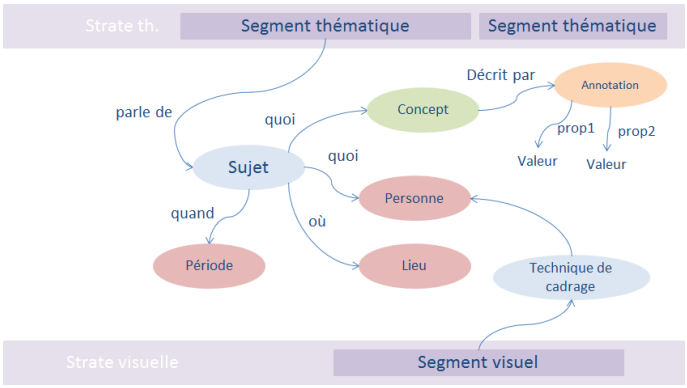


FIG. 2 : Représentation sous forme de graphe d'une description de contenu dans le modèle ASA-SHS avec deux types de strates et d'extraits audiovisuels, l'un portant sur l'aspect thématique et l'autre sur l'aspect visuel

3.3 Édition de graphes sous contraintes ontologiques

L'ensemble des descriptions de ressources multimédias et les entités du domaine prennent la forme de graphes sémantiques au format RDF. Le processus d'édition de ressources de la plateforme Okapi interprète les différentes contraintes logiques portées par les ontologies au format OWL pour guider l'annotateur vers la réalisation de graphes « bien formés ». À chaque nœud du graphe constituant une annotation, correspond un ensemble valide de propriétés et de valeurs de propriétés qui est spécifié dans les ontologies. Cette généricité et cette puissance d'expression impliquent en retour une compréhension certaine pour le modélisateur des logiques en œuvre au sein des langages de représentation des connaissances utilisés. Dans leur fondement, ces langages reposent en effet sur la logique des prédicats du premier ordre, la logique ensembliste et plus spécifiquement les logiques de description.

À la dérivation de types et de propriétés définies précédemment s'ajoute, dans le cas du langage OWL, la possibilité de contraindre pour une classe donnée les valeurs des propriétés héritées de ses classes ascendantes. Ce mécanisme permet de diminuer de manière drastique le nombre de propriétés nécessaires à la modélisation en encourageant la définition de propriétés à la sémantique suffisamment large et « orthogonales »

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

dans les branches hautes de l'ontologie. Ses principes facilitent l'appréhension d'ontologies comportant par opposition un grand nombre de classes.

La conception de modèles simples de taille réduite, est cependant à la portée d'un utilisateur averti et peut se limiter par exemple à la définition de thésaurus servant à un simple « *tagging* » d'extraits. Il est ainsi possible pour un chercheur ou une communauté de chercheurs de mettre au point un modèle documentaire parfaitement adapté à leurs besoins. La conception d'ontologies de taille importante constitue à l'opposé un travail considérable et il est préférable de s'appuyer sur des thésaurus ou des ontologies existantes comme par exemple le CIDOC-CRM. Le langage OWL offre notamment la possibilité de définir une ontologie par couche successive de spécialisation en utilisant un mécanisme hiérarchique d'import de modules.

4 Interfaces d'édition

4.1 Fonctionnement général d'Okapi

La plateforme Okapi se situe à l'intersection de trois disciplines : l'indexation de contenus, la gestion de connaissances et la gestion de contenus Web. Elle tire parti des avantages des normes et standards du Web sémantique (RDF, RDFS, OWL, SKOS (Motik 2012)) pour représenter l'ensemble des données sous forme de graphes de connaissances, pour raisonner et transformer ces connaissances afin d'enrichir et générer de nouveaux contenus.

Okapi fournit un ensemble d'outils destiné à l'analyse sémantique et sémiotique de contenus multimédia (vidéo, image, son) et à la gestion de corpus de vidéos, d'extraits sonores et de parties d'images. Les outils d'analyse permettent l'indexation et la description sémantique de contenu sur la base d'ontologies qui décrivent les objets et classes des domaines étudiés. Les outils de gestion de corpus fournissent des services pour la constitution, la visualisation, l'enrichissement et la valorisation de corpus thématiques.

L'outil de gestion de connaissances Okapi fournit des services d'édition, de stockage, de recherche et de suggestion d'entités nommées. Ces dernières regroupent habituellement les descriptions de personnes physiques, d'organisations, de lieux, de périodes ; d'autres types d'entités pertinentes peuvent être documentés en fonction du domaine modélisé. Ces entités peuvent être alignées avec d'autres entités stockées dans des bases de connaissances ouvertes sur le Web comme DbPedia, Wikidata, HAL ou Rameau. Cet alignement permet d'assurer l'interopérabilité d'Okapi avec l'écosystème du *Linked Open Data* (Bizer 2009).

À l'extrémité de la chaîne de valorisation, l'outil de publication Okapi prend en compte les caractéristiques du domaine étudié et les préférences utilisateurs pour générer divers produits de médiation (portail, site Web thématique, parcours narratif, etc.). Il permet à un auteur de créer des portails thématiques sans aucune compétence technique, en favorisant un travail d'*authoring* logique et en assurant la mise en forme du produit final. L'auteur peut ainsi constituer ses corpus et fournir le contenu thématique de sa publication ; Okapi y applique un ensemble de règles de transformation et génère le produit de médiation souhaité.

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

4.2 Édition de modèles

La plateforme Okapi offre son propre outil de modélisation compatible avec les langages OWL, SKOS et SKOS-XL. Cet outil peut être connecté à une base Okapi distante pour créer de nouveaux modèles de description, ou modifier les modèles existants voire importer des modèles tiers. Les fonctionnalités offertes peuvent se comparer à celles d'outils disponibles comme « Protégé », « *PoolParty* » ou « *TopBraid Composer* » tout en proposant des modalités d'édition de formules ensemblistes de contraintes de propriétés ainsi qu'une visualisation en couleur des concepts, utile pour différencier les apports des différents modules ontologiques constituant une ontologie domaine donnée (Figure 3).

L'outil de modélisation d'Okapi gère également les thésaurus au format SKOS ou SKOS-XL en autorisant leur création ex-nihilo ainsi que l'import et l'extension de thésaurus existants. De manière générale, l'outil fournit des interfaces pour les opérations de création, modification et suppression d'individus et leur mise en relation.

L'extensibilité de la plateforme en matière de modèles rend également possible l'intégration de descripteurs calculés par des algorithmes, comme la transcription de la parole, la segmentation en locuteurs, la segmentation paroles / musiques et bruits, la reconnaissance de textes dans le flux vidéo. Elle permet également de les interroger, de les évaluer et de les corriger.

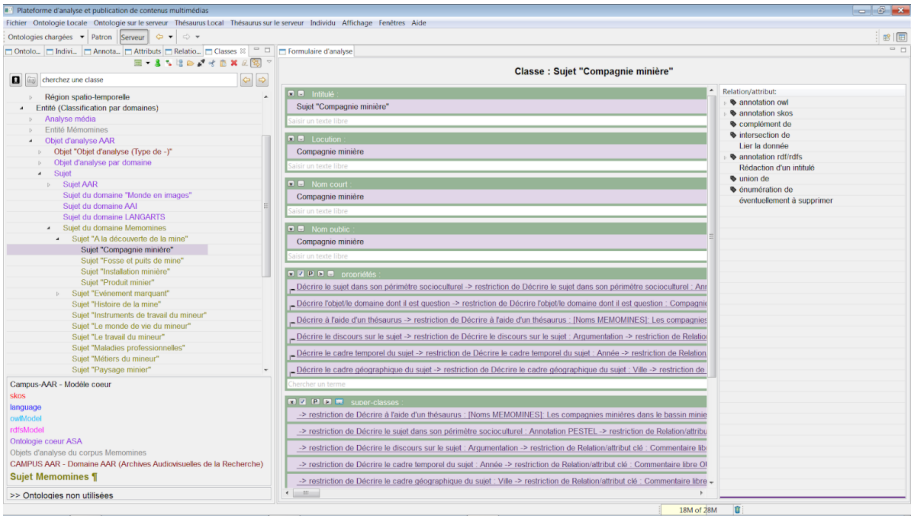


Fig. 3 : Outil de modélisation d'Okapi connecté à une base distante. Exemple d'édition de classes de « sujets » relatifs au domaine MEMO-MINES

4.3 Description de contenu

L'analyse d'un document multimédia (vidéo, image, son) dans Okapi est l'activité qui consiste à décrire sémantiquement ce document. Nous entendons par description sémantique toute description réalisée en fonction d'une ontologie de domaine et dont le résultat est un graphe de connaissances décrivant ce document. Cette analyse est

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

souvent réalisée manuellement par des utilisateurs spécialistes du domaine étudié. De manière alternative ou complémentaire, des descriptions issues d'algorithmes d'analyse automatique du contenu peuvent être intégrées dans la plateforme et mises à disposition des utilisateurs.

L'interface de saisie de données utilisée pour la description de contenu, le peuplement de la base de connaissances ou encore la réalisation de portails et de sites thématiques est constituée d'un formulaire Web dynamique entièrement dirigé par les ontologies de domaines utilisées. Des algorithmes de classification prennent en compte les contraintes OWL2 exprimées dans l'ontologie ainsi que le voisinage du nœud du graphe sémantique en cours d'édition pour inférer les propriétés disponibles et les plages de valeurs autorisées pour compléter le graphe. Ces algorithmes permettent une classification approchée avec un relâchement de contraintes qui autorise la prise en compte de la non-complétude transitoire des données au moment de l'édition d'un graphe. Ce mécanisme est abondamment utilisé pour contraindre en cascade les associations particulières entre les objets structurants (descriptions, strates et segments) et les objets du domaine. La Figure 4 montre le formulaire présenté pour la description thématique d'un extrait dans le cadre du projet ANR Memo-Mines. L'interprétation des contraintes définies dans l'ontologie oriente l'annotateur vers un ensemble de sujets propres à ce domaine particulier.

La description d'un document multimédia peut être réalisée suivant plusieurs axes. Selon Stockinger *et al.* (2015), la description peut concerner l'aspect thématique et consiste à identifier et décrire les sujets qui sont évoqués dans le document, les lieux et dates qui leur sont liés ainsi que les personnes qui y sont impliquées. Elle peut également s'étendre aux aspects sonores et visuels, et concerne les objets que l'on entend et voit à l'écran ainsi que les techniques de captation sonores et visuelles (plan large, plan serré, etc.) utilisées pour la réalisation de ce document. Enfin, la description rhétorique consiste à identifier la façon dont les sujets sont présentés dans le document. Plusieurs sujets peuvent être abordés dans un même document multimédia et y avoir des ancrages temporels (vidéo, son) ou spatiaux (image) différents. Okapi tient compte de ces caractéristiques et fournit des outils et interfaces permettant de découper virtuellement un document en plusieurs segments, et décrire séparément chacun d'entre eux comme présenté dans la Figure 4.

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

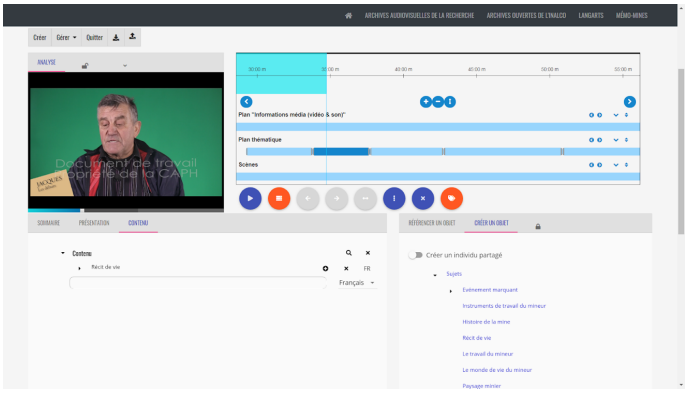


FIG. 4 : Interface de description audiovisuelle comportant sa « ligne de temps » éditable et le formulaire multi-onglets contrôlé par ontologie

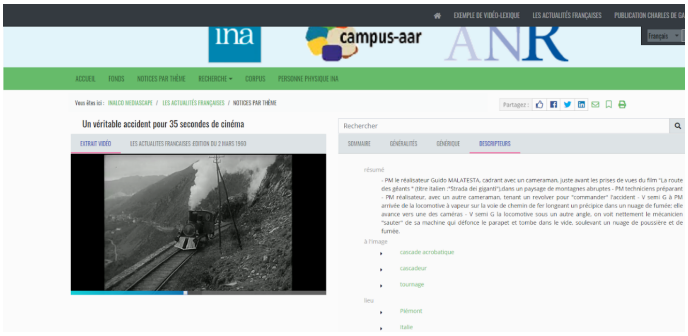


FIG. 5 : Reprise d'écran du portail Okapi du projet ANR ANTRACT portant sur le fonds des actualités filmées (1200 émissions du fonds INA des Actualités Françaises, presse filmée)

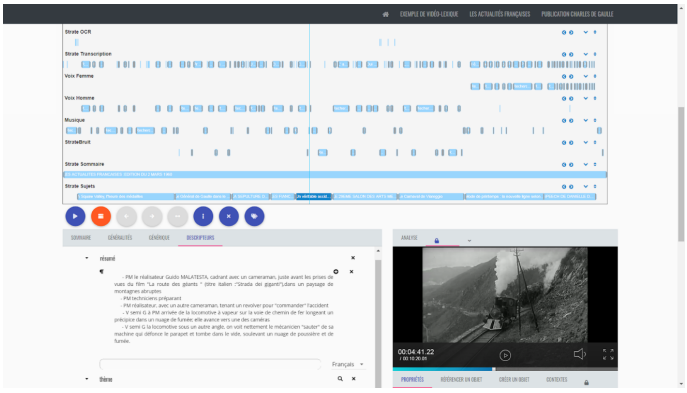


FIG. 6 : Écran d'annotation en ligne montrant un ensemble de descripteurs calculés automatiquement ainsi que la documentation INA des sujets. Une segmentation manuelle contrôlée par le type de strate peut compléter la description. Projet ANR ANTRACT

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

Okapi offre aussi un ensemble d'outils pour la description d'images. À l'instar de la description de vidéos, les utilisateurs peuvent créer plusieurs strates de description qui agissent comme des calques pour délimiter des régions d'image différentes, suivant le type de description envisagée (thématique, visuelle, etc.). La Figure 7 montre l'interface principale de découpage virtuel d'images en plusieurs régions de différentes formes (rectangulaires, circulaires ou à main levée). Chaque région peut être décrite à l'aide d'un formulaire spécifique.



Fig. 7 : Annotation de régions d'image dans le cadre du projet Mémo-Mines

4.4 Constitution de corpus

Un corpus est un ensemble d'extraits vidéo et/ou régions d'image qu'il est également possible d'annoter avec ses propres métadonnées. Le corpus fournit un contexte d'annotation particulier laissant possible la sur-annotation contextuelle des extraits, ainsi que leur mise en relation. L'utilisateur peut ainsi ajouter des relations rhétoriques entre deux éléments du corpus. Par exemple, un extrait vidéo (E1) qui introduit, contredit ou exemplifie un autre extrait (E2). Ces métadonnées seront exploitées par le processus de publication pour proposer une mise en page qui respecte et reflète l'intention de l'auteur du corpus. Dans cet exemple, le processus de publication peut établir un ordre d'affichage des deux extraits en montrant E1 avant E2.

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

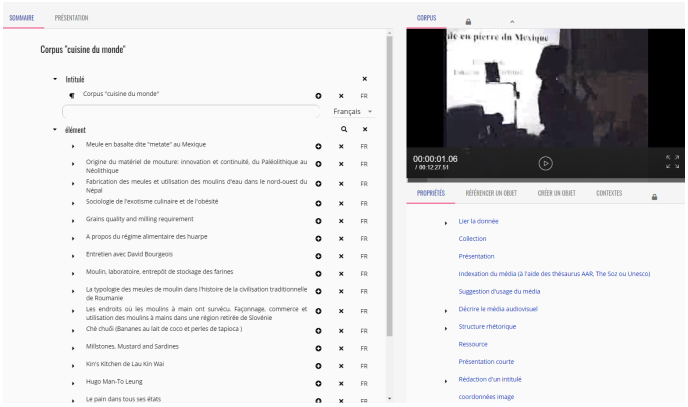


Fig. 8 : Interface d'édition de corpus.

Okapi offre également des outils pour la valorisation des corpus. Une page de portail peut être créée pour l'affichage et navigation dans des corpus (Figure 9). Elle montre la liste des éléments d'un corpus, leurs métadonnées ainsi que les extraits vidéos et régions d'image correspondants. Un outil de montage est également proposé par la plateforme et permet d'attribuer un ordre d'affichage à chaque élément d'un corpus et de générer une page Web qui visualise le montage de ces éléments. Enfin, les corpus créés dans Okapi peuvent être intégrés dans une publication auteur de type vidéo-lexique ou parcours narratifs pour illustrer ou appuyer des propos autour d'une thématique donnée.

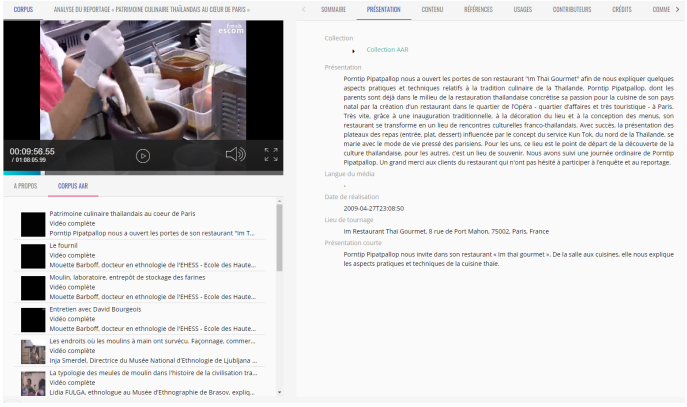


Fig. 9 : Visualisation d'un corpus sur le portail.

4.5 Gestion de la base de connaissances

L'analyse sémantique d'un document multimédia dans Okapi est représentée par un graphe de connaissances qui décrit ce document. Ces connaissances sont stockées dans Okapi sous forme d'une base de connaissances et suggérées aux utilisateurs pendant leur travail d'analyse. Elles concernent essentiellement deux types d'objets : entité nommée et concept. Les entités nommées sont des individus et objets identifiés du

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

monde réel et peuvent être de différents types : personne physique, organisation, lieu géographique, période temporelle, événement, etc. Les concepts sont plutôt des abstractions d'objets du monde réel comme « voiture », « bâtiment », etc. En langage documentaire, l'ensemble de ces concepts constituent un thésaurus.

Okapi fournit des interfaces d'édition et de visualisation d'entités nommées.

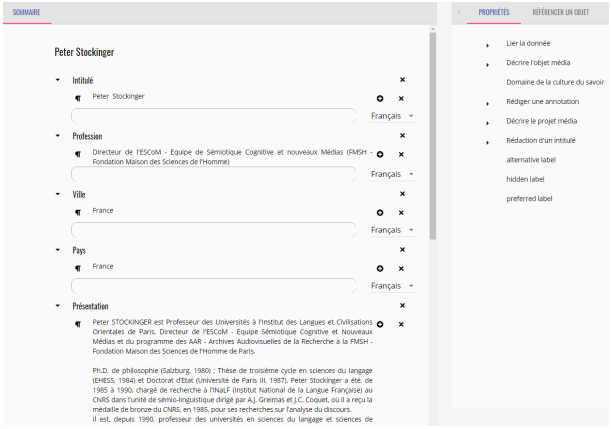


Fig. 10 : Interface d'édition d'entités nommées, le formulaire est propre au type de l'entité éditée : ici, une personne physique.

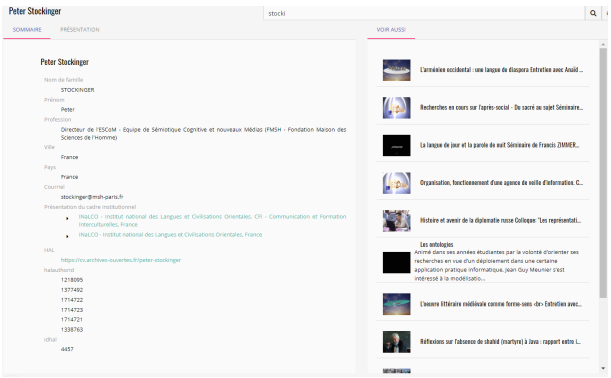


Fig. 11 : Visualisation de la documentation d'une entité nommée en mode navigation avec accès aux ressources la référençant.

Okapi peut facilement s'intégrer dans un écosystème de bases de connaissances ouvertes disponibles sur le Web (*Linked Open Data*). Des référentiels existants, comme Rameau, peuvent être utilisés dans Okapi pour annoter des documents multimédias et alternativement, la base de connaissances Okapi peut être alignée sur d'autres bases de connaissances comme DbPedia ou WikiData. Cela rend les données Okapi interopérables avec des données ouvertes disponibles sur le Web et les documents Okapi atteignables et accessibles par les termes d'un autre référentiel. Cet alignement permet aussi d'enrichir la base Okapi par des données contenues dans d'autres bases de

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

connaissances. Ces dernières étendent les capacités de navigation dans les données et autorisent la réalisation d'études quantitatives inenvisageables sur la seule base des données produites.

5 La plateforme Inalco–MediaScape

Si l'INA défend un point de vue autour de la généricité au travers des ontologies ASA (Atelier de Sémiotique Audiovisuelle) et AAR (Archives Audiovisuelles de la Recherche), le méta-portail MediaScape déployé à l'Inalco est quant à lui, un prototype en cours de développement s'appuyant sur Okapi et remplissant deux fonctions essentielles : l'archivage et l'analyse. Le portail MediaScape se veut un outil orienté utilisateur et regroupe plusieurs projets de recherche centrés sur la valorisation de corpus audiovisuels.

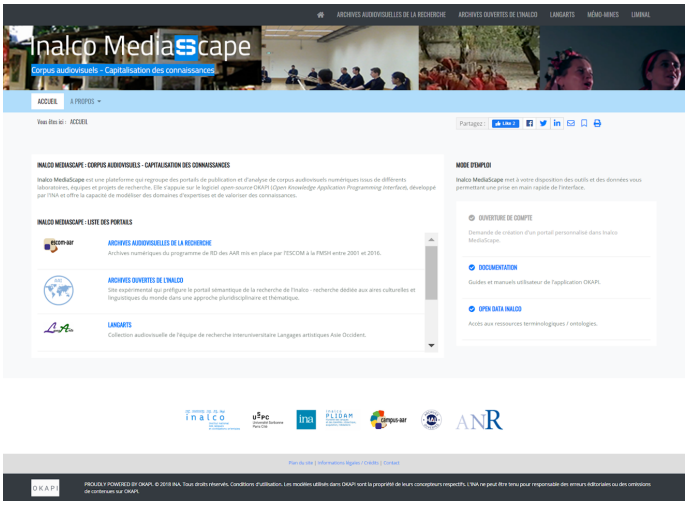


Fig. 12 : Capture d'écran du portail Inalco-MediaScape

Au sein de la plateforme Inalco-MediaScape, quatre portails sont en cours de développement, dont le plus abouti est celui des AAR (Archives Audiovisuelles de la Recherche). Ce programme qui s'est déroulé de 2001 à 2016 au sein de la Fondation Maison des Sciences de l'Homme regroupe les analyses des corpus vidéos réalisées pendant quinze ans au sein de ce programme.

Construit sur le même modèle, le portail des Archives Audiovisuelles de l'Inalco (AAI) revêt un enjeu important pour l'Inalco : il présente les documents audiovisuels réalisés autour des principales thématiques développées par l'Inalco en lien avec les langues et les cultures du monde. Cet espace est le futur site des Archives Ouvertes de l'Inalco (AOI) qui hébergera les analyses réalisées sur le fonds appartenant à l'institution.

Langarts est également une collection audiovisuelle de l'équipe de recherche interuniversitaire « Langages artistiques Asie-Occident ».

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

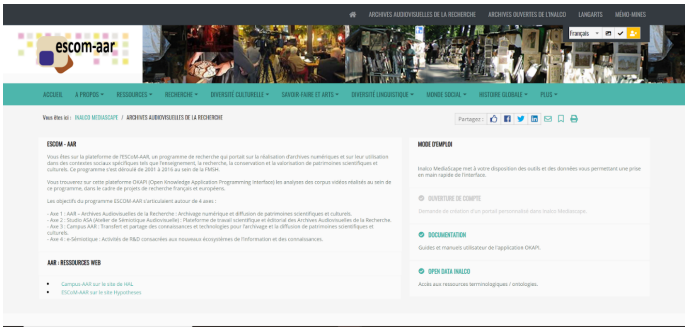


Fig. 13 : Capture d'écran du portail des AAR

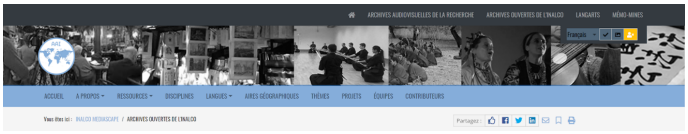


Fig. 14 : Capture d'écran de la bannière du portail des AAI

Le projet ANR Mémo-Mines porte sur la constitution et l'exploitation des corpus audiovisuels documentant le monde des mines du nord de la France, autrement dit, la conservation de faits ou d'événements constituant un patrimoine culturel en danger. Les actions de recherche menées dans le cadre de ce projet ANR sont également disponibles au sein d'un carnet hypothèses.

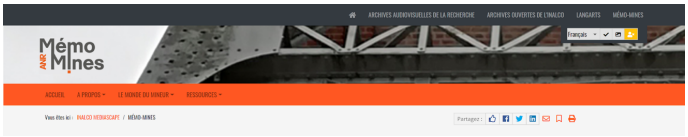


Fig. 15 : Capture d'écran de la bannière du portail Mémo-Mines.

Sont donc accessibles des domaines de connaissance différents et spécialisés : le patrimoine (Mémo-Mines), les archives des SHS (AAR), les langues et les cultures du monde (Langarts et AAI).

La plateforme Inalco-MediaScape s'adapte ainsi aux spécificités de chaque domaine de connaissances tout en utilisant le même modèle de données. Aussi, d'autres portails et d'autres projets (portés par l'Inalco ou par des initiatives individuelles) peuvent être ajoutés. Les chercheurs sont donc invités à archiver leurs données sur HAL ou des plateformes respectant le standard OAI (*Open Archives Initiative*), les langages Dublin Core ou TEI (*Text Encoding Initiative*) permettant l'échange de métadonnées. Via Okapi, les données sont « moissonnées », les vidéos référencées et une classification, une annotation des publications ou des republications peuvent être réalisées.

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

La plateforme Inalco-MediaScape prévoit un mode d'emploi avec, à terme, une possibilité d'ouvrir des comptes utilisateurs pour alimenter et analyser leurs fonds. Une page de documentation donne accès aux guides et tutoriels vidéos.

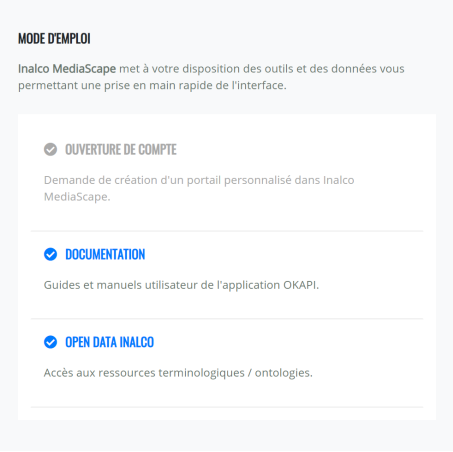


Fig. 16 : Capture d'écran du mode d'emploi présent sur le portail Inalco-MediaScape

Partisan de l'Open Data, initiative nationale portée par Madame la Ministre Vidal qui, en juin 2018 a décidé de faire de la science ouverte une de ses priorités, l'Inalco s'inscrit dans ce paradigme en rendant accessibles et disponibles les métadonnées et les ontologies de la plateforme. Dès à présent, environ 500 analyses et près de 1.500 heures de vidéos analysées sont à la disposition de la communauté.

5.1 Inalco-MediaScape : faire parler les données

Inalco-MediaScape s'inscrit dans le paradigme des humanités numériques appliquées aux recherches, à l'enseignement, à la formation et à l'expertise des aires culturelles et linguistiques. Il est dédié plus particulièrement à la problématique de la collecte et du traitement de corpus de données visuelles, audiovisuelles et sonores générées par la recherche, au service de la recherche et de l'enseignement. Le potentiel de développement d'un tel projet nous semble considérable puisque l'on peut dire que la majeure partie des données sur Internet sera dans les prochaines années des données audiovisuelles. La question qui se pose est celle de savoir comment intervenir activement sur un média audiovisuel pour le transformer en une ressource appropriée à des contextes d'usages spécifiques comme l'enseignement, la recherche ou la valorisation d'un patrimoine scientifique.

La plateforme Inalco-MediaScape poursuit donc l'objectif de faire parler les données, faire parler les archives, en faisant en sorte qu'elles se comportent un peu comme des acteurs qui répondent à nos horizons d'attentes. Il s'agit d'une approche sémiotique des archives, considérant que chacune d'elles forme un univers de sens thématisé, exposé, traité, exemplifié par les objets média qui en constituent le fonds. Il s'agit de « faire parler » les corpus audiovisuels mais aussi les objets média, les vidéos ainsi que les images, le texte, etc. d'une archive.

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

Nous entendons par archive, un ensemble de données qui à la fois documentent, conservent et transmettent le discours d'un acteur social (d'une institution, d'un groupe, d'une communauté, voire d'une personne) sur un domaine, un objet ou encore une époque. De notre point de vue, Okapi remplit exactement cette mission d'archives et répond à un besoin des chercheurs d'exploiter des données dans des directions parfois diverses.

L'usage courant de l'archive est celui de la simple consultation des données qu'elle contient pour s'en inspirer, pour trouver des nouvelles idées ou pour s'en servir sous forme d'une « bonne pratique » par exemple.

L'Inalco a pour ambition d'utiliser ou, plutôt, de réutiliser un corpus de données archivées dans le cadre d'une activité de transmission d'une culture, orale ou écrite.

Pour illustrer notre propos, le projet scientifique DHonsite, *Humanités Numériques et Littérature numérique : Lire ensemble, Reading Together* vise à proposer des interfaces interculturelles de consultation d'un corpus Gallica de productions, relatif à des missions françaises au Bénin, principalement aux 18e et 19e siècles. La conférence internationale qui s'est tenue à Cotonou les 2-3 mai 2019 a permis d'évaluer ces propositions en immersion africaine et de rechercher plus généralement les conditions d'une lecture commune.

L'idée est de créer ensemble des médiations ou des modes de communications communs en incluant les différences culturelles. La question posée est celle de savoir s'il est possible de créer des interfaces ou des dispositifs numériques permettant des relations entre le point de vue africain et le point de vue occidental. Dans cette perspective, l'Inalco, médiateur reconnu entre espaces francophones et non-francophones, permet d'interroger l'interculturalité afin de comprendre et de modéliser les formes de médiations nouvelles.

L'Afrique a une longue tradition orale que l'audiovisuel retranscrit fidèlement, constituant ainsi le médium de choix pour faire entrer l'oralité dans le numérique. Le document audiovisuel, la vidéo, permet non seulement de « saisir » et de rendre compte, mais aussi de constituer un corpus d'études, de l'analyser et de le valoriser dans le cadre de diffusions futures. La plateforme Okapi permet de révéler des points de vue de communautés de recherche à partir de corpus audiovisuels mais aussi d'images et de textes. L'analyse de corpus rendue possible donne l'opportunité de révéler les points de vue africains et occidentaux sur un même support. Okapi offre donc ces conditions d'un « lire ensemble » en stimulant ou produisant de « l'intelligence collective ».

5.2 La promesse d'Okapi

Ce prototype cherche à valoriser les corpus audiovisuels, dans le champ des humanités numériques, avec une vision et des valeurs défendues par les équipes qui s'y investissent.

Nous en avons sélectionné trois principales.

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

- La transmission : le processus est de nourrir Okapi de ressources et connaissances pour en faire un réservoir de mémoire mobilisable par différents auteurs pour exprimer différents points de vue, sur ces contenus et selon une intentionnalité donnée.

- La co-construction des savoirs : outre le fait que la plateforme en elle-même mobilise différents rôles (informaticiens, graphistes, anthropologues, historiens, management de projet), elle permet une analyse collaborative des savoirs et une médiation des ressources. Plusieurs personnes peuvent en effet intervenir sur Okapi en même temps sur les mêmes contenus. Pendant la construction même de l'outil Okapi, les utilisateurs sont placés en position de « beta testeurs » : il s'agit d'une démarche de co-participation de ceux qui produisent les données et proposent des améliorations de fonctionnalités.

- Le partage : le partage des ressources constitue un enjeu majeur. Un acteur peut profiter des analyses et des connaissances d'un autre acteur de la plateforme. Okapi est un lieu de partage de ressources et de connaissances, mais aussi de controverses, de confrontations de points de vue autour d'un concept. Il constitue ainsi un outil pertinent pour l'éducation à l'image.

6 Conclusion

Okapi est une plateforme intégrée de modélisation, description et valorisation d'archives audiovisuels, favorisant le travail collaboratif des utilisateurs sur des documents communs.

La multiplication des documents et des supports nous incite à repenser les outils de lecture et d'écriture, afin de rester des « lecteurs créatifs » (Manguel, 1997), voire des « citoyens créatifs » (Farge, Rutherford, 2019) dans notre approche de la lecture et de l'utilisation des outils numériques.

En proposant des ateliers collaboratifs à destination des étudiants et en mettant à disposition des corpus sur lesquels ils peuvent échanger, Okapi devient un puissant outil d'éducation aux médias. Ces valeurs inscrites dans l'outil lui-même, proposent aux utilisateurs d'être réceptifs, sensibles aux médias qu'ils manipulent, favorisant ainsi le développement d'une culture numérique, voire d'une culture de l'outil (Farge, 2015), préfigurant nos pratiques et nos manières de faire, l'enjeu étant une prise de conscience collective de l'importance de la lecture critique de l'information.

Okapi nous offre cette opportunité et relève le défi de constituer un corpus de connaissances, en rendant les archives vivantes. La valeur du partage, permet l'émergence non pas uniquement d'une culture commune mais d'une conscience commune favorisant l'interaction et la manifestation d'une économie du sens, c'est-à-dire la participation active à la création de richesses notamment relationnelles dans le respect de l'altérité.

Remerciements

Les auteurs remercient Peter Stockinger pour sa contribution à la réalisation du portail Inalco-MediaScape et plus précisément à l'élaboration des modèles de description et à la suggestion de nouvelles fonctionnalités de la plateforme Okapi.

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799

Financements

La plateforme Okapi a été et est partiellement subventionnée par l'ANR dans le cadre de plusieurs projets de recherche CAMPUS-AAR (ANR-13-CORD-0016-01), MEMOMINES[1], ANTRACT[2] et LIMINAL[3].

Références

Beloued, Abdelkrim ; Peter Stockinger ; Steffen Lalande, 2017, "Studio Campus AAR : A Semantic Platform for Analyzing and Publishing Audiovisual Corpuses". In *Collective Intelligence and Digital Archives*, John Wiley & Sons Inc., p. 85-133.

Beloued, Abdelkrim ; Steffen Lalande ; Peter Stockinger, 2015, « Modélisation et formalisation RDFS/OWL d'une ontologie de description audiovisuelle ». Dans *Archives audiovisuelles et valorisation du patrimoine à l'ère du numérique* sous la direction de Tarek Ouerfelli, Les Cahiers du Numérique, vol 11/3, p. 39-70.

Bizer, Christian ; Tom Heath ; Tim Berners-Lee, 2009, "Linked data – the story so far". Int.J. Semantic Web Inf. Syst., 5,1-22.

Farge, Odile ; Bianca Rutherford-Iglesias, 2019, « Pour une ville numérique et humaniste : entre espaces de pouvoir et pratiques artistiques de résistance urbaines ». Dans *Communication, City and Public Spaces. Les Éditions de l'Immatériel, collection « Écritures du Monde »*, p. 240-262.

Farge, Odile, 2015, « La "rhétorique de la conception" : pour une conscientisation du rôle de l'outil dans la formation d'une culture numérique ». Dans *Interfaces Numériques*. Volume 4, No.3, p. 525-540.

Hunter, Jane, July 30 – August 1, 2001, "Adding Multimedia to the Semantic Web – Building an MPEG-7 Ontology". In : *Proc. of International Semantic Web Working Symposium (SWWS)*, Stanford.

Lalande, Steffen ; Kamenka Staykova ; Michel Chein ; Alain Gutierrez ; Valia Saraydarova, et Danail Dochev, 2009, "Using domain knowledge to speed up the annotation of digital content with conceptual graphs". In *Cybernetics and Information Technologies*, 9(3).

Manguel, Alberto, 2010, *A Reader on Reading. New Haven*. Yale : University Press.

Motik, Boris ; Peter F. Patel-Schneider ; Bijan Parsia, 11 December 2012, "OWL 2 Web Ontology Language" eds. W3C Recommendation.

Nanard, Marc ; Jocelyne Nanard, Nov 2005, Saphir : « Un Cadre de Référence pour une Spécification des Hypermédias par Intention ». Dans *HT2PTM'05 : Hypermedias Hypertexts Products Tools and Methods*, Paris, France. Hermès, 10 p.

Stockinger, Peter ; Steffen Lalande ; Abdelkrim Beloued, 2015, « Le tournant sémiotique dans les archives audiovisuelles : Vision globale et éléments conceptuels de mise en œuvre ». Dans *Archives audiovisuelles et valorisation du patrimoine à l'ère du numérique* sous la direction de Tarek Ouerfelli, Les Cahiers du Numérique, VOL 11/3 – p. 11-37.

Troncy, Raphaël ; Jean Carrive ; Steffen Lalande ; Jean-Philippe Poli, October 2004, "A motivating scenario for designing an extensible audio-visual description language". In *CORIMEDIA'04*, Sherbrooke, Canada.

[1] https://anr.fr/Projet-ANR-16-CE38-0001__ANR-16-CE38-0001
[2] https://anr.fr/Project-ANR-17-CE38-0010__ANR-17-CE38-0010
[3] https://anr.fr/Project-ANR-17-CE41-0004__ANR-17-CE41-0004

Published in : Louise Ouvrard et Odile Farge (dir.), *Corpus audiovisuels*, Editions des archives contemporaines, ISBN : 9782813003799, doi : 10.17184/eac.9782813003799