

Le thesaurus, instrument de l'interopérabilité sémantique : faisabilité d'un rapprochement des vocabulaires thématiques dans le cadre de la fusion de portails d'accès à des collections cinématographiques

Guénaël Eveno

▶ **To cite this version:**

Guénaël Eveno. Le thesaurus, instrument de l'interopérabilité sémantique : faisabilité d'un rapprochement des vocabulaires thématiques dans le cadre de la fusion de portails d'accès à des collections cinématographiques. domain_shs.info.docu. 2016. mem_01476045

HAL Id: mem_01476045

https://memsic.ccsd.cnrs.fr/mem_01476045

Submitted on 24 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

Ecole Management et Société-Département CITS

INTD

MEMOIRE pour obtenir le Titre enregistré au RNCP

"Chef de projet en ingénierie documentaire et Gestion des connaissances"

Niveau I

Présenté et soutenu par

Guénaël Eveno

le 07 décembre 2016

Le thesaurus, instrument de l'interopérabilité sémantique

Faisabilité d'un rapprochement des vocabulaires thématiques dans le cadre de la fusion de portails d'accès à des collections cinématographiques

Jury : Claire Scopsi, Directrice de Mémoire

Laurent Bismuth, Responsable de Stage

Promotion 46

Paternité Pas d'Utilisation Commerciale - Pas de Modification

A Marie Annick Eveno, ma mère

Une femme exceptionnelle.

REMERCIEMENTS

Je tiens à remercier Claire Scopsi pour m'avoir guidé dans cette première mission de chef de projet, et pour ses conseils dans la rédaction de ce mémoire.

Un grand merci à Elodie Gilbert pour son accueil chaleureux et sa disponibilité quels que soient mes questionnements. Et aussi pour le café !

Merci à Laurent Bismuth pour nos échanges durant le stage et pour avoir facilité la phase d'étude de besoins.

Merci à Martine Vignot pour avoir facilité les entretiens à la Cinémathèque Française.

Merci aux équipes du CNC, particulièrement au service AGDC, et aux équipes du traitement des collections de la Cinémathèque Française, pour leur disponibilité et pour leurs bons conseils.

Merci à Sylvie Dalbin et Cécille Kattnig pour leurs apports respectifs à ma réflexion et à ce mémoire.

Merci à Lucie Gay pour sa relecture et ses corrections avisées.

Merci à Vincent Rappeneau, Boris Blanckemane, Adib Kassas et Anna Marenelly pour avoir fait vivre le Fonds Interstagiaire de Soutien Technique.

Merci enfin à mon père pour son soutien.

NOTICE

EVENO, Guénaël. Le thesaurus, instrument de l'interopérabilité sémantique : Faisabilité d'un rapprochement des vocabulaires thématiques dans le cadre de la fusion de portails d'accès à des collections cinématographiques. Mémoire professionnel INTD, Titre 1, Chef de projet en Ingénierie Documentaire et Gestion des Connaissances. Conservatoire national des arts et métiers – Institut National des Sciences et Techniques de la Documentation. 2016. 104 p. Promotion 46.

Résumé : Ce mémoire présente les enjeux et le processus de création d'une interopérabilité entre différents thesauri. Il décrit les spécificités du thesaurus, son évolution d'outil de recherche d'information à outil de structuration sémantique aidant à créer l'interopérabilité sémantique et un accès thématique unique à de multiples ressources.

A travers l'exemple de la fusion de portails donnant accès à des collections patrimoniales de films et de « non-films », il développe les étapes nécessaires à la réflexion du chef de projet dans l'étude du rapprochement de plusieurs vocabulaires afin de pouvoir présenter des requêtes sur tous types de documents.

Descripteurs : Thésaurus, films, recherche d'information, portail documentaire, interopérabilité, web sémantique, recherche fédérée, patrimoine culturel, étude de faisabilité, analyse de besoin.

Abstract : This Research Paper explains the issues and the process of creating interoperability between several thesauri. It develops thesaurus' distinguishing features with other vocabularies, its evolutions from an information retrieval tool to a knowledge organization system helping to create semantical interoperability and a unified thematical access to multiples ressources.

By the case of the merging of two webportals giving access to « films » and « non-films » patrimoines collections, il emphasizes the necessary steps for the chief of project's reflections in order to map different vocabularies.

Keywords : Thesaurus,films, information retrieval, documentary webportal, interoperability,linked data, cultural heritage, feasibility study, requirements analysis.

TABLE DES MATIERES

Remerciements 3

Notice 4

Table des matieres 5

Table des illustrations 8

INTRODUCTION 9

PREMIERE PARTIE : ROLES DU THESAURUS DANS L'ACCES A L'INFORMATION 11

1. Un outil de recherche d'information 12

1.1 Fonctions et caractéristiques d'un thesaurus 12

1.1.1. Du thesaurus dictionnaire au thesaurus documentaire 12

1.1.2. Caractéristiques du thesaurus 12

1.1.2.1 Un langage pour indexer et rechercher les documents 13

1.1.2.2 Un langage contrôlé et normalisé 14

1.1.2.3 Un langage structuré et ordonné 14

1.1.2.5 Un langage combinatoire 15

1.2 Les alternatives au thesaurus 16

1.2.1 Thesauri et classifications 16

1.2.2 Thesauri et taxonomies 17

1.2.3 Thesauri et ontologies 18

1.2.4 Thesauri et listes d'autorités (vedette matière) 19

1.2.5 Thesauri et indexation collaborative 20

1.3 Thesauri et système de recherche d'information 21

1.3.1 Thesaurus et besoins du grand public : un mariage difficile ? 21

1.3.2 La recherche libre contre la recherche contrôlée 24

1.3.3 Balisage sémantique et « Microdonnées » 25

1.3.4 Le thesaurus en appui aux systèmes de recherche d'information 26

2. Un système d'organisation des connaissances (SOC) pour l'interopérabilité sémantique 29

2.1 Des contextes propices à l'interopérabilité sémantique 29

2.1.1 Définition et portée de l'interopérabilité 29

2.1.2 Interopérabilité et recherche fédérée 30

2.1.2.1 Interrogation de sources de données hétérogènes 30

2.1.2.2 Un exemple d'interopérabilité à la recherche : Le projet OTAREN 31

2.1.3 L'essor du web de données liées 32

2.1.3.1 Origines et applications du web de données liées 32

2.1.3.2 Les outils du web de données liées 35

2.1.3.3 Les principes du web sémantique 38

2.2 La normalisation, vecteur de l'interopérabilité des langages documentaires 39

2.2.2 ISO 25964-1 : Rendre les thesauri interopérables 41

2.2.2.1 Des supports techniques 42

2.2.2.2 Une distinction claire entre les concepts et les termes 43

2.2.2.3 Le multilinguisme 43

2.2.2.4 La reconnaissance de SKOS 44

2.2.2.5 Plus de spécification dans les relations 44

2.2.2.6 La création de domaines thématiques – subject areas 44
2.2.3 ISO 25964-2 : Rendre les vocabulaires interopérables 45
2.2.3.1 Définition et portée des alignements 45
2.2.3.2 Des recommandations pour les alignements 45
2.2.3.3 Exemple de relation entre les vocabulaires 47
2.2.3.4 Utilisation des alignements à l'indexation et à la recherche 47
2.3 Applications concrètes de l'interopérabilité sémantique des thesauri 48
2.3.1 Coût et avantages d'un thesaurus interopérable 48
2.3.2 L'interopérabilité sémantique au sein du web de données culturelles 49

DEUXIEME PARTIE : AU CŒUR DE L'INTEROPERABILITE DES VOCABULAIRES FILMS ET NON FILMS 53

3. Le thesaurus, enjeu du rapprochement documentaire du CNC et de la Cinémathèque Française 54
3.1 Les Archives Françaises du Film et la Cinémathèque Française, deux visions de la valorisation du patrimoine cinématographique 54
3.1.1 Une rivalité historique 54
3.1.2 Evolution documentaire des Archives du Film et description des films 55
3.1.2.1 Du stockage à la valorisation 55
3.1.2.2 La base documentaire Lise 57
3.1.2.3 Le traitement documentaire, maillon d'une orientation vers la valorisation des collections 59
3.1.3 Evolution documentaire de la Bibliothèque du Film (BiFi) et de la cinémathèque française 60
3.1.3.1 Définition et développement des fonds non-films 60
3.1.3.2 Création du thesaurus *Cinédoc* et indexation iconographique 61
3.1.3.3 Création du portail *Ciné-ressources* et du site web 62
3.2 Le projet plateforme, un outil de rapprochement des pratiques documentaires des équipes 64
3.2.1 Contexte et enjeu de la mutualisation 64
3.2.2 Genèse et développement du projet plateforme 65
3.3 Les enjeux d'une recherche fédérée thématique au niveau interne 66

4. Conduire l'étude de faisabilité d'un thesaurus film et non-films : Méthode et ajustements 69
4.1 Importance d'une étude préalable à la construction d'un thesaurus 69
4.1.1 Mobiliser autour du projet en interne 69
4.1.2 Intégrer le thesaurus dans un système 69
4.1.3 Documenter la construction du thesaurus 70
4.2 Choix Méthodologiques inhérents au projet 72
4.2.1. Un projet au sein du projet 72
4.2.2 Harmoniser les vocabulaires 73
4.2.3 Une position d'expert extérieur 74
4.3 Analyser le besoin 75
4.3.1 Pourquoi analyser le besoin ? 75
4.3.2 Définir une typologie d'utilisateurs 76
4.3.2 Elaboration, conduite et synthèse d'entretiens 76
4.4 L'importance d'établir un état de l'art des langages documentaires 78
4.4.1 Collecter les langages existants 80
4.4.2 Evaluer les langages collectés 80
4.5 Analyser l'existant et l'environnement documentaire 81
4.5.1 Utilité de l'analyse de l'existant 81
4.5.2 Des outils pour auditer les langages contrôlés « actuels » 83

4.5.2.1 Les occurrences d'indexation des termes 83
4.5.2.2 Des outils qualitatifs 83
4.6 Etablir un benchmark, difficultés et atouts 84
4.7 Analyser les contraintes 85

5. Des perspectives pour la construction et la gestion d'un thesaurus films et non films 86
5.1 Le livrable de faisabilité, concrétisation de l'étude 86
5.1.1 Objectifs et structure du livrable de faisabilité 86
5.2 Vue sur les préconisations 87
5.2.1 Scénario 1 : Fusion des vocabulaires films et non-films 87
5.2.2 Scénario 2 : Alignement des vocabulaires films et non-films via un outil externe 89
5.2.4 Scénario 3 : Alignement des vocabulaires films et non-films via l'outil Adlib 90
5.2.4.1 Intégration de thesaurus dans Adlib : l'exemple du British Film Institute 91
5.2.4.2 Application aux vocabulaires de *Lise* et *Cinéressources* 92
5.2.4.3 Exemple d'alignement 93
5.3 Perspectives de gestion et de maintenance pour les services documentaires des partenaires 94
5.3.1 Intégrer le thesaurus à un outil de gestion 94
5.3.2 Une évolution des langages documentaires de Lise et Cinéressources 95
5.3.2.1 Un thesaurus pour décrire les films 95
5.3.2.2 Une adaptation des vocabulaires non-films 97

Conclusion 98

BIBLIOGRAPHIE 99

ANNEXES 108

Annexe 1 – Note de Mission 109

Annexe 2 – Schema « enjeux du thesaurus commun » 112

Annexe 3 – Vue sur les fonctionnalités thesaurus d'adlib for windows 114
Recherche incluant les termes associés 122

Annexe 4 – Evaluation des collections films et non films des partenaires 132

TABLE DES ILLUSTRATIONS

Figure 1 - Exemple d'ontologie représentant un véhicule _____ *18*
Figure 2 - Schéma d'un concept OTAREN en 2010-2011 _____ *32*
Figure 3 - Le premier "layer cake" du web sémantique _____ *33*
Figure 4 - Schéma de l'ontologie FOAF _____ *37*
Figure 5 - Représentation d'un concept SKOS (W3C) _____ *40*
Figure 6 - Schéma simplifié du modèle de données ISO 25964-1:2011 _____ *42*
Figure 7 - Proportion des entretiens menés dans les institutions partenaires _____ *77*

Tableau 1 - Exemple de relations dans la norme ISO 25964-1 _____ *44*
Tableau 2 - Informations données par l'étude de faisabilité sur la construction du thesaurus _____ *72*
Tableau 3 - Lignes directrices sur l'orientation du livrable _____ *87*
Tableau 4 - Test des fonctions de gestion et maintenance de thesaurus _____ *95*



INTRODUCTION

La mission décrite dans ce mémoire s'est déroulée entre le 1er juin et le 23 septembre 2016, entre Bois d'Arcy et Bercy. Elève ingénieur documentaliste mais néanmoins cinéphile, je fus chargé par un heureux concours de circonstances d'étudier la faisabilité d'un projet de thesaurus commun porté par la direction du patrimoine du Centre national du Cinéma et de l'image animée (*CNC*) et la *Cinémathèque Française*. Ma mission était d'évaluer les possibilités d'un thesaurus transversal, langage documentaire d'indexation et d'interrogation globale des collections, qui intégrerait une plateforme de valorisation commune aux deux institutions patrimoniales, fruit de la fusion des deux portails actuels. Elle allait me mener vers des territoires au sein desquels un thesaurus peut ne plus être une œuvre indépendante et farouchement spécialisée, mais une œuvre composite faite de relations entre différents référentiels d'indexation.

D'outil « réservé » au documentaliste, le thesaurus a du évoluer lorsqu'il lui a fallu intégrer les logiciels documentaires. Il se frotte depuis dix ans à des problématiques plus larges liées à l'interopérabilité et au web sémantique. Cette ouverture le met en concurrence avec d'autres langages et techniques d'indexation, qui subissent également l'évolution des instruments de recherche d'information ou, dans une plus faible mesure, qui sont directement issus de cette évolution. De ce fait, l'ingénierie d'un nouveau thesaurus doit désormais s'accompagner d'études plus appuyées sur ses enjeux et ses possibilités de développement. Ses enjeux, car ils permettent de mettre en perspective le thesaurus au sein de l'organisation qui l'a créé, mais aussi dans un tout global désormais incontournable. Ses possibilités de développement, car un thesaurus est un investissement coûteux qui doit être rentabilisé et dont les possibilités de maintenance, d'évolution et d'intégration à un outil de gestion doivent être étudiées sur le long terme. La création de ce langage documentaire complexe devient dès lors plus intéressante du point de vue du chef de projet, qui n'est pas seulement amené à coordonner le groupe qui élaborera ce thesaurus. Des phases comme l'étude de faisabilité et l'analyse de besoins, qui étaient depuis longtemps préconisées mais réduites à portion congrue deviennent inévitables.

Ce mémoire ne saurait donc se borner à établir un guide pratique de la fusion, ou par ailleurs de l'alignement, de plusieurs vocabulaires[1]. Il ne saurait pas plus

[1] On emploiera le terme vocabulaire dans son sens générique, comme l'ensemble des termes et concepts appartenant à un langage défini.

apporter un éclairage sur l'ingénierie du thesaurus, d'autant moins que de précédents travaux tels que le mémoire INTD de Lorraine Keller [31] ou le « *guide pratique pour l'élaboration d'un thesaurus documentaire* » de Michèle Hudon [30], plus fortement ancrés dans l'élaboration, réussiront bien mieux à guider le lecteur désireux de connaître la méthode de construction d'un thesaurus. Il a pour but d'évaluer les évolutions du périmètre d'action du thesaurus dans la recherche d'information, ainsi que les conséquences de ces évolutions sur la la méthodologie d'un chef de projet amené à étudier la faisabilité d'un projet de thesaurus en 2016. Face aux nouvelles techniques d'indexation matière, le thesaurus a-t-il encore une raison d'être comme outil de référence pour la recherche documentaire, et si oui, quels rôles peut-il jouer ? Nous consacrerons les deux premières parties de ce mémoire à cette mise en perspective.

Le cas de la fusion des collections films et non-films des institutions cinématographiques patrimoniales françaises est unique, mais il s'insère dans un contexte culturel global de mutualisation. De par la diversité de ses acteurs et des collections à traiter, il constitue un exemple riche de ces nouveaux contextes auxquels les outils classiques de recherche documentaire peuvent être appliqués. Il constituera donc un fil rouge cohérent sur lequel s'appuiera ce mémoire.

Les parties suivantes s'attacheront à mettre en évidence les adaptations d'une étude de faisabilité de thesaurus dans un cadre d'interopérabilité. Dans un premier temps, il développe la posture peu banale d'un chef de projet au cœur d'un mécanisme complexe qui impliquera de plus en plus souvent des équipes informatiques et des intervenants extérieurs. Mais aussi la transversalité inhérente à des projets toujours plus nombreux d'accès à des collections diverses. Ce mémoire est destiné à donner des clés d'action à un chef de projet confronté à la faisabilité d'un rapprochement de langages documentaires lié à des collections différentes, connaître les outils et donner une idée des questions à se poser lorsqu'une recherche fédérée sur celles-ci est nécessaire. L'exemple des adaptations sur les méthodes établies pour les films et non-films pourraient les guider, de même qu'un regard critique sur mon expérience pourrait les détourner de certaines erreurs.

Enfin, il ne saurait être question de conclure sans décrire les préconisations livrées pour ce cas, différents scénarii conçus pour guider la décision des partenaires du futur portail des institutions cinématographiques patrimoniales françaises.

PREMIERE PARTIE : ROLES DU THESAURUS DANS L'ACCES A L'INFORMATION

1. Un outil de recherche d'information

1.1 Fonctions et caractéristiques d'un thesaurus

1.1.1. Du thesaurus dictionnaire au thesaurus documentaire

En grec ancien, Thesaurus signifie « trésor ». Plusieurs formes du terme sont aujourd'hui acceptées dans le vocabulaire courant : Une forme « thésaurus » avec accent utilisée dans la littérature francophone et une autre forme latine « thesaurus » sans accent qui emploie le pluriel « thesauri », adoptée à l'internationale et que nous utiliserons dans ce mémoire.

Le terme thesaurus existait dans la littérature médiévale, mais son premier emploi officiel fut en 1531 pour le *latinae linguae thesaurus* de Robert Estienne, imprimeur de François Ier. Il est alors employé jusqu'au XIXème siècle comme un dictionnaire de mots dans les langues anciennes[2]. De nombreux *thesaurus dictionnaires* apparaîtront jusqu'au XVIII ème siècle.

Précurseur du thesaurus documentaire, l'évêque John Wilkins crée un dictionnaire amélioré qui prend en compte les relations entre les mots dans le but d'organiser les objets du référentiel d'histoire naturelle de la Royal Society. C'est Peter Mark Roget qui donna à l'outil documentaire son acception contemporaine en 1852, à travers le *« Roget's Thesaurus of English words and phrases »*, réunion de concepts hiérarchisés en synonymes et notions connexes. Cependant, on constate que les thesauri ont connu leur essor après la seconde guerre mondiale[3], dans le sillage du développement des sciences de l'information et du développement des techniques informatiques. Leur stabilisation prend en compte de nouveaux besoins de limitation du nombre de concepts et d'harmonisation des vocabulaires des auteurs, des indexeurs et des utilisateurs. [1, MENON].

1.1.2. Caractéristiques du thesaurus

Dans leur thesauroglossaire, Danièle Degez et Dominique Ménillet définissent le thesaurus comme une « ***Liste organisée*** *de termes* ***contrôlés*** *et* ***normalisés*** *(descripteurs et non-descripteurs) servant à* ***l'indexation des documents et des questions dans un système documentaire****. Les descripteurs sont reliés par des*

[2] Le *Dictionnaire françois latin* de 1539 du même auteur est le premier lexique à entrées en français désignant les équivalents latins.

[3] Le premier thésaurus opérationnel est celui des termes de la chimie en 1959 *Thesaurus of Engineering Terms.*

relations sémantiques *(génériques, associatives et d'équivalence), exprimés par des signes ou symboles conventionnels. Les synonymes (non-descripteurs ou termes interdits) sont reliés aux descripteurs par une seule relation d'équivalence* » [2, DEGEZ, MENILLET].

Très complète, cette définition établit le contexte dans lequel l'outil est utilisé et elle comporte l'ensemble des caractéristiques qui différencient le thesaurus d'autres langages documentaires (listes d'autorités, taxonomies, classifications…). Nous allons dans un premier temps explorer plus en détail ces caractéristiques.

1.1.2.1 Un langage pour indexer et rechercher les documents

La première fonction du thesaurus est de donner accès à des collections. Il permet de rechercher de l'information correspondant au sujet des documents au sein de ces collections. Le thesaurus intrique donc deux opérations qui correspondent à deux niveaux du traitement documentaire.

- L'indexation, qui consiste à traduire le contenu du texte. Elle est effectuée en général par des documentalistes, qui se présentent comme des médiateurs entre l'auteur/le contenu des textes et les utilisateurs. Ils devront donc anticiper les besoins et les demandes dans la constitution de l'indexat[4]. [3, PIERRE]

- La recherche, qui consiste à la formulation de questions par un utilisateur. Celui-ci est confronté à la masse des documents et il doit trouver des réponses à sa question. Le thesaurus permet de faire le tri en sélectionnant les documents pertinents, du fait qu'ils correspondent aux mots clés sélectionnés, à des termes synonymes ou à des termes connexes.

Exemple : Le thesaurus *Cinédoc* est employé par la cinémathèque française pour indexer des collections liées au cinéma. Il est spécialisé dans le domaine du cinéma. Un utilisateur du thesaurus *Cinédoc* qui cherche des informations sur « les requins au cinéma » pourra retrouver un article sur le sujet du fait que le descripteur *« requin (thème) »* du thesaurus aura été associé à l'article par le documentaliste lors de l'indexation du périodique.

Lorsqu'un organisme décide d'opter pour un thesaurus pour organiser ses documents, il doit donc se demander de prime abord : Quel type de collections le thesaurus va t'il indexer ? Quels seront les utilisateurs de ce thesaurus ?

[4] Ensemble des termes ou indices résultant de l'indexation d'un document.

1.1.2.2 Un langage contrôlé et normalisé

Les termes et concepts utilisés pour décrire les documents au sein du thesaurus sont soumis à des normes, formalisation de bonnes pratiques existantes au moment de leur élaboration. Tous les organismes s'appuyant sur ce langage indexent alors les documents de la même façon. Voici ces principaux cadres normatifs :

- AFNOR Z 47-100 (1981) : Etablissement d'un thésaurus monolingue
- ISO 2788 (1986) : Principes directeurs pour l'établissement et le développement du thesaurus monolingue
- AFNOR 247-101 (1990) : Thesaurus multilingue
- ISO 25964-1 (2011) : Pour la recherche documentaire
- ISO 25964-2 (2013) : Thesaurus et interopérabilité avec d'autres vocabulaires

Les normes en vigueur imposent entre autres des formes préférentielles pour l'emploi des descripteurs (l'emploi de substantifs, du singulier, l'exclusion des mots vides) et elles encouragent l'emploi de notes d'application pour l'indexation. Le Thesaurus Cinédoc de la Cinémathèque française a été *« réalisé en conformité avec la norme AFNOR Z-47-100 ».*

Par opposition à des mots clés libres, le thesaurus a pour fonction d'unifier les pratiques d'indexation au sein d'une ou plusieurs structures. C'est un référentiel qui préexiste à l'indexation des documents. Afin de vérifier l'opportunité d'utiliser un thesaurus pour une ou plusieurs structures, il faudra évaluer l'intérêt de contrôler le langage dans la ou les structures.

1.1.2.3 Un langage structuré et ordonné

Les concepts et termes sont organisés entre eux en fonction de relations sémantiques afin d'éclairer leur sens lors de la recherche.

- Les relations d'équivalence permettent de relier plusieurs termes ayant des relations de synonymie et de polyonymie[5] correspondant à différents niveaux de langage d'une même langue, de relier des termes de plusieurs langues différentes renvoyant à un même concept, ou bien des sigles à leur équivalent développé. La relation d'équivalence est généralement formalisée par les sigles EP (employé pour) et EM (employé).

[5] La polyonymie désigne l'emploi de plusieurs termes pour exprimer un seul concept.

- Les relations hiérarchiques décrivent une dépendance entre deux termes. Le terme générique (noté TG) décrit un concept plus global qui incorpore un concept spécifique qui précise la notion (noté TS).

  Ex : TG = Analyse / TS = Analyse textuelle et Analyse filmique

  Un terme générique ne l'est pas dans l'absolu mais il est qualifié comme tel en fonction de son environnement sémantique. Un générique peut avoir plusieurs spécifiques mais un spécifique ne peut pas avoir plusieurs génériques. Ce fait est de plus en plus contourné la polyhiérarchie[6].

- Les relations d'association (notés TA) permettent de relier un descripteur à un autre, non-synonyme, mais qui peut être intéressant dans le cadre d'une même recherche. Il s'agit d'une relation de renvoi qui peut aussi se matérialiser par *voir aussi*.

Ces relations s'entrecroisent dans un contexte plus général qui vise à organiser le thesaurus en sous-thesauri afin d'accéder plus facilement à certains types de concepts. Ainsi un thesaurus peut-être organisé en facettes ou en sous-thèmes. Le thesaurus *Cinédoc* a choisi cette deuxième option. Il est découpé en neuf champs sémantiques distincts et 4 listes.

1.1.2.5 Un langage combinatoire

La grande particularité du thesaurus est de permettre de combiner les descripteurs afin d'affiner une recherche, ce qui équivaut à la création de nouveaux descripteurs à la recherche. C'est ce qui fait du thesaurus un représentant des langages post-coordonnés, par opposition aux classifications ou aux listes de vedette matière, dont les termes sont combinés a priori. L'avantage de ce système est d'éviter de répertorier trop de concepts, tout en exprimant de nombreux sujets.

La recherche booléenne est particulièrement adaptée à cette caractéristique du thesaurus, ce qui lui a permis de pouvoir très vite évoluer dans un contexte informatisé afin de soutenir les recherches avancées.

Un thesaurus sera donc conseillé pour un système privilégiant les recherches avancées. Il faudra aussi exploiter cette possibilité de combinaison dans le cas d'un trop grand nombre de descripteurs.

[6] Qualité d'un langage documentaire dans lequel chaque terme peut avoir plusieurs termes génériques de niveau immédiatement supérieur. *« Thesauro-glossaire des langages documentaires »*

1.2 Les alternatives au thesaurus

Chaque langage documentaire peut être utile à un contexte particulier. Les avantages et inconvénients de chacun de ces outils doivent être comparés au besoin pour décider de l'opportunité d'employer un thesaurus ou un autre langage.

1.2.1 Thesauri et classifications

« Les classifications sont une conséquence directe du foisonnement de l'édition lors du XIXème siècle » (1, MENON). L'accroissement considérable des collections des bibliothèques conduisit les professionnels à concevoir des systèmes classificatoires rationnels et thématiques pour ranger les ouvrages dans les rayons, et donc repérer facilement des documents connexes. Une classification permet d'attribuer au document un indice alphanumérique qui le situe dans un domaine de connaissance.

La plus répandue, la classification Dewey, a été créée par Melvil Dewey en 1873. En tant que classification décimale, elle comporte dix classes, et chaque classe est elle-même divisée en dix autres classes. De par sa simplicité, son utilisation est particulièrement prisée pour le libre accès aux documents en bibliothèques. La principale critique à cette classification est qu'elle ne peut pas s'appliquer à de nombreuses institutions qui possèdent des collections très spécialisées dans un domaine particulier. En effet, la répartition se fait d'abord par disciplines, puis par sujets. De plus, elle ne prend pas en compte le fait que le document puisse appartenir à plusieurs disciplines (3,WATRELOT).

La classification décimale universelle (CDU) a été créée en 1905 par Paul Otlet et Henri Lafontaine. Elle est composée de neuf classes qui s'inspirent de la classification Dewey, dont une classe vide[7]. La CDU a pour avantage de présenter un système de notation assez simple et flexible qui lui permet de décrire très précisément les documents. Elle propose un indice allongé et d'une syntaxe à ponctuation (séparateurs) permettant de mettre en relation plusieurs concepts, ce que n'autorise pas la Dewey. Ainsi les autres chiffres pourront exprimer les facettes supplémentaires d'un sujet. Cette constitution très précise peut s'avérer utile lors d'une recherche informatisée, notamment pour l'attribution d'identifiant mais elle ne permet de classer physiquement des ouvrages. La CDU connaît depuis quelques années une certaine régression compte tenu de l'irrégularité de ses mises à jour.

[7] Un document est mis à disposition pour aller plus loin sur la CDU.
http://www2.fr.ch/bcu/N/la_BCU/cooperation/Bib_uni/SCANT_Generalites.pdf

Contrairement au thesaurus, une classification est un langage non combinatoire et pré-coordonné. C'est donc un système rigide et figé, dont l'évolution se fait à la marge ou via une refonte complète. La notation décimale permet néanmoins d'étendre et d'aménager la classification et de prendre en compte l'évolution des progrès de la science. Qu'ils soient représentés par des indices ou des catégories, les langages classificatoires bénéficient d'une grande simplicité qui a établi leur attractivité auprès du public, sur le web et au sein des systèmes d'information des entreprises.

Une classification hybride, la classification à facettes, propose d'organiser de manière syntaxique la classification des documents selon les facettes universelles de celui-ci, mais il n'a pas connu de succès dans la pratique. Le modèle fut élargi aux thesauri. La norme en vigueur prévoit qu'une facette indépendante de la hiérarchie puisse devenir une des sous-divisions du thesaurus. [5, MANIEZ].

1.2.2 Thesauri et taxonomies

Le terme taxonomie s'appliquait originellement à tous type de classification, puis il a connu d'autres significations lors de son appropriation par les systèmes d'information, puis par internet.

Une taxonomie se définit désormais comme « un cadre d'organisation pour des ressources numériques de toutes natures (donc pas forcément des documents) destinés à en permettre une présentation ordonnée et y donnant accès par navigation hypertextuelle. »[8]. C'est un langage classificatoire et contrôlé qui introduit une relation de hiérarchie qui permet d'élargir ou de resserrer une recherche à partir d'un terme. Selon l'angle de vue de l'organisme ou du chercheur définissant la notion, la taxonomie peut s'étendre du répertoire de site web[9] au schéma d'organisation arborescent des contenus d'un site sur un mode hypertexte. Les taxonomies s'attachent essentiellement à organiser les métadonnées « sujet » des ressources auxquelles elles se rapportent. Elles peuvent donc être à la base d'un système de catégorisation automatique de document ou bien faciliter la recherche d'un terme en fonction de ses relations hiérarchiques avec d'autres termes.

[8] D'après le vocabulaire de l'ADBS

[9] Le site Dmoz, qui existe depuis 1998, est un bon exemple de répertoire du web https://www.dmoz.org/World/Fran%C3%A7ais/

1.2.3 Thesauri et ontologies

L'ontologie voit son origine dans la philosophie, se définissant comme l'étude de l'être en tant que tel et de ses propriétés intrinsèques.

Les ontologies, au pluriel, sont des outils qui firent leur apparition dans l'intelligence artificielle à la fin des années 80. Les ingénieurs avaient besoin d'un modèle pour les domaines utilisés dans les systèmes de raisonnement. Selon Tom Gruber[10], *« Une ontologie est une spécification partagée d'une conceptualisation »*. On déduit de cette définition la nature formelle de l'ontologie, qui s'assimile à un cadre conceptuel de raisonnement à destination des systèmes informatisés. La modélisation proposée par l'ontologie définira pour la machine une représentation qui lui permettra d'analyser, via des propriétés, attributs et relations, le sens des concepts d'une spécialité. Ainsi l'ontologie obtenue décrit le monde de ce domaine du point de vue du concepteur de l'ontologie, et spécifiquement de sa culture.

Jacques Chaumier décrit les ontologies comme : *« Une ou plusieurs taxinomies ordonnées en classes et sous-classes composées d'instances représentant les individus ou objets ; les types d'attributs ou propriétés qui peuvent être attachés à ces objets ; les types de relations entre les concepts d'une taxinomie ; des axiomes ou des règles d'inférence permettant de définir les propriétés de ces relations »* (5, CHAUMIER).

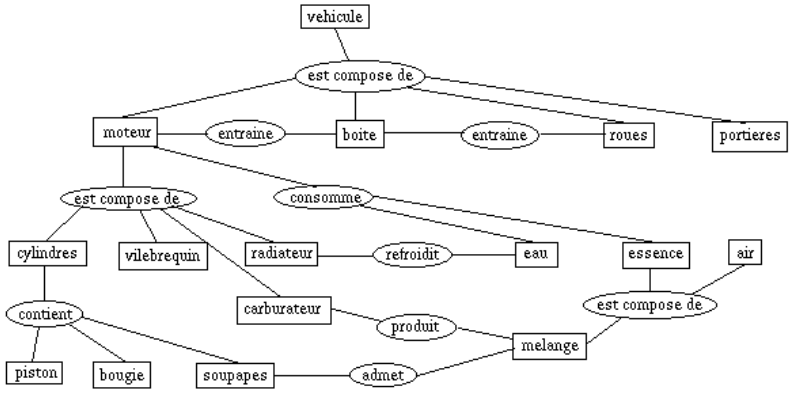


Figure 1 - Exemple d'ontologie représentant un véhicule[11]

[10] Tom Gruber est un informaticien et scientifique américain pionnier de l'intelligence artificielle.
[11] http://liris.cnrs.fr/alain.mille/enseignements/IGC_M2_2008/session8/rapc2/Rapc_Session2_Cas_base_de_cas.html

L'assimilation des ontologies aux langages documentaires classificatoires s'explique par le fait qu'elles servent à normaliser un corpus de termes et à expliciter les liens sémantiques qui les lient.

Les ontologies partagent avec les thesauri des termes et des relations, mais aussi une normalisation et des phases de construction analogues. Comme le souligne Yola Polity[12], « *Dans les deux cas, il s'agit d'un vocabulaire contrôlé, utilisé et validé par les acteurs d'un domaine. Dans les deux cas, ce vocabulaire est structuré et doté de relations sémantiques entre les termes qui le composent. Mais les ressemblances s'arrêtent là car la sémantique des objets et des relations dans une ontologie est une sémantique formelle qui n'est pas destinée à être interprétée par des êtres humains* […]. *Leur caractère formel les rend aptes à alimenter des traitements et des raisonnements menés par des automates*. » Quels que soient ses développements, on ne saurait se servir du thesaurus Mesh[13] pour faire effectuer des raisonnements à une application. Les relations d'un thesaurus sont en effet trop peu développées pour servir de référentiel.

Bruno Bachimont ajoute que, « *les ontologies n'étant pas sans rapport avec les terminologies, on peut trouver dans les thésaurus des ressources pour amorcer une ontologie. Mais il faut prendre garde au fait que ce ne sont que des "ressources pour" et pas des "embryons de"*. » (5,CHAUMIER). Socle du web sémantique, les ontologies sont au final des schémas de données au sein desquels la terminologie d'un langage documentaire, par exemple un thesaurus, peut être entre autres utilisée comme ressource (*voir 3.2.1*).

1.2.4 Thesauri et listes d'autorités (vedette matière)

Les listes de vedettes matière apparaissent au début du XXème, alors que les catalogues imprimés permettent la diffusion des informations bibliographiques et permettent l'émergence des premiers réseaux de bibliothèques. Un répertoire de vedette matière peut se définir comme une *« Suite ordonnée de termes représentant systématiquement le sujet ou l'un des sujets d'un document, plus particulièrement dans un catalogue de bibliothèque »* (2, DEGEZ,MENILLET). Le plus notable en France est RAMEAU, qui appartient à la BNF.

[12] Dans : Gérard Henneron, Rosalba Palermiti, Yolla Polity (dir). *L'organisation des connaissances : approches conceptuelles*. Paris : L'Harmattan, 2005.
[13] Très développé, le Mesh (INSERM) est utilisé pour représenter le domaine médical. http://mesh.inserm.fr/mesh/

RAMEAU est un langage d'indexation pré-coordonné, composé de termes reliés entre eux et d'une syntaxe qui définit les règles de création de vedettes-matières. Il est structuré sur trois niveaux : terminologique, sémantique et syntaxique. Il y a les termes retenus, nommés vedettes-matières, et les termes exclus.

Une liste de vedettes matière est conçue pour cataloguer une collection documentaire en bibliothèque alors qu'un thesaurus est conçu pour indexer un corpus de documents. Son unité de base est la liste, soit un sujet qui renvoie à plusieurs concepts alors que le thesaurus a pour unité le descripteur, un seul concept. Une liste d'autorité sera utile pour normaliser l'accès commun d'une multiplicité de fonds à des sujets particuliers, mais sa gestion est particulièrement lourde.

1.2.5 Thesauri et indexation collaborative

L'indexation collaborative n'est ni contrôlée, ni normalisée. Les pratiques de folksonomies et de *tagging* sont nées d'amateurs du web souhaitant s'approprier l'espace collaboratif du réseau.

Le premier échelon de l'indexation collaborative est l'indexation par l'auteur. Quand un individu ou une organisation met en ligne un site web, il choisit lui-même les descripteurs qui caractérisent le contenu de son site pour le référencer sur les autres sites. Les internautes qui vont consulter ce site pourront aussi le référencer, dans un second temps, avec leurs propres termes. Les folksonomies se présentent sous forme de nuages de mots qui apparaissent de taille variable, les plus gros étant les occurrences les plus fréquentes pour une communauté donnée. Chacun peut décider d'utiliser les mots proposés par d'autres ou d'en incorporer de nouveaux. Les tags peuvent porter entre autres sur contenu de la ressource.

L'intérêt des folksonomies vient de leur adaptation à l'espace du web, que ne possèdent pas les thesauri. *« Indexer un corpus avec un thésaurus est naturel pour un documentaliste. Mais ce n'est pas le cas d'un utilisateur d'Internet qui a pour simple but d'organiser les références et les sites qui l'intéressent. Ce qui forme un atout pour une utilisation professionnelle peut alors apparaître comme un handicap pour une utilisation personnelle »* [7, FRANCIS, QUESNEL]. Il vient aussi du nombre croissant de documents sur le web qui oblige une autogestion en matière d'indexation. Chaque ressource décrite par les internautes économise ainsi du travail à l'indexation manuelle ou automatique. Les folksonomies répondent aussi à un besoin communautaire d'apposer des points de reconnaissance dans l'espace

du web[14]. Elles sont donc un bon moyen pour impliquer des usagers d'une institution dans les contenus de son portail web.

Contrairement aux thesauri, les folksonomies ne sont pas structurées. Elles ne peuvent donc pas éviter l'homonymie et de ce fait, ne facilitent pas l'accès à l'information. En l'absence de normes, la manière d'orthographier les mots interfère aussi dans la qualité de l'indexation. La liberté de choix des termes entraîne la possibilité de se trouver, dans sa liste de mots clés, en présence de termes écrits de différentes manières mais ayant le même sens (exemple : *Open-Source, Open Source* ou *OpenSource*). Enfin, l'individu va décrire la source de manière subjective, en fonction de l'intérêt qu'il lui porte, par conséquent certains sujets seront sciemment ou non évacués. Au final, l'indexation collaboratif parait adaptée au public, mais ne permet pas d'obtenir des résultats de recherche objectivement pertinents, d'où leurs recul dans les organisations au profit des langages contrôlés.

1.3 Thesauri et système de recherche d'information

1.3.1 Thesaurus et besoins du grand public : un mariage difficile ?

« Les thésaurus traditionnels n'ont plus leur place dans les systèmes modernes de recherche d'information ». Telle était la motion discutée lors de la conférence-débat de l'ISKO UK organisée le 19 février 2015[15]. Cette question n'est pas neuve. Les thesauri ont accompagné l'évolution de l'accès à l'information et au document à distance. Elle s'est posée lors du développement des OPAC[16] à la fin des années 80 [8, LE MAREC], puis elle est revenue avec la démocratisation de l'accès au web et le développement des portails documentaires. Sa prégnance s'explique dans le fait que le thesaurus est un outil qui nécessite une médiation.

Pour rappel, un thesaurus est utilisé dans le cadre de deux opérations : l'indexation et la recherche. L'indexation est une forme d'appropriation du langage contrôlé par le documentaliste, qui doit comprendre le territoire du vocabulaire et repérer, puis choisir le descripteur le plus approprié par rapports aux notions sélectionnés dans le document pour élaborer l'indexat. A force d'utilisation, le documentaliste deviendra un expert du thesaurus, d'autant plus qu'il concernera souvent un domaine de

[14] Le hashtag de Twitter est un bon exemple de folksonomie.
[15] Pour un résumé des débats : http://dossierdoc.typepad.com/descripteurs/isko-uk.html
[16] Online Public Access Catalog

spécialité. Lors d'une activité de recherche d'information, le chercheur est dans une posture différente. Il se retrouve avec un besoin d'information qu'il aimerait satisfaire en interrogeant en une seule fois la totalité des fonds aptes à lui donner l'information qu'il recherche. Cet utilisateur final n'indexe pas et il n'a pas construit le thesaurus. Il n'est pas un professionnel de la documentation, et bien souvent il n'est pas un spécialiste du domaine [9, DALBIN]. La médiation du documentaliste, qui maîtrise le langage et le système documentaire interrogé, a permise l'essor du thesaurus et constitué un pont entre les fonctions d'indexation et de recherche. Le développement de l'accès à l'information à distance coupe de cette médiation et enclave les positions respectives de l'indexeur et du chercheur, ne pouvait que remettre en cause l'utilisation d'outils complexes comme le thesaurus.

Cette difficulté s'est vérifiée avec la démocratisation d'internet, déplaçant la question de la recherche de documents au sein d'une collection à la problématique de recherche d'information au sein d'une multitude de ressources. Le 20 septembre 2007, la journée d'étude ADBS *« optimiser l'accès à l'information, une opportunité pour les langages documentaires ? »* fut l'occasion de faire le bilan de la recherche d'information dix ans après l'émancipation du web et de constituer un profil du chercheur grand public. Les habitudes de recherches faisaient alors état d'une *« contamination entre les moteurs de recherche sur internet et les autres systèmes de recherches d'information »* [10, MENON]. Elles se caractérisaient par :

- Un ou deux moteurs de recherche privilégiés.
- Peu de mots dans les requêtes
- Pas d'utilisation des opérateurs booléens.
- La seule consultation des résultats les mieux classés.
- Peu d'utilisation de la recherche multillingue.
- Peu d'aise avec les résultats cartographiés
- Plus d'aisance avec la catégorisation des résultats a posteriori.

Les pratiques se sont depuis établies, renforcés par l'usage de routine des moteurs de recherche commerciaux. Bien que les méthodes de référencement sur ces moteurs aient évolué, l'utilisateur cherche toujours à fuir la complexité, quitte à ne pas obtenir des résultats exhaustifs. Il se satisfera des résultats les mieux classés et

ne se préoccupera pas du taux de précision apporté par le résultat de sa requête. Si un filtre des résultats est opéré, il se fera dans un second temps.

Cet état de fait pénalise fortement les langages documentaires, de même qu'un excès de complexité fut un frein à l'utilisation pertinente des classifications à facettes ou des vedettes matières[17]. Les langages documentaires les plus perfectionnés sont de fait les moins répandus sur le web. Les thesauri ont beau chercher à se rapprocher de l'utilisateur, notamment à travers les équivalences linguistiques qui tendent à faire coller les descripteurs au niveau de langage naturel du public ciblé, ils continuent de pâtir de la complexité de leur réseau sémantique.

Ce constat signifie t'il que le thesaurus est perdu à la cause du grand public et ne peut plus quitter les centres de documentation ? Les thesauri ont-il perdu la bataille du web ?

Pour nuancer, on dira que certains thesauri sont aujourd'hui disponibles via internet et ils utilisent les possibilités hypertextuelles et les graphiques de représentation afin de se rendre plus accessibles. Les raisons sont peut-être à chercher dans l'absence de prise en compte du profil du chercheur plus que dans l'outil.

En 2004, dans une expérience orientée utilisateurs, Jane Greenberg testait 42 étudiants de MBA (Maitrise en administration des affaires), qui avaient été auparavant exposés à des recherches en bibliothèque et dans les systèmes d'information [11, GREENBERG]. Deux méthodes de recherche avaient été proposées par la chercheuse :

- Une *« méthode automatique »* qui permettait d'aligner les requêtes en langage naturel de l'utilisateur, sans qu'il ait connaissance des implications du thesaurus dans le système.

- Une méthode interactive reposant sur un accès au thesaurus et une sélection de termes.

Contre toute attente, 52,4% des participants étaient en faveur de combiner ces deux méthodes, sans doute satisfaits par les résultats de la première, mais désireux d'avoir plus de documents. Seuls 9,5% étaient en faveur d'une forme automatique de recherche dans laquelle ils laissaient toute initiative au système. L'étude concluait sur les bienfaits de l'initiation à l'outil, proposée en préliminaire de

[17] Michel Mingam, RAMEAU, Bilan, perspectives, bbf 2005, vol 50 n°5. P 38-47

l'expérience, ainsi que l'avantage que constituait une proposition de termes à la recherche pour compléter une recherche via ses propres termes puisée dans un thesaurus. De nombreux systèmes d'information ne mettaient pas assez en valeur la recherche par thesaurus dans leur interface, ou bien la nimbaient de jargons de documentalistes sans aucune ressource explicative, ce qui n'incitait pas à avoir recours à cet outil.

La désaffection du grand public pour les thesauri peut également s'expliquer par la quasi-absence de recours à la recherche avancée sur le web. Or, le thesaurus, langage combinatoire par excellence, n'a jamais eu meilleure raison d'être que dans une recherche avancée, par laquelle l'utilisateur trouverait un instrument idéal pour filtrer et préciser sa recherche.

En 2009, la BPI aboutissait aux mêmes conclusions dans l'évaluation de la compréhension de son portail documentaire, son ergonomie et son adéquation aux besoins des usagers [12, BOURGEAUX, FRESNEAU]. Les résultats révélèrent que les utilisateurs étaient déroutés par la multiplicité des chemins d'accès et des possibilités de recherche de l'interface, par le jargon professionnel et que l'absence de visibilité de certaines fonctions appréciées expliquait leur délaissement. Les deux auteures de l'étude militent pour remettre l'utilisateur au centre du système. *« En s'inspirant des logiques de la conception centrée utilisateur, fondées sur la compréhension en amont du contexte d'utilisation et sur l'évaluation régulière des solutions mises en œuvre au regard des exigences prédéfinies et des réactions observées, nos projets d'informatique documentaire gagneront en efficacité, en efficience, en capacité d'attrait et de satisfaction : autrement dit, en utilisabilité. ».*

1.3.2 La recherche libre contre la recherche contrôlée

Le recours au langage naturel et à la recherche plein texte sont à la convergence de plusieurs tendances [9, DALBIN].

- L'insatisfaction des utilisateurs d'information face à la complexité et la multiplicité des interfaces systèmes. Ils souhaitent être informés rapidement et simplement.

- Dans un contexte d'expansion considérable des documents numériques, la notice du document a perdu son statut central. Elle a laissé peu à peu la

place au contenu du document. L'utilisateur veut être informé plus qu'il ne veut avoir accès à une ressource. Il ne veut plus obtenir un lot de références, mais plutôt les informations utiles contenues dans le document.

- Le temps des bases homogènes est désormais révolu. *« La recherche plein texte facilite en apparence un accès simplifié et unifié à des ressources multiples, traitées en des lieux différents et de différentes natures »*.

Les techniques de recherche plein texte sur les notices étaient déjà opérationnelles dans les années 1980, notamment par les serveurs de banque de données professionnelles qui proposaient des opérateurs de proximité entre les termes [13, DALBIN]. L'informatique documentaire des entreprises s'est aussi très vite orientée vers le texte intégral.

Parallèlement se développent la classification et l'indexation automatique des documents en réponse à une indexation humaine jugée longue et coûteuse, et dont les résultats ne sont pas scientifiquement vérifiables. Couplés aux logiciels documentaires, les logiciels de traitement automatique de la langue (TAL) ont apporté de nombreux progrès aux techniques d'indexation libres. Ces logiciels étudient la relation des mots entre eux au sein d'une phrase pour en saisir le sens. Ils permettent de faire une analyse interprétative des données en fonction de l'analyse de la place et du sens des mots. Ils sont très utiles sur des corpus homogènes en termes de vocabulaires et de structures des documents, mais restent limités pour des corpus comme ceux proposés sur le Web. S'ils se perfectionnent désormais d'éléments sémantiques, les logiciels de TAL ne rendent toujours pas compte du sens des concepts, contrairement aux langages documentaires.

Un langage libre ne permet pas non plus de contrôler les problèmes d'homonymie, de synonymies et de polynymie comme un thesaurus sait le faire. *« La recherche libre reporte sur l'usager du système l'essentiel de la charge du travail intellectuel de distinguer les synonymes, homonymes et la polynymie et ce pour chaque recherche, au lieu d'être consigné de manière permanente dans le langage documentaire ou dans les champs d'indexation des documents » [1, MENON]*.

1.3.3 Balisage sémantique et « Microdonnées »

Une autre opportunité se trace dans structuration des documents, qui permettrait d'incorporer les métadonnées aux balises, afin que les sujets des documents

puissent être extraits par les moteurs de recherche. Cette structuration a franchi un pas décisif grâce au balisage sémantique.

Pour être en mesure de proposer des résultats de recherche pertinents, il faut qu'un moteur de recherche puisse lire le contenu de toutes les pages web qu'il a indexées. Le balisage d'un document HTML permettait d'établir la fonction de parties du document. Créés en 2005, les « microformats » visent à structurer les contenus par rapport à des "types" sémantiques (calendrier, revues/critiques, ...) tout en conservant la simplicité de production d'HTML. L'arrivée du HTML 5 en 2014 a ouvert les fonctions des balises HTML en les enrichissant d' « attributs » particuliers**.** Grâce à ces attributs, les robots d'indexation du moteur de recherche ne lisent pas seulement une chaîne de caractère, mais comprennent le sens du mot. Le balisage sémantique leur permet ainsi de proposer des résultats adaptés, mais surtout de privilégier les sites ayant balisé sémantiquement leur contenu[18].

Afin d'organiser ce balisage sémantique, Schema.org a été initié par Google en juin 2011, puis rejoint par Bing, Yahoo ! et le portail russe Yandex. Le site est un corpus d'éléments de métadonnées contenues dans les balises HTML, sélectionnés, structurés et encodés selon un formalisme propriétaire. Des applications permettent aux webmasters des sites de marquer facilement leurs pages avec des microdonnées, ce qui permet d'enrichir progressivement le schéma[19]. En partant d'un film, on peut par exemple identifier des attributs de ce film, ou d'une œuvre plus générale, comprenant son genre et son sujet. Ainsi un tagging thématique et un rassemblement de contenus sur un thème est-il tout à fait possible.

Même si les schémas répondent à des ontologies et à des modèles proches, le balisage sémantique ne se confond pas avec le web de données liées. Cependant, des initiatives de rapprochement des vocabulaires ont eu lieu, encouragées par schema.org et le W3C[20]. Ainsi les méthodes tendent à devenir complémentaires.

1.3.4 Le thesaurus en appui aux systèmes de recherche d'information

Le thesaurus est apparu à la même époque que l'informatique, d'abord comme un dictionnaire de descripteurs, puis comme un réseau de relations. Dès les années 50, il a pu soutenir l'accès thématique à de grandes banques de références

[18] Blog Descripteurs - http://dossierdoc.typepad.com/descripteurs/2011/12/schemaorg.html
[19] Le schéma est disponible en entier sur http://schema.org/docs/full.html
[20] Nogales Alberto et al, A linking from schema.org microdata to the web of linked data : an empirical assessment, Computer standards and interfaces. Volume 45, March 2016, Pages 90–99. < http://www.sciencedirect.com/science/article/pii/S0920548915001440 >

scientifiques qui prenaient appui sur les premiers ordinateurs [13, DALBIN]. A une époque d'intense production documentaire, thesaurus et informatique devenaient incontournables pour rendre accessibles ces ressources via les réseaux de télécommunications. Dans les années 80, l'essor de la microinformatique a établi une impulsion supplémentaire autant au niveau des centres documentaires qu'au niveau des entreprises, soutenant les possibilités de rechercher et sélectionner les descripteurs au sein des thesaurus. L'informatique a contribué à installer le modèle *indexation-thesaurus-recherche* que l'on connaît toujours aujourd'hui.

Mais parallèlement, des réflexions se sont constituées sur la possibilité d'établir des thesauri à la recherche, indépendants du thesaurus à l'indexation moins tournés vers la formulation d'une requête. L'utilisateur se trouve en effet plus à l'aise dans l'exploitation du résultat de la recherche (*voir 1.3.1*). Envisagé comme appui du public, le thesaurus se dépouille cependant de certains de ses attributs et développe certaines déviances, comme une grande tendance à la polyhiérarchie ou une déposession de ses relations hiérarchiques.

Dans le cadre d'une recherche plein texte, Sylvie Dalbin conseille de n'exploiter que les équivalences au sein des thesauri afin de relier les descripteurs de l'indexat des notices aux non-descripteurs du thesaurus [9, DALBIN]. Il serait ainsi possible au chercheur d'établir une requête sur les champs de type « descripteurs » basés sur le langage contrôlé ainsi que sur les champs de type « titre » et « résumé » contenant du langage naturel. Cette solution peut se voir comme un compromis entre la simplicité du plein texte et les avantages du langage contrôlé. Elle peut également tirer une force de la sur-pondération d'un mot clé[21] en ce qu'il qualifie mieux conceptuellement un texte qu'un résumé ou un titre.

Mais cette simplification se fera au détriment des fonctions hiérarchiques du thesaurus, qui permettent d'éviter le silence dans une recherche. Ceci est particulièrement vrai pour les thesauri géographiques qui permettent de mettre en perspective le sujet d'un article dans un contexte géographique. Prenant pour exemple le cinéma, un utilisateur qui souhaiterait trouver tous les documents d'un fonds traitant de la nouvelle vague britannique (courant influencé par le documentaire prenant place dans les années 60, et désigné rétrospectivement par référence à la nouvelle vague française), il se verra afficher les notices répondant aux descripteurs « nouvelle vague » et « Angleterre ». Or, la notion est moins

[21] Technique qui consiste à attribuer une prépondérance au descripteur du thesaurus

consacrée que la nouvelle vague française, donc il se peut qu'aucun document ne ressorte de cette recherche. Avec une extension hiérarchique de la recherche, le système aurait pu prendre en compte également le terme *« Europe »* dans une logique expansion automatique ascendante. Il aurait pu ainsi sélectionner des documents qui parlaient de la nouvelle vague française en Europe, comprenant la nouvelle vague britannique.

En 2007, la spécialiste soulignait que dans une logique de se libérer de l'indexation et de mieux servir les chercheurs d'informations, le thesaurus devaient se positionner en appui des moteurs de recherche [10, MENON] :

- Au moment de la formulation de la requête, une aide à la sélection de termes à travers les relations d'équivalence ou comme filtres pour choisir ses clés de recherche.

- Au moment du traitement de la requête en enrichissant la requête des synonymes, par recherche fédérée dans le cas d'alignement possible entre plusieurs langages.

Au moment de la présentation des résultats, la structure du thesaurus peut permettre d'offrir des choix pour affiner ou étendre la recherche. Il est également possible d'organiser un lot de résultats en fonction des thèmes, domaines ou facettes d'appartenance des descripteurs. Le tri des documents par facettes sera plus en accord avec les besoins des publics qu'une recherche avancée.

En dépit de sa mauvaise presse, le thesaurus peut donc régulièrement enrichir les recherches documentaires en ligne. Ainsi en 2011, une étude de la *Semantic Web Company* constatait que son utilisation se faisait de plus en plus courante dans les moteurs de recherche d'entreprise[22]. Cette revalorisation tardive de l'outil va de paire avec l'essor du web de données liées qui favorise son emploi comme système d'organisation des connaissances.

[22] Moteurs de recherche : Vers un usage banalisé des thésaurus ? – Blog Antidot <http://blog.antidot.net/2011/06/29/moteurs-de-recherche-vers-un-usage-banalise-des-thesaurus/>

2. Un système d'organisation des connaissances (SOC) pour l'interopérabilité sémantique

En tant que représentant le plus qualifié des langages documentaires, le thesaurus a dû évoluer en phase avec les systèmes d'information. L'appui qu'il a su apporter à la recherche au sein de ces systèmes atteste de son utilité pour désambiguiser les termes et incorporer des liens sémantiques dans une recherche. Nous avons pu constater que la complexité de l'outil a pu décourager les utilisateurs des systèmes d'information et les internautes. Or, cette complexité n'existe pas pour les machines. Bien au contraire, les systèmes de recherche d'information modernes se nourrissent de représentations formelles. Les langages contrôlés ont pu notamment investir le web à travers des schémas de données et les ontologies destinées à communiquer en bonne intelligence en transmettant un référentiel de compréhension des requêtes aux applications. L'essor du web sémantique dans les années 2000 a permis de matérialiser cette base. C'est dans ce contexte de sémantisation du web que se crée une dynamique de normalisation visant à rendre les langages interopérables. Afin de comprendre les raisons de la normalisation récente sur l'interopérabilité des thesauri et des autres langages documentaires, nous allons nous pencher sur les contextes qui lui ont été favorables : les portails et l'essor du web sémantique, puis nous développerons cette notion au cœur de la réforme structurelle du thesaurus.

2.1 Des contextes propices à l'interopérabilité sémantique

2.1.1 Définition et portée de l'interopérabilité

Patrice Landry définit l'interopérabilité comme « la capacité de plusieurs systèmes à communiquer entre eux sans ambiguïté et à échanger de l'information sans difficulté »[23]. La norme ISO 25964-2 :2013 *« Thesauri and interoperability with other vocabularies »*, qui applique l'interopérabilité aux langages documentaires traduit l'interopérabilité comme la « capacité de deux ou plusieurs systèmes ou composants à échanger de l'information et d'utiliser l'information qui a été échangé ».

L'interopérabilité intervient donc lorsqu'on veut unifier les pratiques d'une structure, ou bien lier des systèmes hétérogènes qui semblent organisés de manière anarchique. Elle se distingue de la compatibilité en ce que l'interopérabilité peut

[23] Patrice Landry, Multilinguisme et langages documentaires : le projet MACS en contexte européen, Documentation et Bibliothèques, avril/juin 2006, p 121

s'effectuer entre plus de deux systèmes et elle suppose ensuite d'utiliser cette communication à une fin particulière.

2.1.2 Interopérabilité et recherche fédérée

2.1.2.1 Interrogation de sources de données hétérogènes

Un portail est un site web qui offre un point d'accès unique à de multiples services et ressources documentaires. Le portail documentaire possède une fonction centralisatrice de ressources, comme tout autre portail, mais il tient sa distinction du fait qu'il met l'accent sur la recherche fédérée. *« Une recherche fédérée permet d'incorporer dans les résultats d'une requête un ensemble de bases de données documentaires hétérogènes et de recueillir les résultats de cette requête dans un affichage unique, sur la seule interface du portail. (…) Ce mode de recherche correspond aux pratiques des usagers du web, plus à l'aise avec une seule interface d'interrogation, plutôt que de multiples interrogations »* [14, GIUSTI].

Interroger des sources hétérogènes de données en une interface commune soulève le problème des relations entre les données des sources intégrées. Lorsqu'il existe un thesaurus transversal à toutes les bases, tous les répertoires de ressources pourront être interrogés de façon pertinente. En l'absence de thesaurus commun, il se posera toujours le problème du bon mot clé à employer pour obtenir une recherche exhaustive. Différents répertoires de ressources peuvent utiliser différents langages documentaires (thesaurus ou autres). Ce cas de figure peut également se poser dans le cadre de la fusion de deux portails, comme nous le verrons dans la partie 2 de cette étude avec l'exemple du CNC et de la cinémathèque française. Dans ces hypothèses, un même concept pourra être représenté par des descripteurs hétérogènes selon le langage utilisé. Ainsi on se retrouvera probablement avec un taux de rappel faible[24] ou du silence, du fait que la requête n'aura pas prise en compte l'ensemble des descripteurs pertinents des langages des différentes bases.

Mais que faire lorsqu'il est impossible de lier toutes ces ressources en un thesaurus commun ?

[24] Le taux de rappel mesure de l'efficacité d'un système d'indexation et de recherche à partir du ratio entre le nombre de documents pertinents trouvés lors d'une recherche documentaire et le nombre total de documents pertinents existant dans le système. C'est un indicateur de mesure du silence.

Des ressources peuvent en effet se révéler trop hétérogènes pour admettre un langage documentaire qui soit globalement cohérent pour toutes les décrire. Un thesaurus est bien plus efficace lorsqu'il est adapté aux fonds qu'il décrit et à un public spécifique, et l'accès à une base ne saurait être suffisant pour mutiler un langage documentaire utilisé depuis de nombreuses années. *« Il y a en effet une apparente contradiction entre la nécessité d'élaborer des langages documentaires utilisés par le plus grand nombre possible de centres de documentation pour échanger des références documentaires, et les besoins particuliers de catégories d'utilisateurs spécifiques »* [15, FEYLER]. Il faut alors réfléchir à dissocier le langage d'indexation de l'interface d'interrogation, et ainsi rendre interopérables les différents langages existants. Il est conseillé dans un premier temps d'identifier les vocabulaires utilisés dans l'ensemble des bases de données concernées afin de déterminer s'il est possible d'utiliser un vocabulaire contrôlé commun à toutes, et si cela n'est pas le cas, il faudra établir une table d'équivalence entre les langages [16, RAVET p.89]. Cette table d'équivalence permettra de lier les termes équivalents ou très proches des différents langages afin qu'ils puissent être pris en compte, directement ou indirectement dans la recherche de l'utilisateur, soit en lui proposant ces descripteurs, soit en les incluant automatiquement. Ainsi la recherche fédérée inclura t'elle des documents de toutes les bases.

2.1.2.2 Un exemple d'interopérabilité à la recherche : Le projet OTAREN

Le projet OTAREN est un bon exemple d'application de ce système d'équivalences terminologiques. Créé en 2004, le projet envisageait de développer pour différentes catégories d'utilisateurs une interface de recherche permettant d'accéder harmonieusement au thesaurus MOTBIS et le répertoire RAMEAU, tous deux utilisés pour des recherches sur l'éducation. L'outil prototype OTAREN proposait d'établir dans un premier temps des équivalences entre les descripteurs de ces deux langages documentaires.

En 2007, le projet s'agrandit des langages documentaires anglophones spécialisés dans l'éducation : Les thesaurus multilingues de l'ONU, de l'UNESCO, le GEMET (thesaurus multilingue environnemental) et le TEE (thesaurus européen de l'éducation). On remarque qu'en se plaçant dans le cadre anglophone, l'interopérabilité par équivalences linguistiques permet de dépasser la division du thesaurus monolingue et multilingue proposée dans la norme Z 47-100 pour interroger sur des bases linguistiques différentes. Le système d'équivalences évitait de passer par une phase de modification de l'indexation initiale et semblait de ce fait

une solution plus pertinente et efficace. Nous nous retrouvons donc dans une dissociation de l'indexation et de l'interface de recherche. Ces interfaces sont néanmoins cohérentes avec les langages d'indexation préexistants qui demeurent les véritables outils d'indexation propres à chaque répertoire de ressources [15, FEYLER]. Le projet OTAREN tirait son intérêt du fait qu'il suivait l'évolution des standards de l'interopérabilité établis par le W3C[25] et les normes en gestation dans l'ISO. Ainsi proposait-il déjà en 2010 la notion de *mapping* (alignement) entre les concepts qui sera au centre de la norme ISO 25964-2.

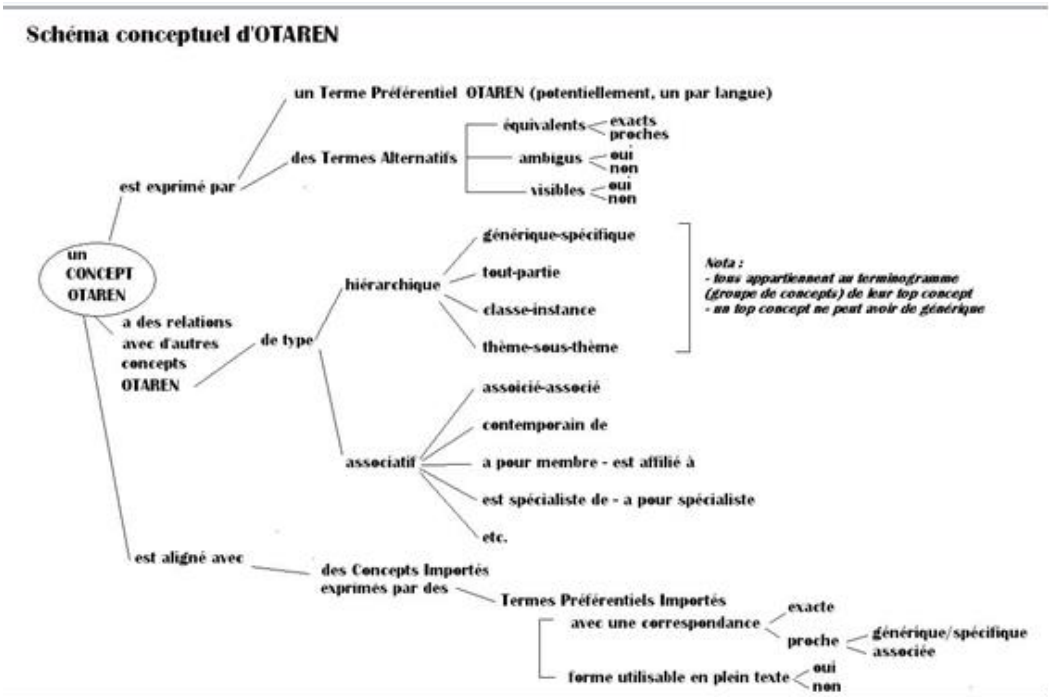


Figure 2 - Schéma d'un concept OTAREN en 2010-2011

2.1.3 L'essor du web de données liées

2.1.3.1 Origines et applications du web de données liées

En septembre 1994, lors de la première conférence www au CERN, Tim Berners Lee définit les nouvelles directions du W3C[26] axées sur le besoin de sémantique pour le web. Il explique que les machines devraient relier automatiquement les données du web aux choses du monde réel.

[25] World Wide Web Consortium. Il s'agit de l'organisme qui établit les standards du web. https://www.w3.org/

Dans une feuille de route détaillée en 1998, il présente le web sémantique comme *« une base de données globale à l'échelle du réseau à laquelle les machines applications de recherche auraient accès afin de mieux appréhender les données ».* Ainsi elles pourraient mieux interagir avec les personnes. Afin de démontrer que son idée n'était pas que théorique, il présenta une première mouture du layer cake, briques technologiques nécessaires à construire le web sémantique, qui sont encore utilisées aujourd'hui.

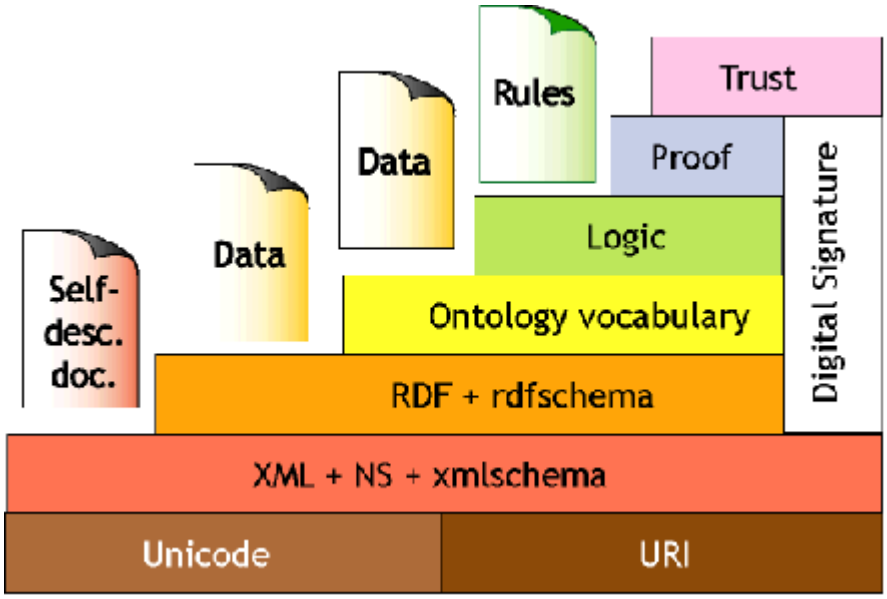


Figure 3 - Le premier "layer cake" du web sémantique

Le W3C entama alors un processus de normalisation pour le web sémantique, mais un climat de défiance vis-à-vis des intelligences artificielles fit traîner les opérations. C'est au milieu des années 2000 que la conjoncture se trouva la plus propice au développement du projet. Tim Berners Lee détruisit l'ambiguité qui subsistait dans le terme *« sémantique »* en parlant du *« linked data »*[27]. La vocation de relier les données entre elles afin qu'une machine ou un humain puisse explorer le web est dès lors plus explicite.

« Si une entreprise met en oeuvre le Web sémantique, toute personne pourra accéder aux informations que cette entreprise a stockées sur ses produits et aussi aux informations stockées par d'autres entreprises sur les mêmes produits. Si quelqu'un cherche des photos sur un sujet et qu'il a besoin de récupérer le nom du photographe, les droits à payer, la définition de l'image etc., il accédera en une

[27] Nous utiliserons la traduction la plus commune de linked data, web de données liées.

seule recherche aux photos et à ces informations, alors qu'avec le Web actuel il doit les chercher successivement dans plusieurs sites d'images. »[28]

C'est DBpedia, créé en 2007 par deux universités allemandes, qui constitue la première réalisation et la première réussite du web sémantique. Le projet met à disposition les données structurées extraites de Wikipedia selon les standards établis par le W3C. Les chercheurs de DBpedia récupèrent les informations présentes dans les « infobox » de Wikipedia. Ces infobox rassemblent, sur le côté droit de l'article, un certain nombre d'informations de manière à peu près normalisée. Les chercheurs ont extrait ces informations, les ont converti au format du web sémantique (RDF) selon *« une ontologie mise au point pour chaque type d'objet »* et les ont intégrées à un entrepôt RDF[29].

L'intérêt de cette technologie est de pouvoir faire appel aux données de Wikipedia depuis un autre site. Les données liées de Wikipédia ont pu ensuite être utilisées par de nombreuses applications, dont la plus fameuse est le *knowledge graph* de Google. Activé en France en 2012, il concrétise pour le grand public une pratique d'utilisation des données de l'encyclopédie Wikipedia via DBpedia, à laquelle il ajoute d'autres sources d'entités et de relations. *« Le Knowledge Graph permet de chercher et d'obtenir instantanément des informations pertinentes sur des lieux, des bâtiments, des objets, ou des personnes »*[30]*.* Il répond ainsi à une demande d'information liée à une requête formulée sur le moteur de recherche Google en exploitant les données du web. Le but avoué est d'anticiper les questions de l'utilisateur du moteur de recherche.

Le web de données liées permet donc de émerger du sens à travers la structuration des données et la description de la logique qui les relie, afin qu'une machine puisse décoder ce sens et l'utiliser à des fins définies. Il s'agit d'un manuel universel donné à une application ou un moteur de recherche afin qu'il puisse lier entre elles les données disséminées sur le web, pour ensuite utiliser cette compréhension. Le web

---

[28] [27] BERNERS-LEE Tim. Le web va changer de dimension. Propos recueillis par Marie-Laure Théodule [en ligne]. La Recherche, Novembre 2007, n°413 [Consulté le 16 septembre 2016].< http://www.larecherche.fr/2-tim-berners-lee-%C2%AB-le-web-va-changerde-dimension-%C2%BB >

[29] DBPedia ou la puissance du RDF au profit du savoir, Got. Les petites cases. Publié le 11 février 2007, consulté le 10 octobre 2016. http://www.lespetitescases.net/dbpedia-ou-la-puissance-du-rdf-au-profit-du-savoir

[30] https://france.googleblog.com/2012/12/lancement-en-france-du-knowledge-graph.html

of linked datas nécessite de rendre interopérables les données entre elles à travers un certain nombre de procédures unifiées et normalisés.

2.1.3.2 Les outils du web de données liées

Les fondations du web sémantique s'appuient toutes sur l'architecture du web. Il ne s'agit pas de réinventer le système, mais d'utiliser l'environnement existant pour développer un nouvel espace [18, Ministère de la culture].

- **Une URI pour décrire un concept**

L'URI (uniform ressource identifier) est la base d'identification d'un élément au sein du web, mais aussi un chemin d'accès[31]. C'est un système d'identifiant à la syntaxe normalisé. Une URI à la particularité d'être unique et pérenne.

Une aura une ou plusieurs représentations en fonction du format que peut traiter ou lire le client qui exécute la requête. Dans un cas d'utilisation d'un navigateur (client humain), c'est le format HTML qui guide la représentation de l'URI. C'est cette page HTML qui rend l'information lisible par l'homme.

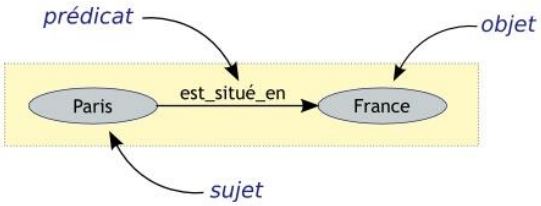
La ressource est l'entité conceptuelle identifiée par l'URI. Ainsi dans les théories de la communication, l'URI peut être vue comme le signifiant, la matérialisation de ce concept du point de vue de la machine. C'est elle qui donne son unicité au concept.

- **Une grammaire : Le modèle RDF**

Un modèle est une sorte de grammaire intégrée. Le modèle RDF (*ressource description framework*) est la grammaire de base du web sémantique. RDF permet d'encoder des données de relations entre les entités afin qu'une machine puisse traiter et analyser ces données. De ce fait, c'est une application particulière du modèle entité-relation. Les relations présentées dans ce modèle ont la particularité d'être à la fois orientées et typées.

RDF va s'appuyer sur les URI pour identifier les entités, qui seront alors assimilées à des ressources. Une relation de deux entités (Paris et France) et une relation (est situé en) représentera ce qu'on appelle un triplet RDF (sujet-objet-prédicat). La machine concevra qu'une URI = Une ressource.

[31] Une URL est une forme d'URI



Ainsi une URI sur le web renverra à une entité du monde réel, en l'occurrence Paris.

Pour pleinement saisir le signifié de l'URI (signe), la machine devra s'appuyer sur une représentation à laquelle lier ce signe. C'est le contexte crée par l'encodage de la relation sujet-predicat-objet qui permet de donner du sens à cette URI-sujet. Ce sens peut être multiple si l'URI-sujet est liée à plusieurs objets par plusieurs prédicats. Ainsi, sur la page Paris de Dbpedia[32].

- Paris peut être situé en France
- Paris peut être une capitale
- Paris peut accueillir une population de parisiens.

L'objet du triplet peut être représenté par une chaîne de caractère, une date ou un entier tel un signe peut être associé à un sens. Dans ce cas, l'objet n'a pas la forme d'une adresse URI mais de la chaîne de caractère le représentant dans le langage courant. http://fr.dbpedia.org/page/Paris devient « Paris » avec la relation « a pour nom ». De ce fait, la machine sait lier la chaîne de caractère à cet URI.

Chaque sujet ou objet d'un triplet RDF peut à son tour devenir le sujet ou l'objet d'une autre ressource, ce qui mène à la formation de l'unité supérieure du triplet : le graphe RDF. L'ensemble des triplets reliera ainsi un concept à plusieurs unités de sens qui seront intelligibles par les machines et définiront peu à peu ce concept.

- **Un vocabulaire : l'ontologie OWL/RDFS**

Là où le modèle est une grammaire, les ontologies seront utilisées comme un vocabulaire, une sorte de dictionnaire pour enrichir le modèle. Le W3C emploie les vocabulaires RDFS et OWL. Comme nous l'avons déjà vu, les ontologies aidé la recherche en intelligence artificielle afin de faire assimiler aux machines des domaines d'apprentissage.

Le diagramme ci-dessous représente l'ontologie FOAF, qui a pour domaine les personnes et les relations qu'elles entretiennent entre elles.

[32] Disponible à l'adresse http://fr.dbpedia.org/page/Paris

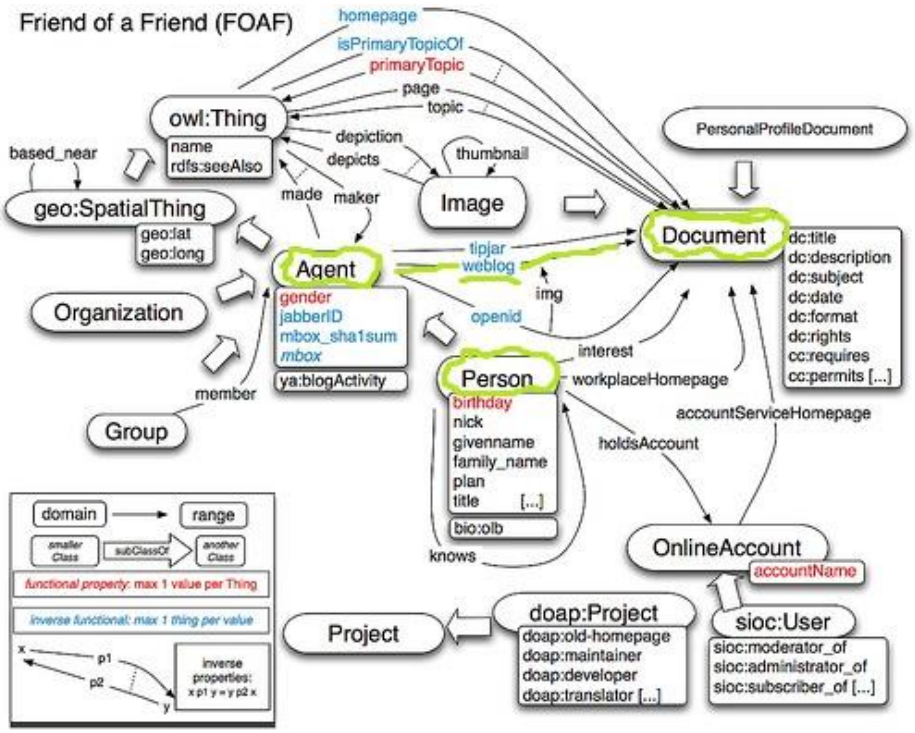


Figure 4 - Schéma de l'ontologie FOAF

Une ontologie décrit :

- **Des classes** qui sont des types d'entités de l'ontologie. Elles sont représentées en vert sur le schéma, **Agent**, **Person** ou **Document**. Chacune de ces classes sera traduite par un URI.

- **Des propriétés** qui se caractérisent par un domaine (le sujet de la relation) et un co-domaine (l'objet de la relation). Sur le schéma, **la propriété page Web relative à un agent (weblog)** a pour domaine la classe agent et pour co-domaine la classe document.

- **Une logique**, matérialisée dans l'ensemble des règles associés à une classe ou une propriété. Par exemple, la définition d'une classe supérieure, dont la propriété s'appliquera à toutes les autres classes.

- **Des instances**, individus d'une ou plusieurs classes.

- **Une syntaxe : RDF/XML, N-triples…**

Enfin, le modèle RDF nécessite d'un élément de langage permettant de retranscrire les triplets RDF, un code permettant la communication des triplets entre humains et

machines. Ce code a plusieurs formalisations, dont l'une, RDF/XML s'inspire du langage XML.

Dans la représentation RDF[33] ci-dessous, nous pouvons lire que Brad Pitt, qui est un acteur américain, a pour fils aîné Maddox Jolie Pitt. Nous remarquons aussi que certaines entités viennent de l'ontologie FOAF.



Exemples de syntaxe N-Triples

Il sera ensuite possible d'interroger le RDF à travers le langage de requête SPARQL[34]. Le résultat reçu sera exprimé en flux XML. Cette possibilité peut également être exploitée par les systèmes de recherche d'information, le but étant la possibilité d'interrogation de données structurées.

2.1.3.3 Les principes du web sémantique

Tim Berners Lee a édicté quatre grands principes au web de données liées [18, Ministère de la culture] :

- utiliser des adresses URI unique pour identifier les ressources ;
- utiliser des adresses URI HTTP qui existent sur le Web. Une Erreur HTTP 404 indique que l'URI utilisé est peu fiable et ne doit pas être réutilisé pour décrire d'autres données ;

[33] Nikolaos Konstantinou, Dimitrios Emmanuel Spannos, Technical Background. Slideshare. Slide 15 of 120.
[34] Langage établi par le W3C.

- fournir à travers l'URI des renseignements lisibles par les humains et par les machines. En utilisant le mécanisme de redirection HTTP (code 302) et la variable User-Agent contenue dans les en-têtes des requêtes HTTP, un serveur peut afficher une page en XML ou RDF pour une machine ou une page HTML pour le navigateur d'une personne ;

- ajouter des URI externes aux données pour améliorer la découverte d'autres informations sur le Web.

Ces principes indiquent sans ambigüités que le web de données liées ne compte pas créer un autre web, mais une extension du web de documents existant à partir du protocole HTTP.

Pour rendre les données interopérables, le W3C a donc dû mettre en place des standards et protocoles communs que suivront désormais toute institution qui souhaitera lier ses données à d'autres jeux de données sur le web. Les langages documentaires, et particulièrement les thesauri ont un rôle à jouer dans cette interopérabilité en tant que systèmes d'organisation des connaissances, ou en anglais Knowledge Organisation system (KOS).

2.2 La normalisation, vecteur de l'interopérabilité des langages documentaires

2.2.1 SKOS, un schéma pour l'interopérabilité sémantique

Les thesauri sont de véritables réserves de vocabulaire spécialisés. Ils peuvent donc se révéler très utiles pour appréhender les différentes terminologies liées à une spécialité. Cependant, ils possèdent une consistance structurelle trop basique pour permettre de lier tous ces termes et concepts et transmettre la compréhension humaine du monde à un logiciel de recherche d'information. Un autre problème réside dans le fait que plusieurs thesauri reliés entre eux peuvent avoir des termes et relations qui auront des sens différents d'un thesaurus à l'autre.

Afin de faire du thesaurus un référentiel sémantique, Il sera donc nécessaire de le modéliser, de préciser les relations entre les termes et d'offrir un cadre d'interopérabilité pour relier les termes des thesauri entre eux. Le cadre offert par les ontologies paraît particulièrement adapté à ce but. Les ontologies sont spécialement faites pour exprimer les règles et relations d'un domaine spécifique.

Elles peuvent également contenir des relations suffisamment explicites pour être compréhensible des machines. Initié en 2003, puis repris comme standard du W3C le 18 août 2009, SKOS (simple knowledge organization system) est une ontologie sur le modèle de OWL (voir 3.1.3.2) qui organise l'interopérabilité des langages documentaires. Les systèmes d'organisation de connaissances (ou langages documentaires) dans le périmètre de cette ontologie comprennent les thesaurus, taxonomies, classifications et listes d'autorités. Sa syntaxe est donc compatible avec l'ensemble de ces langages documentaires.

Le formalisme de SKOS repose également sur les graphes RDF [19, LENART] :

- Un concept au centre du graphe, qui peut contenir des propriétés : indications sur le concept, termes préférentiels ou alternatifs, équivalents dans d'autres langues, termes cachés pratiques pour les variantes avec fautes d'orthographes, les notes, l'image.
- Des relations sémantiques plus fines
- Un autre concept objet.

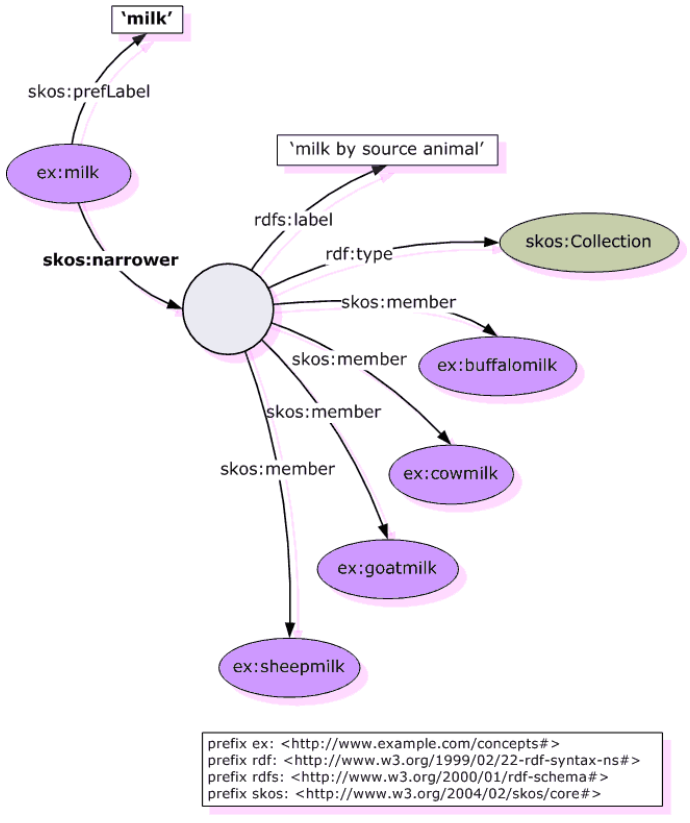


Figure 5 - Représentation d'un concept SKOS (W3C)

Des classes et des propriétés de base forment le noyau de SKOS (Skos Core) afin de définir un fondamental commun aux différents langages et aux standards du web sémantique. Il est ainsi possible de représenter plus finement les relations, par exemple pour rendre interopérables deux thesauri, ou bien de préciser les attributs d'un concept. L'utilisateur peut également choisir de n'exploiter qu'un sous-ensemble de fonctionnalités suivant ses besoins. On ne se trouve dès lors pas nécessairement confrontés aux coûts de maintenance que soulèvent les normes plus complexes du web sémantique. Cependant, le fait que les propriétés soient optionnelles fait peser sur le seul concepteur le maintien de l'intégrité de son application [20, BOYDENS].

Un nombre grandissant de langages contrôlés ont déjà été transposés au Web par le biais de SKOS. En 2015, le thesaurus AGROVOC qui a été converti en SKOS en 2009 [21, SOERGEL] est en relation avec pas moins de 16 langages, parmi lesquels RAMEAU, DBpédia, le GEMET et le STW[35].

Ces liens entre concepts permettront par exemple d'envisager des extensions de recherches depuis une description de fonds concernant l'économie (via le STW) vers des fonds concernant l'environnement (via le thesaurus GEMET). Dans une autre idée, les concepteurs du thesaurus W décrivant les archives ont pu lier les données de leur thesaurus aux données de DBpedia afin qu'il soit utilisable par des applications utilisant les données de l'encyclopédie [18, Ministère de la culture].

SKOS s'est très vite imposé comme un standard d'échange incontournable, d'autant plus qu'il a été reconnu par l'organisme de normalisation ISO dans le cadre de la nouvelle norme sur l'interopérabilité des thesauri.

2.2.2 ISO 25964-1 : Rendre les thesauri interopérables

La norme ISO 25964 représente un tournant dans l'histoire des thesauri car ce langage documentaire n'avait pas été normalisé depuis 1990 et la norme AFNOR 247-101 sur le thésaurus multilingue. Cette évolution trouve son origine dans l'essor du web sémantique, dans l'adoption du format SKOS, mais surtout dans une adaptation nécessaire aux sciences de l'information qui rendait les anciennes normes archaïques.

[35] La liste de ces thesauri est disponible sur le site d'AGROVOC.
http://aims.fao.org/fr/agrovoc/linked-open-data

En 2005, Le *National Information Standard Organzation* (NISO), organisme de normalisation américain, adopte la norme ANSI/NISO Z39.19. Dans le même temps, La *British Standard Institution* adopte la norme BS 8723. Ces deux normes étendent le périmètre des normes de gestion de thesaurus aux synonyms rings, aux taxonomies et aux lexiques. Elles normalisent également pour la première fois l'interopérabilité en préconisant d'établir des ponts entre les concepts, les termes et les relations qui appartiennent à deux vocabulaires différents. Toutefois le problème des formats d'échange et des protocoles n'est pas abordé puisqu'ils n'étaient pas encore définis en 2005. C'est sur la base de ces deux normes que le projet de norme internationale ISO 25964 a vu jour. Elle a fait l'objet d'un travail commun mené par deux comités, l'un international et l'autre français, composés de spécialistes des langages documentaires[36].

La première partie de la norme ISO 25964 est publiée en 2011 [22, NISO]. Elle ne connaît pas de traduction française, mais un livre blanc en exposant les grandes lignes est publié par l'AFNOR, sous la plume entre autres de Sylvie Dalbin, Nathalie Yakovleff et Hélène Zysman [23, DALBIN et al.]. Plusieurs constats nouveaux sur la construction du thesaurus font état d'une évolution vers l'interopérabilité.

2.2.2.1 Des supports techniques

Une représentation par schéma UML des principes directeurs de la norme est incluse afin de permettre sa compréhension par les informaticiens.

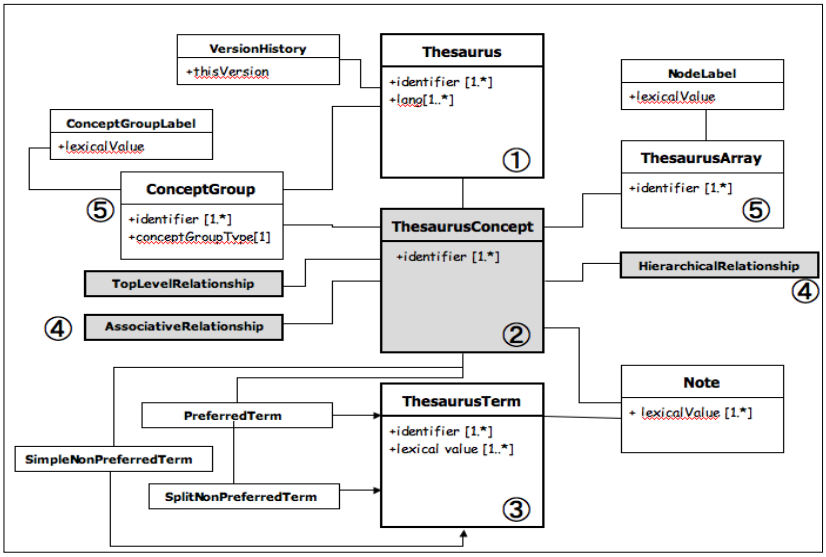


Figure 6 - Schéma simplifié du modèle de données ISO 25964-1:2011

[36] Dans le groupe français figuraient Sylvie Dalbin, Danièle Dégez, Dominique Ménillet et Michel Mingam.

Un schéma au format XML est également inclus, ainsi que trois clauses portant sur l'intégration des thesauri dans les applications, les formats d'échange et les protocoles [22, NISO, articles 16,17 et 18]. Ainsi la nécessité que le thesaurus soit utilisé et compris par les machines autant que par les humains est-elle désormais consacrée.

2.2.2.2 Une distinction claire entre les concepts et les termes

La norme dissocie la sémantique (le concept) et la terminologie (multiples expressions du concept). Cette distinction est importante en ce qu'elle attribue un identifiant unique (l'URI) au seul concept, lui permettant une identification stable dans le temps. Ainsi pourrons-nous toujours trouver sur le web une information sur ce concept. Evolutifs dans le temps, les termes peuvent être des équivalents de langage, mais aussi des équivalents linguistiques. Pratique pour la détermination des identifiants sémantiques, cette distinction entre termes et concepts est peu comprise par les spécialistes. Laurence Maroye met en garde contre les pratiques déterministes dangereuses sur le long terme que peut amener un identifiant immuable attribué à des concepts relatifs et changeants [24, MAROYE].

Ainsi le thesaurus est-il amené à faire évoluer son vocabulaire en fonction du réel, y compris dans l'expression des concepts. Il est difficile de représenter une réalité en perpétuel mouvement dans un langage stable et intelligible. Que faire en cas d'évolution du concept dans le réel ? Le signifié et le signifiant ne forment-ils pas un tout ? Les relations hiérarchiques et relationnelles devront-elles être aussi ajustées en fonction de l'évolution du thesaurus.

2.2.2.3 Le multilinguisme

La norme concerne à la fois les thesauri multilingues et monolingues. A contrario de la France qui avait érigé deux normes distinctes (NF Z47-100 et Z47-101), la norme ISO incorpore les spécificités des thesauri multilingues. Elle propose quatre degrés d'équivalence entre termes de langues différentes : L'équivalence exacte, l'équivalence proche, l'équivalence partielle et La non-équivalence.

Les thesauri multilingues ont alors la particularité de représenter un concept par des termes issus de différentes langues naturelles. Ce postulat est également critiqué car il implique que les langues ont toutes la même structure conceptuelle. Or, les idées peuvent varier d'une langue ou d'une culture à l'autre. « *La relativité linguistique suppose que nos idées, nos conceptions sont conditionnées par la*

langue dans laquelle nous nous exprimons » [24, MAROYE]. La norme inclut le cas de relation non-symétriques, c'est-à-dire de signification différente des concepts entre les langues. Elle préconise alors d'avoir recours à des alignements conformément à la deuxième partie de la norme. Elle réserve cette première partie pour tous les autres cas.

2.2.2.4 La reconnaissance de SKOS

Le format d'échange SKOS, qui porte la même distinction entre terme et concept, est suggéré par la norme comme format d'échange de thesauri (voir 2.2.1).

2.2.2.5 Plus de spécification dans les relations

Les relations établies dans les normes précédentes sont toujours présentes dans cette nouvelle norme. Elles peuvent cependant être typées afin que les machines puissent mieux assimiler les particularités d'un cadre de relation. Ainsi une relation hiérarchique (TG/TS) pourra être spécifiée par différents types de hiérarchies : relation générique, relation partitive, relation d'instance ;

Relation hiérarchique	Générique	Partitive	D'instance
Terme Générique	TGG - Oiseau	TGP - Vaisseau sanguin	TGI - Montagne
Terme Spécifique	TGS - Moineau	TSP - Veine	TSI - Alpes

Tableau 1 - Exemple de relations dans la norme ISO 25964-1

2.2.2.6 La création de domaines thématiques – subject areas.

Il est possible de créer des groupes de concepts ou de regrouper les concepts par facettes. Ces regroupements ne sont guère utiles dans le cadre de l'interopérabilité des langages documentaires, mais ils permettent d'enrichir les dispositifs d'accès à l'information en filtrant ou rassemblant les résultats de recherche.

Le regroupement par domaines ou thématiques permet d'opérer une classification parallèle aux relations des thesaurus, qui lui est indépendante : Une organisation chronologique ou géographique des concepts, une organisation par discipline ou par domaine d'activité métier. Chaque groupe de concepts peut former un microthesaurus. Les microthesauri peuvent admettre une hiérarchie interne.

Le regroupement par facettes fait référence à la classification à facettes de Raganathan [5, MANIEZ] Il se matérialise par la constitution de groupes de concepts de même rang autour d'une facette particulière. L'utilisation de catégories fondamentales objets, matériaux, agents, actions, lieux, temps est fréquente. Elles admettent aussi des spécifications.

2.2.3 ISO 25964-2 : Rendre les vocabulaires interopérables

2.2.3.1 Définition et portée des alignements

Dans le cadre de la recherche d'information ou de document, le principal but de l'interopérabilité entre les vocabulaires est de faire correspondre une requête utilisant un vocabulaire spécifique à un autre vocabulaire. Désigné comme *« mapping »* au sein de la norme, cette correspondance peut se traduire par alignement. Les deux langages étant rendu interopérables par ces alignements, les résultats de recherche comprendront des documents issus des deux fonds, bien qu'ils utilisent à la base différents langages documentaires.

Alors que la première partie de la norme rendait possible l'interopérabilité entre thesauri à l'aide de préconisations formelles, la norme ISO 25964-2 établit l'interopérabilité entre différents langages documentaires et un ou plusieurs thesauri, en proposant des équivalences entre les concepts des vocabulaires de ces langages. Les vocabulaires entrant dans le champ de ses préconisations sont les thesauri, les schémas de classification, les taxonomies, les listes d'autorités, les ontologies et terminologies. Selon la norme, ils ont tous pour point commun de décrire des concepts, bien que nous parlions de classes pour les classifications ou de catégorie pour les taxonomies [25, NISO, p.16].

Outre la définition de recommandations pour effectuer ces alignements, la norme établit des préconisations pour les alignements de chacun des vocabulaires et le niveau d'alignement possible de ces langages avec des thesauri. Ces recommandations peuvent être utiles au chef de projet amené à rapprocher différents langages, incluant des thesauri.

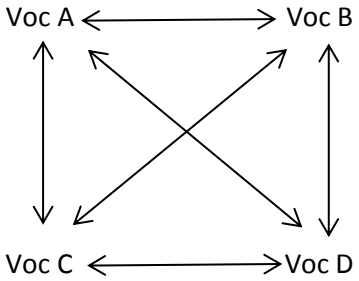
2.2.3.2 Des recommandations pour les alignements

La norme propose trois modèles d'alignement, en fonction du type de langage à disposition et des options d'alignement. Elle énumère également les facteurs de

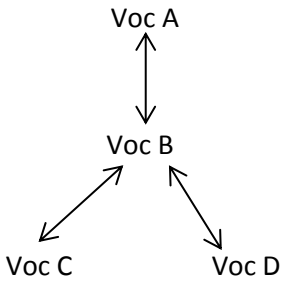
décision qui peuvent conduire à adopter l'un ou l'autre des modèles, voire à combiner les modèles [25, NISO, p.20].

- Le modèle 1 renvoie à des langages qui partagent la même structure hiérarchique et associative entre concepts, comprenant l'interopérabilité entre deux thesauri. Dans ce cas, il est possible de se satisfaire des recommandations incluses dans ISO 25964-1 et de se passer d'alignements.

- Le modèle 2 concerne des langages qui ne possèdent pas la même structure et qui ne sont pas de même nature. Les correspondances entre concepts peuvent être directs et réciproques sur les différents vocabulaires.



- Le troisième modèle, « en hub », propose des alignements indirects en prenant un vocabulaire de référence sur lequel seront alignés les concepts des autres vocabulaires. Le vocabulaire de référence devra incorporer tous les concepts présents dans les vocabulaires satellites. Il sera parfois nécessaire de créer un langage documentaire.



Il est également possible de faire des alignements sélectifs, notamment dans le cas où il y'a peu de recoupements entre les vocabulaires.

2.2.3.3 Exemple de relation entre les vocabulaires

Les alignements peuvent être constitués par des relations de hiérarchies, d'association ou d'équivalence. Lorsqu'un thesaurus se rend interopérable avec d'autres thesauri ou vocabulaires, la relation la plus utile est l'équivalence entre les concepts. Nous ne traiterons donc que de celle-ci, comme un exemple de la profondeur des alignements préconisés. La norme conseille différents types d'équivalences, parmi les suivantes [25, NISO, p.26] :

- Une équivalence simple : Les concepts sont identiques. Ex : Téléphone mobile, téléphone cellulaire. La relation est exprimée par EQ. Il faut dans ce cas faire attention aux concepts homonymes qui peuvent révéler des réalités différentes.

- Une équivalence composée : Un concept complexe issu d'un vocabulaire peut correspondre à une combinaison de plusieurs concepts dans d'autres vocabulaires. Exemple : « blé génétiquement modifié » peut renvoyer aux termes « modification génétique » et « blé ». L'alignement se fera dans une direction. Il existe deux types d'équivalence composée, la première correspond à notre exemple, en ce qu'elle concerne deux termes mis en intersection. Elle est notée EQ+. La seconde est cumulative car les termes correspondant ne représentent que des parties du terme lié. Exemple : « Voies navigables » peut être aligné sur à la fois « rivière » et « canal ».

Les équivalences entre vocabulaires ont différentes profondeurs, quelle que soit la relation établie. Elles peuvent être exactes, inexactes ou partielles. Cette inexactitude peut dépendre des milieux culturels et de contextes d'utilisation.

2.2.3.4 Utilisation des alignements à l'indexation et à la recherche

Un alignement peut être utilisé à l'indexation ou à la recherche. A l'indexation, les termes indexés avec un vocabulaire A sont convertis aux termes correspondants dans le vocabulaire B. Cette opération peut s'effectuer pendant l'indexation routinière ou en conversion de la collection complète, en supplément de petites mises à jour des vocabulaires et/ou d'ajouts de documents à la collection.

A la recherche, les termes d'origine ne sont pas changés. Pour utiliser le vocabulaire B lors de la recherche dans une collection indexée en A, la source et les vocabulaires cibles pour les alignements doivent être établis en direction inverse. Cela permet des requêtes comprenant des termes du vocabulaire B d'être

converties au terme correspondant du A. Les alignements sont incorporés aux processus de recherche et pourront se matérialiser par des propositions de terme.

Parfois, les *mappings* sont appliqués de manière automatique, mais une intervention humaine sera nécessaire pour tous les cas qui ne relèvent pas de l'équivalence simple. Par exemple, lors d'une recherche d'informations sur la protection de l'enfance (voc A), si le terme n'existe pas dans le vocabulaire B, le chercheur se verra proposer un mapping inexact *« safeguarding children »* ou un mapping associé *« child care »*. Il pourra choisir l'un ou les deux termes en fonction de ce qui est plus proche de son besoin.

2.3 Applications concrètes de l'interopérabilité sémantique des thesauri

2.3.1 Coût et avantages d'un thesaurus interopérable

Créer un thésaurus conforme à la norme ISO 25964 est un investissement coûteux, autant qu'une ontologisation d'un thesaurus existant via SKOS. Dans l'étude de faisabilité de cette ontologisation, il faudra prendre en compte les avantages sur le long terme liés à l'interopérabilité de son thesaurus. Ils sont matérialisés, entre autres par le retour sur investissement (ROI) que peut apporter cette conversion.

- Un identifiant unique aura été créé pour chaque concept, ce qui permettra de désambiguiser les concepts.

- Le thesaurus sera plus cohérent du fait de relations précisées et contraintes. Il permettra ainsi d'éviter des ambigüités entre relations. Ainsi il sera possible d'éviter qu'un terme se retrouve en équivalent alors qu'il décrit une relation hiérarchique spécifique [21, SOERGEL].

- Le thesaurus pourra servir de base à des extensions de requêtes sur d'autres fonds (schémas de métadonnées, langages contrôlés, taxonomies) et sur plusieurs disciplines.

- Il sera possible de faire des extensions de requêtes aux concepts liés, dans plusieurs langues, et d'anticiper une demande d'information sur la base d'une requête précédemment établie.

- Les liens sémantiques établis permettent de répondre à une question en langage naturel. Le système serait capable de détecter l'ambiguité d'un terme et de demander une clarification à l'usager.

- Le thesaurus pourra être un outil d'aide au *text mining* sur le web avec une orientation de sens. Il pourra ainsi combler les défauts sémantiques de cette technique tout en conservant ses avantages.

Il existe un rapport entre spécialité et interopérabilité qui ne doit pas être négligé [23, MAROYE]. Plus le langage sera précis et spécifique à une organisation, notamment à travers ses notes d'utilisation, moins il pourra être interopérable avec d'autres vocabulaires, puisque non cross-disciplinaire. L'interopérabilité s'effectuerait donc au détriment de la richesse du langage. Il appartient ainsi au gestionnaire du thesaurus de peser entre la spécificité du langage contrôlée envisagée pour le besoin de l'utilisateur dans son domaine de connaissance et l'opportunité d'une interopérabilité. Il faut enfin garder à l'esprit qu'une évolution du thesaurus en interne doit être possible, car l'évolutivité est importante pour un retour sur investissement (ROI).

Le coût de la gestion et de la maintenance du langage sont également à prendre en considération. Il existe désormais des systèmes de création, de gestion et de maintenance de thesauri qui ont été créés sur la base de la norme et qui permettent de développer des alignements. Ces logiciels peuvent être gratuits. De nombreux modules de gestion de thesaurus au sein des outils de gestion intègrent des fonctions d'ordinaire dévolues aux seuls logiciels autonomes. Nous pourrons constater que c'est le cas du logiciel Adlib, utilisé par le CNC et la cinémathèque française pour construire leur portail commun (voir *Annexe 3*).

2.3.2 L'interopérabilité sémantique au sein du web de données culturelles

Dans le cadre de ses missions de démocratisation du patrimoine culturel, le ministère de la culture a activement veillé depuis la fin des années 2000 à mettre en œuvre des moyens allant dans le sens d'une ouverture et d'une intelligibilité des données en vue de leur réappropriation par ses publics.

La logique qui sous-tend cette politique est de se réapproprier l'espace culturel, mais aussi de repenser la médiation au cœur des professions de la culture. La documentation a longtemps été un département en amont, au service de l'inventaire et de l'archivage patrimonial des collections, principalement à destination des

érudits. Elle s'opposait aux missions de de médiation culturelle tournées vers l'extérieur et spécifiques aux publics. Un décloisonnement s'opère désormais, qui place la documentation au centre de cette médiation. Très actifs en terme d'innovation, les musées ont contribué à porter ce changement en replaçant au centre des œuvres des éléments documentaires qui permettaient de les comprendre : *« L'exposition est conçue afin de les mettre en regard avec des documents qui fournissent par exemple des informations sur le processus créatif, les conditions de leur découverte, les étapes de leur valorisation ou encore l'état actuel des connaissances lié à leur interprétation »* [26, DESPRES LIONNET]. Autant d'informations liées au dossier documentaire de l'œuvre désormais ouverte aux publics. Les expositions virtuelles offrent par ailleurs aux langages documentaires une place particulière : Listes d'autorité, index et thesaurus organisent et structurent l'information pendant la visite, prenant le pas sur les choix commissaires d'expositions. Ces outils reposant sur des normes documentaires fixent désormais *« les modalités d'apparition des objets et leur mise en ordre à l'écran »*.

Cette évolution de la place de la documentation dans la médiation culturelle se double du mouvement de l'open data culturel. L'open Data public repose sur le principe que les données produites par les agents publics doivent être librement accessibles à tous aux fins de consultation, de modification, de mise en forme et de réutilisation. Initié en 2009, ce mouvement a trouvé son écho en France via *Etalab*, organisme sous la direction du premier ministre[37]. Cette mise à disposition a été plus timide dans le domaine de la culture, du fait des contraintes liées au droit d'auteur, mais les encouragements prodigués par le ministère de la culture ont impulsé de nombreux projets, dont le plus fameux est SemanticPedia[38]. L'enjeu de l'open data culturel se situe dans un nouveau rapport à la médiation qui passe de la transmission de l'information culturelle à l'appropriation. L'usager est désormais co-producteur de la connaissance [27, SAJUS, LEROI]. Les interventions du médiateur et de l'usager deviennent croisées, rendant propice une appropriation des données via des applications et des outils qui peuvent être développés par tous types d'acteurs. De ce fait, la valorisation du patrimoine passe par la mise à disposition des données et par leur ordonnancement. Aux fins de rendre plus efficace

[37] Etalab a établi une licence ouverte pour la libération des données sur internet, puis accéléré le mouvement via une mise à disposition portail data.gouv.fr

[38] Adaptation française de DBpedia.

l'appropriation de ces données, il ne s'agit plus uniquement de les libérer, mais de faciliter leur utilisation.

Lancé en 2008, le programme Hadoc est né de ce déplacement qui porte la valorisation au niveau de la production des données culturelles. *«Le ministère a fait le pari que déplacer l'indispensable effort d'amélioration de la qualité des données le plus en amont possible dans la phase de production pouvait répondre aux besoins des métiers de disposer des données de référence, comme à celui des usagers d'accéder à des contenus de qualité sémantiquement cohérents et immédiatement compréhensibles »* [28, BRIATTE]. L'un des piliers de ce programme est l'harmonisation des vocabulaires dans le secteur de la culture. Son but est de décloisonner les métiers et les organisations et de sortir de la logique de silos existante. Ce décloisonnement passe par l'interopérabilité des langages documentaires utilisés par les acteurs pour les différentes bases de données patrimoniales, via des processus d'alignements.

Afin d'effectuer ces alignements, le Ministère de la Culture et de la Communication a initié le développement d'un outil de création et de gestion de thesaurus autonome, *Ginco*. Cet outil s'appuie sur la norme ISO 25964-1 et il a pour but de transformer les thesauri liés à la culture en référentiels sémantiques traduits en SKOS. Ainsi les vocabulaires produits par le ministère peuvent-ils devenir interopérables. Ils peuvent être téléchargés au format SKOS sur la plateforme *« data culture »* qui utilise les technologies du web sémantique[39]. Ils peuvent de ce fait être interrogés par des API et permettre des extensions de recherche sur la globalité des fonds dans les langages sont présents dans Ginco.

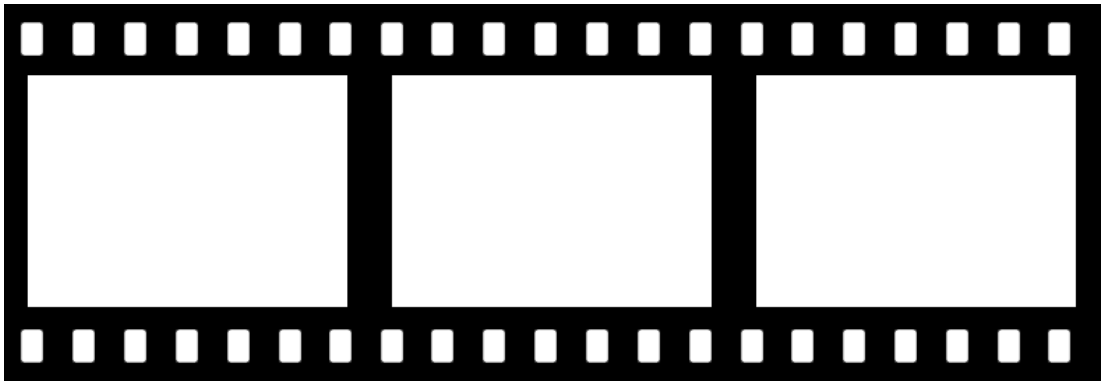
Ginco propose des fonctionnalités liées aux outils de gestion de thesauri traditionnels[40], mais il appartient à une nouvelle génération de ces logiciels qui intègrent la nouvelle norme, et donc intègrent dans leurs fonctionnalités l'interopérabilité des langages documentaires. Ils sont donc le produit d'un compromis entre l'ouverture et la spécificité, entre la progression vers l'interopérabilité des ressources et la conservation d'une certaine intégrité des thesauri au sein des institutions.

[39] A l'adresse http://data.culture.fr/thesaurus/

[40] Ginco, un éditeur SKOS Open Source, Sparna Blog, thomas. http://blog.sparna.fr/2013/09/19/ginco-un-editeur-skos-open-source/ consulté le 07/08/2016. Mise en ligne le 19 septembre 2013

Hier au statut de projet, l'interopérabilité sémantique est désormais au cœur des préoccupations en termes de gestion et de valorisation patrimoniale. Rendre interopérables les langages documentaires à une échelle globale dans l'optique du web de données semble être la voie d'avenir. Cependant l'interopérabilité s'effectue également à des échelles plus condensées, par des regroupements d'acteurs. A travers l'exemple d'une étude de faisabilité menée sur l'harmonisation des langages documentaires du CNC et de la cinémathèque française, nous pourrons remarquer que l'interopérabilité des langages documentaires peut être un vecteur de rapprochement des institutions et de diversification du rôle du chef de projet.

DEUXIEME PARTIE : AU CŒUR DE L'INTEROPERABILITE DES VOCABULAIRES FILMS ET NON FILMS



3. Le thesaurus, enjeu du rapprochement documentaire du CNC et de la Cinémathèque Française

3.1 Les Archives Françaises du Film et la Cinémathèque Française, deux visions de la valorisation du patrimoine cinématographique

3.1.1 Une rivalité historique

Connu comme le pays qui a vu naître le cinéma[41], la France fut également précurseur dans le domaine de la conservation des films. En 1936, Henri Langlois, George Franju et Jean Mitry créèent la cinémathèque française, association privée qui se présente comme une salle de cinéma et un musée. Henri Langlois considère déjà la conservation d'un film comme indissociable de sa diffusion. *« Notre tâche principale est celle de la propagation de la culture par le film »*[42]. On retrouve encore aujourd'hui cette empreinte muséale dans l'identité de la cinémathèque française.

Longtemps premier secrétaire de la Fédération Internationale des Archivistes de films (FIAF), Langlois s'oppose à la vision résolument documentaire d'Ernest Lindgren, premier conservateur du national film Archive, qui considère que, plus que leur diffusion, le premier devoir était de conserver les films, une conservation qui passait par la description. Archiviste de profession, Lindgren œuvre pour la création de catalogues décrivant les films et il murit déjà l'idée de normes internationales de description afin de créer un fichier mondial. Henri Langlois a sauvé de la destruction de nombreuses pellicules du cinéma des premiers temps provenant des studios Pathé et Gaumont et il les a rassemblées avec les moyens du bord, dans une méfiance des outils documentaires. Lors des visites des Archives du Film, on raconte encore que sa mémoire du cinéma suffisait à elle seule pour retrouver n'importe quel film, ce qui le faisait détenir une forme de pouvoir sur ses collections.

L'absence de réelle gestion documentaire et financière de la cinémathèque française alerta le ministère chargé des affaires culturelles d'André Malraux, qui accordait une large subvention à l'association. Il entreprit de devenir majoritaire dans le conseil d'administration, concédant à Langlois une position de directeur artistique. Ce dernier refusa, ce qui ouvrit le conflit qui amorça la séparation entre la

[41] La première projection publique et *payante* du cinématographe *eut lieu* le 28 décembre 1895 à Paris et marque la *naissance* officielle du cinéma.
[42] Cette phrase fut prononcée par Henri Langlois au Congrès de la FIAF de Copenhague en 1948.

cinémathèque française et les archives patrimoniales, connu sous le nom d' *«affaire Langlois »*. Le 9 février 1968, Henri Langlois acquis à sa cause de nombreuses personnalités pour faire entendre sa voix. Il parvint ainsi à être rétabli dans ses fonctions. Le Ministère décida en réaction de créer son propre établissement de conservation patrimoniale, *les Archives du film*, qu'il installa à la batterie de Bois d'Arcy, lieu où étaient conservées les collections films de la cinémathèque française depuis 1960. Nous avons donc à faire à deux institutions marquées par des vues différentes sur la conservation des films : L'une est une institution de valorisation de par son origine, l'autre a été créée dans le but d'opérer un contrôle étatique sur les conditions de conservation des films.

3.1.2 Evolution documentaire des Archives du Film et description des films

3.1.2.1 Du stockage à la valorisation

En 1961, un décret est publié qui interdit la circulation, la distribution et la projection des copies de films cinématographiques établies sur support inflammable, les « films nitrate »[43]. Dès 1967, l'État y entreprend la construction de différents bâtiments de stockage à la batterie de Bois d'Arcy : le premier, destiné à la Cinémathèque française, accueille des rayonnages conçus pour l'archivage des films sur support de sécurité acétate, le support officiel qui suivit l'interdiction de la pellicule nitrate. Aujourd'hui, 11 000 mètres carrés de bâtiments à température et hygrométrie constantes abritent des films déposés par des sociétés de production et de distribution des films, des laboratoires commerciaux, des cinéastes, des particuliers, mais aussi par la Cinémathèque française, la Cinémathèque de Toulouse, celle de Grenoble, les autres cinémathèques régionales, l'INA et même les archives nationales.

Suite à l'*Affaire Langlois*, les Archives du Film furent chargées de conserver en ces lieux les films confiés en dépôt[44], ou dont l'Etat possédait la propriété. Le dépôt légal qui était attribué à la bibliothèque nationale lui fut également confié en 1977 pour les films français, et en 1993 pour les films internationaux exploités en France, rendant de ce fait exponentiel le nombre de dépôts de bobines. Les années 70 avaient déjà été une époque de stockages en masse de films, dont beaucoup ne sont toujours

[43] Ils furent interdits par la loi puisque hautement inflammable, d'un feu considéré comme inextinguible. Ils ont provoqué divers incendie, dont celui du bazar de la charité.
[44] Comprenant les films de la Cinémathèque Française qui étaient sur support nitrate.

pas catalogués et décrits. Le service Analyse et Gestion Documentaire des collections (AGDC), au sein duquel j'ai effectué mon stage, œuvre encore à l'identification de ces films, souvent à l'occasion de la réunion de corpus. Ces travaux sont souvent l'occasion de découvrir des raretés. Les films sur support nitrate sont stockés dans des cellules spéciales. Il faudra attendre 1990 pour que soit mis en place un plan de sauvegarde et de restauration de ces films, consistant au transfert en nombre sur un support de conservation *(« support safety »)*, le polyester. Ce plan permit de sauvegarder et de restaurer 15 000 films de fiction et documentaires.

Les Archives Françaises du Film ont longtemps obéi aux seuls impératifs de stockage dans des conditions permettant une conservation maximale des supports. Dans cette optique, il n'y a pas eu pendant très longtemps de description documentaire des films et la gestion des dépôts de film était rudimentaire. En 1994, un système de code barre et de gestion informatique des stocks fut mis en place, permettant de recenser les mouvements et les emplacements. Le système TESS recensait les dépôts par déposants. *« C'était du pure inventaire, une analyse technique des éléments (ndr : les bobines de pellicule déposées), on ne faisait pas de documentaire ».*[45]

Le plan nitrate a amorcé la question de ce qui devrait être fait des films restaurés, comment rentabiliser l'argent investi. A l'occasion de ces restaurations, les équipes sortaient la totalité des éléments et ils vérifiaient s'ils avaient les mêmes contenus. Très vite, il apparut que décrire la différence entre les éléments revenait à décrire les contenus. Une pratique des résumés fut instaurée, mais elle n'était pas systématique, ainsi que des avis artistiques sur les films pour dégager des priorités sur les restaurations. *« Ils nous faisaient confiance en tant que spécialistes pour remonter les pépites »*. Les autres nécessités étant de retrouver l'année et le titre. A l'occasion du centenaire du cinéma, en 1995, sur 1428 films du catalogue Lumière, 1408 furent identifiés et restaurés par le Service des archives du film, en partie grâce à une collaboration internationale. Un catalogue dédié à cette production retrouvée et restituée au public fut édité. En 1996, un embryon d'indexation thématique existait déjà, avec des résumés, des notes sur l'intérêt du film et des fiches remplies avec des mots clés libres. En 1998, des fiches de catalogage étaient

[45] Les citations de cette page ont été recueillies auprès des agents du service AGDC. De nombreux documentalistes du CNC étaient déjà en poste dans les années 90. Ils ont pu m'apporter des renseignements précieux sur l'histoire des lieux lors de mes entretiens d'étude de besoins.

produites à l'analyse, mais peu exploitées. Leurs informations permettaient de valoriser les films dans leur en-tête.

Une première réflexion sur la description des films est menée à la fin des années 90, mais elle n'aboutit pas. La base Laure qu'utilisaient alors les archives du film constitue un « premier niveau » de catalogage : Une carte d'identité du film, des éléments par déposant, une identification par code barre. Lors des travaux préparatoires à la base Lise, une nouvelle réflexion fut menée sur l'indexation thématique. Les agents effectuèrent des formations sur les résumés. Un travail sur un hypothétique thesaurus décrivant les films fut amorcé et différents tests furent effectués. L'idée d'utiliser un thesaurus fut abandonnée au profit d'une indexation par mots-clés, méthodes alors privilégiée dans la profession et qui paraissait plus simple.

3.1.2.2 La base documentaire Lise

Développée à la fin 2004, *Lise* est la base de données des Archives françaises du film. Elle est mise en ligne en 2006 sur le site « www.cnc-aff.fr ». Elle propose une sélection de fiches documentaires renseignant certains des 100 000 films conservés aux Archives françaises du film. En plus du premier niveau de catalogage proposé par Laure, Lise propose un second niveau et plusieurs onglets qui permettent de décrire avec plus de détail les collections de films du CNC, de la cinémathèque française, de la cinémathèque de Toulouse et d'autres partenaires en régions.

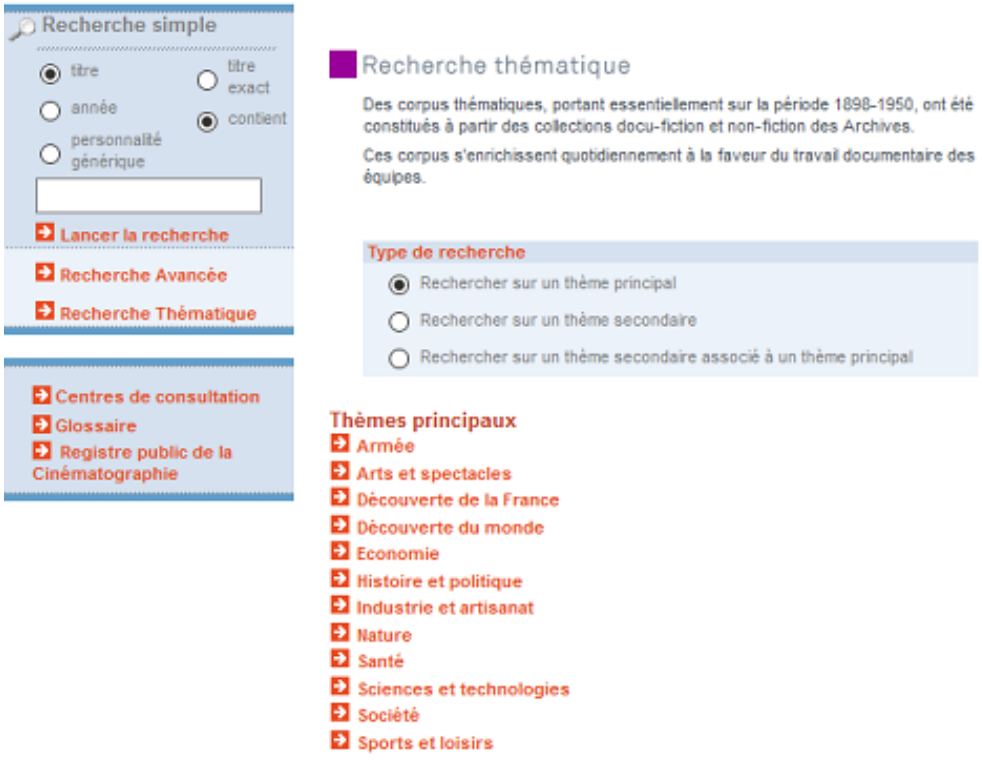
La base Lise comporte dans l'onglet indexation :

- Une liste de mots clés, dont les propositions de descripteurs ont été fermées. Le nombre de termes s'élevait à 8277 en août 2016. Les mots clés n'ont pas été hiérarchisés, bien qu'un pas ait été franchi récemment par la documentaliste du service documentation du *CNC*.
- Un thesaurus géographique, utilisé pour le champ lieux de tournage.
- Des champs « Genre Général » et « Genre précis ».
- Des champs « Thème principal » et « Thème secondaire », constitués principalement lors de la gestion du *corpus Lumière*.
- Un champ « thème de valorisation », crée à l'initiative du Service Accès et Valorisation des Collections (SAVEC) des Archives Françaises du Film.

- Une recherche est possible sur les mots des résumés.

Depuis la création de Lise, il n'y a jamais eu de termes candidats. L'indexation a été libre pendant plusieurs années, puis les mots clés ont été « fermés » du fait d'un trop grand éparpillement et notamment de termes trop spécialisés. Un « nettoyage » est depuis effectué pour fusionner des doublons. Désormais, il est nécessaire de demander à l'administratrice de Lise pour créer de nouveaux mots-clés.

La particularité d'une telle liste ouverte, contrairement à un thesaurus qui préexiste à l'indexation tient dans le fait que les mots clés entrés sont ceux auxquels au moins un film est rattaché. Les termes sont donc représentatifs des fonds films du CNC. La base Lise possède une interface publique avec une vue sur les notices et les résumés des films qui ont passé le deuxième niveau d'indexation. Elle comporte une recherche simple, une recherche avancée et une recherche thématique. Chacun des champs de l'onglet indexation y sont interrogeables pour filtrer la recherche, excepté les thèmes de valorisation.



L'interface publique de recherche thématique de Lise

3.1.2.3 Le traitement documentaire, maillon d'une orientation vers la valorisation des collections

Il ressort de l'évolution des Archives du Film une progression du stockage vers la valorisation des collections. Cette vue sur la valorisation n'est pas nouvelle. Elle est soutenue depuis plusieurs décennies par le Service Accès Valorisation et Enrichissement des Collections (SAVEC). L'objectif du SAVEC est de faire connaître les collections et de les rendre accessibles au public sous des conditions de détentions de droit et conventions avec les déposants des films. Les interlocuteurs du SAVEC sont donc composés à 80% de professionnels. Les collections sont valorisées et diffusées dans le cadre de manifestations culturelles en France et à l'étranger, rétrospectives de cinémathèques, festivals, programmations. Elles sont également enrichies par des dépôts avec visées de restauration[46].

Un niveau de maturité dans la valorisation des collections a été franchi par la création du département Analyse et gestion documentaire des collections (AGDC) en 2007, qui vise déjà une valorisation auprès du grand public. Suite à l'implémentation dans Lise du deuxième niveau d'indexation des collections films, les Archives ont entrepris un travail de fond sur la description des collections afin de mieux guider les utilisateurs dans leurs recherches. L'AGDC gère ainsi l'ensemble des informations d'indexation suivant le catalogage de premier niveau établi par le service inventaire. Il est également en charge des analyses d'éléments (copies, négatifs, contretypes…) qui précèdent les restaurations, assurées les laboratoires internes ou déléguées à des laboratoires privés. C'est aussi à ce service qu'échoue la rédaction d'un résumé et de l'attribution de mots-clés, de thèmes, de sous-thèmes, de genres précis et de lieux de tournage. Ce travail d'indexation permet de constituer des thématiques ou corpus à géométrie variable qui sont autant de clés d'entrée aux collections accessibles aux internautes dans les recherches avancées et thématiques de film sur l'interface public. L'analyse et la gestion documentaire permettent de valoriser au mieux les films. Ils peuvent ainsi être consultées à Bois-d'Arcy si non-numérisés et à la Bibliothèque Nationale de France si numérisés ou encore être programmés par le SAVEC à l'occasion de festivals ou de rétrospectives.

[46] **Éric Le Roy**, « L'accès aux collections des Archives françaises du film-CNC », *1895. Mille huit cent quatre-vingt-quinze* [En ligne], 41 | 2003, mis en ligne le 21 novembre 2007, consulté le 05 septembre 2016. URL : http://1895.revues.org/773 ; DOI : 10.4000/1895.773

3.1.3 Evolution documentaire de la Bibliothèque du Film (BiFi) et de la cinémathèque française

3.1.3.1 Définition et développement des fonds non-films

Suite à *l'Affaire Langlois*, la cinémathèque française a poursuivi les visées de valorisation de ses créateurs en augmentant le périmètre de visibilité de ses collections.

Les films sur support nitrate de la cinémathèque française sont conservés d'une part dans les locaux des *Archives du Film*, et d'autre part au *fort de Saint-Cyr* à Montigny le Bretonneux. Les documentalistes films de l'Institut indexent leurs collections dans *Lise*. Ils utilisent donc les mêmes mots clés et champs d'indexation que ceux utilisés par les Archives Françaises du Film.

Le non-film est une particularité française. On peut définir ainsi toute la documentation papier, le matériel publicitaire, les archives de création qui font partie de l'institution et qui sont liés à des films. On y compte donc les affiches, les photos de tournages, les archives, les costumes, les machines ou les scénarios. Par extension terminologique, dans les cinémathèques de France, les monographies et les périodiques sur le cinéma sont considérés comme des non-films.

Le traitement documentaire des collections non-films de la cinémathèque française se répartit en plusieurs pôles en fonction du fonds. Pour chaque document, nous trouvons un documentaliste référent. Les périodiques, monographies et thèses sont accessibles au public à la Bibliothèque du Film[47]. Les documents iconographiques (affiches, photos) sont rassemblés à l'iconothèque. Les archives sont conservées à l'espace chercheur. Il est à noter que des collections non-films sont également présentes aux archives du film (voir tableau des collections en annexe 4, p.132). Elles comprennent les documents attachés au film au dépôt légal et ceux de la bibliothèque des archives du film, principalement destinés à l'identification des films déposés. La bibliothécaire catalogue ses monographies et périodiques sur le module *Cinédoc 3*, au même titre que les documentalistes de la cinémathèque française et des autres cinémathèques utilisant le portail *Ciné-ressources*.

[47] La Bibliothèque du Film (BiFi) a fusionné avec la cinémathèque française en janvier 2007 et a intégré par la suite les locaux de la rue de Bercy.

3.1.3.2 Création du thesaurus *Cinédoc* et indexation iconographique

Sur le modèle des bibliothèques, le catalogue de la cinémathèque française est fondé sur des listes autorités personnes et films. Transverses à tous les fonds non-films, elles permettent d'effectuer des recherches sur toutes les collections.

Le module *Cinédoc 1* servant à cataloguer les ouvrages de la bibliothèque du film vit le jour en 1995. Il s'accompagna d'une réduction considérable des termes des fiches papier matières qui le précédaient. Une liste d'autorité matières fut alors constituée, avec vérification des candidats descripteurs, sans qu'il fut possible de faire une recherche par mot matière sur l'interface publique *IP-Ciné*.

C'est en 1999 que le contexte apparut favorable pour la création d'un thesaurus matière, avec le basculement des données entre *Cinédoc 1* et *Cinédoc 2*, l'optique de refonte de l'interface public et la volonté de la bibliothèque du film de développer un outil d'indexation sur le cinéma. La production du thesaurus *Cinédoc* démarra en septembre 1999. En décembre 2000, la première version fut validée après les tests de la bibliothèque du film et des différentes cinémathèques partenaires. Le thesaurus fut mis en pratique en janvier 2001 par l'indexation du courant et la reprise du passif[48]. Au Printemps 2002, les lecteurs pouvaient rechercher les ouvrages et les périodiques par mot-matière via la nouvelle interface publique Ciné Ressources.

Le thesaurus *Cinédoc* est un bon exemple de thesaurus spécialisé[49], qui admet quelques perfectionnements, mais satisfait globalement ses utilisateurs.

Une reprise des thèmes du thesaurus iconographique Garnier, utilisé par les musées de France, sert à l'indexation des plaques de lanterne. Ce fonds particulier regroupe des plaques de verre peintes à la main qui étaient projetées plus de deux cent ans avant l'invention du cinéma. Les plaques de verre représentant des machines de cinéma, ces techniques sont indexées en sujet sur *Cinédoc*. Les scènes présentes sur les plaques sont quant à elles décrites par une adaptation du thesaurus Garnier intégré à une base de données maison, *laterna magica*[50]. Cette adaptation se présente en backoffice sous la forme d'une arborescence sources / domaine / thèmes. Suivant un domaine, certains thèmes sont proposés à l'indexeur. La Cinémathèque française souhaite conserver cette adaptation de Garnier car il

[48] Opération visant à effectuer la concordance entre les termes de Cinédoc 1 et ceux du thesaurus
[49] Ceci explique que nous l'avons utilisé pour illustrer la notion de thesaurus dans la première partie.
[50] Le catalogue se trouve à l'adresse http://www.laternamagica.fr/

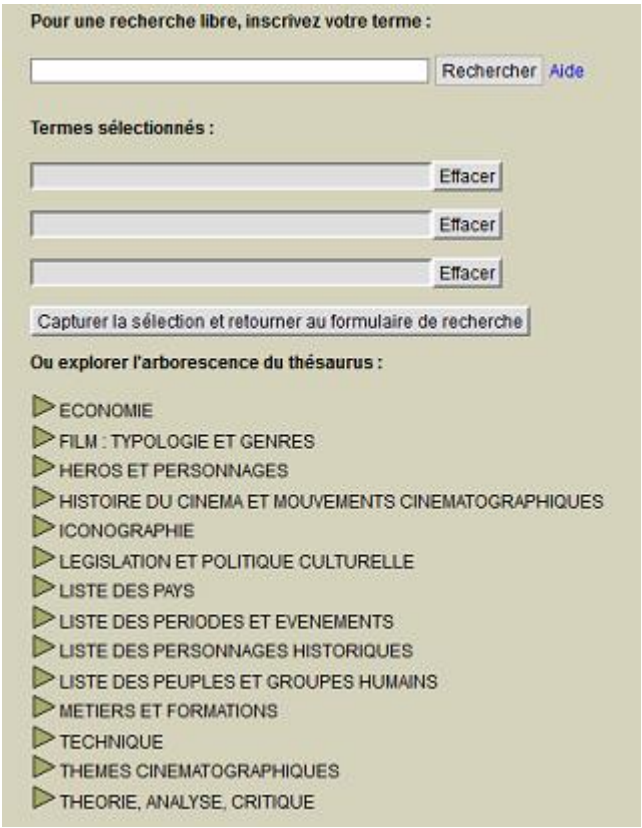
permet de décrire tout type d'image et surtout de réaliser une indexation des documents iconographiques en harmonie avec les musées de France.

3.1.3.3 Création du portail *Ciné-ressources* et du site web

Ciné-ressources est un catalogue collectif des bibliothèques et archives du cinéma. Il fut développé par la Bibliothèque du Film, la cinémathèque de Toulouse et d'autres cinémathèques partenaires, puis repris en interne par la cinémathèque française en janvier 2007. Une interface publique fut lancée le 22 août 2007.

Ces institutions ont adopté les outils de l'application *Cinédoc* et ils cataloguent leurs collections dans une base unique de données documentaires. *Ciné-ressources* permet au public, via un moteur de recherche, de prendre connaissance des collections non film conservées par les institutions partenaires. Chaque référence indique la localisation, unique ou multiple, de l'élément conservé.

Il est possible de rechercher des documents simultanément à travers toutes les collections des cinémathèques partenaires, à partir d'un titre de film ou d'un nom de personne physique ou morale. Il est aussi possible de rechercher un document dans une collection particulière (affiches, archives, dessins de décorateurs et costumiers, ouvrages, articles de périodiques, photographies, revues de presse, vidéos et DVD) en introduisant des critères de recherche spécifiques à cette collection, et d'avoir accès au thesaurus Cinédoc. Enfin, l'interface fait office de portail et permet d'accéder aux catalogues ou répertoires documentaires propres à chaque institution.



Interface d'accès au thesaurus Cinédoc via Cinéressources

La cinémathèque française a également développé un site web permettant de valoriser ses programmations et ses différentes collections, auquel elle apporte un soin particulier. Il est passé à une nouvelle version en 2015, plus ergonomique et responsive design.

Aussi avant de nous attarder sur l'étude de faisabilité d'un thesaurus englobant les films et les documents liés aux films, ces quelques considérations sur les parcours respectifs des acteurs principaux du projet permettent d'établir deux évolutions parallèles. Cette évolution tient des missions dévolues à chacune. Il s'opère néanmoins une conjonction progressive vers des croisements et des visées de valorisation des collections, de laquelle les services documentaires des archives du film sont les corollaires. Un certain degré de maturité dans la valorisation des collections qui rend possible le rapprochement de deux institutions aussi différentes. Il est désormais nécessaire de présenter le contexte de la fusion, très important pour notre étude.

3.2 Le projet plateforme, un outil de rapprochement des pratiques documentaires des équipes

3.2.1 Contexte et enjeu de la mutualisation

Le développement du web 2 .0 a vu naître des moyens d'accès collaboratifs aux connaissances et à la culture. Dans le domaine du cinéma, de nombreuses bases de données sont nées d'initiative amateurs. Aux côtés des célèbres IMDB (internet movie database) et allociné, la page *Wikipédia « liste de bases de données cinématographiques »* recense pas moins de 111 bases de données liées au cinéma, aux périmètres (domaine de spécialité) et provenances (projets propriétaires, privés, associatifs, communautaires) très différents[51].

L'utilisation soutenue de ces ressources par le grand public, au point de faire de ces créations des acteurs majeurs De l'économie numérique, a alerté les organismes officiels. Il n'est d'ailleurs pas innocent que lors de mon enquête sur les pratiques de recherche des professionnels de la documentation, dans une institution comme la direction du patrimoine du CNC, beaucoup aient répondu qu'ils servaient de ces sites, autant que des ressources internes. Quoi de plus normal que de chercher ailleurs des fonctions que son environnement offre à un niveau limité ? Le même constat s'impose au niveau de la documentation sur le cinéma, avec des ressources amateures très riches souvent citées par les professionnels de la documentation telles que le site *calindex*[52], qui dépouille un grand nombre de périodiques cinématographiques. En bon joueur, le portail *ciné-ressources* de la cinémathèque française est le premier à mettre en avant ces initiatives à travers *Cinéweb*, un *« répertoire de ressources sur le cinéma qui propose aux différents publics de la Cinémathèque française des sites web indispensables pour se documenter de manière complémentaire aux ressources papiers et électroniques déjà disponibles sur nos sites Internet ou au sein de la Bibliothèque du film »*[53].

Très complet, ce répertoire atteste d'une reconnaissance d'initiatives officieuses. Cette reconnaissance de concurrents insoupçonnés s'accompagne d'une profonde remise en question des acteurs culturels officiels, qui doivent désormais composer avec, en cherchant leur place dans cet écosystème. A l'instar du développement des webzines ciné qui ont engendré des avatars de magazines en ligne, les bases

[51] Consulté le 10/10/2016. https://fr.wikipedia.org/wiki/Liste_de_bases_de_donn%C3%A9es_cin%C3%A9matographiques_de_l%27Internet

[52] https://calindex.eu/

[53] http://www.cineressources.net/repertoires/repertoires.php?institution=BIFI .

de données cinématographiques allaient être récupérées sur la toile (exemple : Les fiches du cinéma). De fil en aiguille, les grands acteurs publics et privé saisissent les possibilités que peuvent offrir les TIC dans la diffusion publique de leurs catalogues et de leur identité. Il est souvent apparu nécessaire de mutualiser les ressources pour pouvoir faire face aux concurrents du web.

« La mutualisation est le partage par des individus ou groupe d'individus, de biens, de logements, d'équipements ou de moyens de transport de manière à optimiser l'accès à ces ressources et leur rentabilité »[54]

Phénomène prégnant depuis le milieu des années 2000, la mutualisation des ressources a vu se rassembler de nombreux acteurs publics du secteur culturel, dans une période de restrictions budgétaires. La mutualisation peut-être un effet de nécessité ou d'un projet fondé sur des valeurs communes. Emanation irrésistible de son époque, la fusion des portails *Lise* et *Ciné-ressources* participe des deux logiques. A destination du grand public, elle obéit à une volonté de proposer une base de données cinématographique sous contrôle, une alternative de « qualité » à ses concurrents web plus populaires. De ce fait, elle vise à entraîner le public du cinéma de divertissement vers un cinéma de patrimoine. Mais elle part aussi d'une conscience que cette mission ne pourra pas se faire dans un contexte de division entre les Archives du Film et la cinémathèque française.

3.2.2 Genèse et développement du projet plateforme

Le projet d'un portail web mutualisant l'accès aux collections patrimoniales du CNC et des cinémathèques est né en 2011, sous l'impulsion du président du CNC. Le CNC et la cinémathèque française désignèrent chacune un chef de projet en interne, et ils s'allouèrent les services d'assistance à maîtrise d'ouvrage de la société *Doxulting*. Un recueil des besoins, suivi d'un premier cahier des charges fut élaboré avec le concours de l'AMOA, lançant un appel d'offre aux éditeurs de SIGB.

Ce premier appel d'offre fut déclaré caduc faute d'un calendrier réaliste. Un second appel d'offre fut lancé, qui fut remporté par le logiciel *Adlib* de l'éditeur *Axiell*. Développé en collaboration avec le *British Film Institute*, *Adlib* est à l'origine un logiciel de musée qui n'a pas connu d'aussi gros projet que celui du CNC. Il a su néanmoins intégrer le modèle de la norme EN 15907-2010, déclinaison du modèle FRBR pour décrire l'œuvre cinématographique, qui permet d'établir des liens entre

[54] Cette définition provient de l'article de wikipedia « Mutualisation des services ». https://fr.wikipedia.org/wiki/Mutualisation_des_services

l'œuvre, ses variantes, ses manifestations et ses items[55]. Ce modèle de données sera l'ossature de la future plateforme dans l'optique d'un catalogage commun des ressources.

La phase de spécification qui suivit mis en lumière l'ampleur, la complexité du projet et les hésitations d'une collaboration encore fraîche entre les deux institutions. Les années passèrent, qui furent mises à contribution par le CNC et la cinémathèque française pour opérer notamment des nettoyages dans leurs bases. A ma première venue au CNC, en avril 2016, le premier jalon du projet touchait à sa fin, avec la fusion des autorités films et personnes des deux catalogues. Le deuxième s'amorçait en parallèle, avec la construction du portail par lequel le public pourrait interroger la base de données. La société *Cap Gemini* fut choisie pour créer ce portail, qui serait relié par API à la base de données. La nécessité d'un thesaurus transversal aux films et non-films était définie pour le backoffice, mais son intégration au front office était encore incertaine à ce niveau d'avancement du portail. Il est néanmoins porté au cahier des clauses techniques particulières pour la plateforme que *« La recherche devra exploiter les thesaurus thématiques et géographiques, en prenant en compte les relations hiérarchiques (autopostage[56]) et d'équivalence. Chaque thésaurus devra être disponible dans les interfaces de recherche. ».*

3.3 Les enjeux d'une recherche fédérée thématique au niveau interne

Outil de rapprochement, le thesaurus transversal en projet répond autant à une nécessité de toucher le grand public via le portail qu'à un besoin de coopération au niveau interne. Il devrait rendre interopérables les fonds afin de donner des résultats qui intégreraient sur une même recherche sujet des documents films et non-films dans un but d'optimiser les recherches des agents.

Une indexation matière commune sur les films et non-films pourrait faciliter les projets d'exposition et les publications sur des thèmes particuliers liés au cinéma. Le recoupement est parfois nécessaire entre les films et les non-films dans le cadre de l'organisation d'expositions, de programmations thématiques ou à l'occasion de réunions de corpus. Le besoin ce type de recherches sera amené à progresser avec

[55] La norme est disponible à cette adresse : https://www.boutique.afnor.org/norme/bs-en-159072010/identification-des-films-moyens-d-ameliorer-l-interoperabilite-des-metadonnees-ensembles-et-structures-des-elements/article/702587/eu111169

[56] Procédé permettant d'utiliser automatiquement lors de l'indexation des documents ou de la formulation d'une question, les descripteurs appartenant à une même chaîne hiérarchique »

l'augmentation actuelle des programmations des cinémathèques sur des thèmes précis du cinéma (Exemple : la cinémathèque française, *« les élections américaines »* en novembre 2016), qui nécessitent de documentation non-films pour être illustrées.

Dans le cadre de la conservation des films, une recherche non-films peut servir de base pour retrouver des éléments manquants. Les archives Jean Renoir et les éléments de la cinémathèque ont par exemple permis au CNC de reconstruire *« la vie est à nous »* de Jean Renoir. L'importance du lien entre un article de périodique, une monographie traitant d'un film et ce film est mise en avant par la bibliothèque de la direction du patrimoine du CNC dans le cadre de recherches internes liées à l'identification des films.

Un thesaurus matière améliorerait les conditions de recherche et de production de contenus thématiques éditoriaux sur le portail. Les différentes activités des partenaires seront en effet valorisées via la partie éditoriale du portail. Cette « éditorialisation » pourrait se baser sur les termes du thesaurus via des résultats d'enquêtes menées par les documentalistes sur un thème précis (exemple : *« la danse au cinéma »).*

Des indexations a posteriori sur un champ valorisation dédié (différent du champ sujet) pourraient à la fois servir aux documentalistes confrontés à des demandes similaires et aux usagers externes désireux de se renseigner sur un thème particulier. Elles pourraient aussi bénéficier aux recherches sur le portail à travers le thesaurus, en orientant sur des sujets où d'autres sources grand public donneront peu de renseignements.

Un autre enjeu se porte au niveau des processus documentaires, dans les conditions de production et de validation des notices et plus généralement, le « processus qualité » de l'indexation. Le processus de validation en place au CNC qui s'est construit sur une liste libre est lourd et impose une relecture des notices, de nombreuses fusions de termes opérés a posteriori ainsi que des relectures systématiques. Il y'a actuellement 10000 fiches indexées dans Lise et chaque notice a été relue et corrigée. Des mots clés ont été enlevés et repris. Le processus de validation semble particulièrement fastidieux. Un thesaurus pour décrire les films éviterait les contresens et les termes polysémiques.

Dans une problématique intégrant la gestion des connaissances, il est également mis le doigt sur l'expertise des professionnels (particulièrement sur les fonds

iconographiques) qui disposent de fonds qu'ils connaissent très bien et effectuent eux-mêmes la recherche thématique. Il est nécessaire de pouvoir entrer cette expertise sur les thèmes de leurs fonds pour l'exploiter en cas d'absence ou de remplacements sur le poste. Un thesaurus permet d'organiser la connaissance des institutions. Il est donc nécessaire de pouvoir impliquer les différents acteurs de ces institutions dans sa constitution et dans son évolution.

4. Conduire l'étude de faisabilité d'un thesaurus film et non-films : Méthode et ajustements

4.1 Importance d'une étude préalable à la construction d'un thesaurus

La première partie de ce mémoire aura mis en évidence les possibilités d'exploitation d'un thesaurus. Il s'insère désormais dans un contexte *« macro »* qui tend vers une interopérabilité « globale » des vocabulaires, tout en conservant les spécificités liées à son environnement documentaire d'origine. Ces différents niveaux d'exploitation ont très vite renforcée la nécessité, consacrée théoriquement mais souvent éludée dans la pratique, d'une étude de faisabilité préalable à la constitution du thesaurus. Il est à la charge du chef de projet de communiquer sur son travail et de ne pas se laisser embarquer dans la structuration d'un thesaurus sans prendre de distance avec l'ouvrage. L'impression de gagner du temps en se passant d'une étude préalable laissera bien souvent des difficultés dans la conception et la maintenance de l'outil qui seront autant de perte pour les partenaires et de nouveaux moyens correctifs à mobiliser.

4.1.1 Mobiliser autour du projet en interne

Un projet de thesaurus est un investissement important en termes de temps et d'infrastructure. Il doit donc être provisionné financièrement et organisé en amont afin de pouvoir affecter les dépenses et les personnes sur une durée prolongée. Confier à un seul documentaliste spécialisé la construction de l'outil est séduisant, mais cela engendre le risque qu'il ne soit adapté qu'à une seule conception du domaine. Il est donc conseillé d'impliquer plusieurs personnes dans son élaboration, dont certaines seront issues des équipes qui utiliseront le thesaurus à l'indexation. La phase d'analyse des besoins peut être utile pour débusquer ces documentalistes manifestant un intérêt pour le projet et qui pourront ensuite le soutenir auprès de leurs collègues. Elle peut être aussi l'occasion de dégager de nouveaux enjeux qui permettront de soutenir cet investissement auprès des instances décisionnelles. Une étude de faisabilité bien menée permettra d'établir des spécifications et un cahier des charges répondant à de nombreuses interrogations [29, DEGEZ].

4.1.2 Intégrer le thesaurus dans un système

Dans le cadre de la coopération du CNC et de la Cinémathèque Française, la décision de constituer un thesaurus dérive particulièrement de la fusion des portails *Lise* et *Ciné-ressources*. Elle est la conséquence directe du choix du logiciel *Adlib*

pour construire la future plateforme d'accès aux collections patrimoniales. *Adlib* comporte deux bases différentes pour les langages documentaires :

- Une base d'autorités qui va accueillir toutes les listes d'autorités fermées, à savoir les autorités personnes et films une fois que la fusion aura été opérée.

- Une base Thesaurus qui accueille toutes les listes ouvertes. Cette base comprend donc les termes sujets et les différentes listes liées aux collections des partenaires (exemple : Format et support pour les collections films). C'est dans cette base qu'un thesaurus matière commun devra être intégré.

Le logiciel intègre des fonctionnalités thesaurus qui sont des modules plus ou moins développés, mais qui n'atteignent pas le niveau de gestion d'un outil autonome.

D'une façon générale, les spécifications techniques menant au choix d'un SIGB ne s'attardent pas sur les fonctionnalités de l'outil liées à la gestion de thesaurus. Ainsi de mauvaises surprises peuvent apparaître lorsque la nécessité d'employer un langage contrôlé d'indexation se fait jour. En l'espèce, ces fonctionnalités étaient prévues dans le cahier des charges techniques de la solution. La cheffe de projet plateforme coté CNC a ainsi pu me briefer sur le fonctionnement de la base Thesaurus. Cependant, ces fonctionnalités faisaient l'objet d'interrogations parmi les équipes des deux institutions et nécessitaient d'être auditées. Les capacités de l'outil à gérer la construction d'un thesaurus permettent de déterminer le niveau de difficulté qui attendra les équipes chargés de la conception du thesaurus. Elles permettent aussi d'évaluer si d'autres outils seront nécessaires pour compléter celui-ci, et à quels niveaux l'expertise technique et l'implication humaine devra intervenir.

Le travail sur les fonctionnalités Thesaurus d'Adlib fut indissociable des autres phases de l'étude de faisabilité. Il se matérialisa dans un dialogue avec les équipes de l'éditeur. Il se concrétisa par un guide pratique des fonctionnalités thesaurus d'Adlib à destination des chefs de projet et des équipes liées au projet plateforme (Annexe 3, p.114). La découverte de nouvelles fonctionnalités se révéla par ailleurs un critère d'appréciation non négligeable dans la faisabilité d'un des scénarios finaux.

4.1.3 Documenter la construction du thesaurus

Une étude de faisabilité bien menée permet de répondre en amont et de manière documentée à de nombreuses questions sur la construction du futur outil. Le

tableau ci-dessous, inspiré des conseils de l'ouvrage *« Guide pratique pour l'élaboration d'un thesaurus documentaire »* [30, HUDON] donne un aperçu du type de questions qui peuvent obtenir des réponses durant ces phases. Chaque question est surlignée par une couleur indiquant la phase de l'étude de faisabilité qui permet d'y répondre : En bleu l'analyse des besoins, En jaune l'étude de faisabilité, en vert la caractérisation des contraintes internes.

<table>
<tr><th><u>Questions trouvant des réponses dans l'étude de faisabilité</u></th><th colspan="2"><u>Implications dans la construction du thesaurus</u></th></tr>
<tr><td rowspan="2">La ou les disciplines décrites sont-elles des domaines émergents ou des disciplines établies ?</td><td colspan="2">Choix des sources pour la collecte des termes</td></tr>
<tr><td>Domaine établi : encyclopédies, dictionnaire, littérature grise</td><td>Domaine émergent :

Sélection d'un corpus de documents</td></tr>
<tr><td>Quel type de documents seront traités ? Des monographies, des articles de périodique, des images fixes ou animées ?</td><td colspan="2">Le nombre de descripteurs et le niveau de spécificité varient en fonction de la nature du document indexé.</td></tr>
<tr><td rowspan="2">La production documentaire est-elle importante ? Y'a-t-il un fort accroissement des collections ?</td><td colspan="2">Niveau de développement du vocabulaire pour distinguer entre sources d'information de sujets voisins.</td></tr>
<tr><td>Oui : Vocabulaire développé</td><td>Non : Vocabulaire peu développé</td></tr>
<tr><td rowspan="2">L'utilisateur principal est-il un documentaliste ou un spécialiste de l'information ? un système automatisé de traitement de la langue naturelle ? le grand public ?</td><td colspan="2">Niveau de complexité de la structure du thesaurus</td></tr>
<tr><td>Oui : relations plus complexes</td><td>Non : Relations moins complexes</td></tr>
</table>

<table>
<tr><td rowspan="2">Le thesaurus est-il destiné aux spécialistes d'une discipline ? Au grand public ?</td><td colspan="2">Niveau de spécialisation des termes / choix des termes préférés et rejetés</td></tr>
<tr><td rowspan="2">Les questions soumises au système d'information sont-elles générales ou précises ?</td><td colspan="2">Niveau de spécificité du vocabulaire</td></tr>
<tr><td>Générales : Vocabulaire général</td><td>Précises : Vocabulaire spécifique</td></tr>
<tr><td>Les utilisateurs visent-ils le rappel ou la précision dans les résultats ?</td><td>Rappel : Vocabulaire général, moins de hiérarchies</td><td>Précision : Vocab spécifique, hiérarchies profondes</td></tr>
<tr><td rowspan="2">Le système d'information devra t'il traiter un grand nombre de requêtes ?</td><td colspan="2">Meilleure productivité du système</td></tr>
<tr><td>Oui : Vocabulaire étendu et spécifique</td><td>Non : Vocabulaire moins étendu</td></tr>
</table>

Tableau 2 - Informations données par l'étude de faisabilité sur la construction du thesaurus

4.2 Choix Méthodologiques inhérents au projet

Il y'a un socle commun à toute étude de faisabilité d'un projet de thesaurus. Ce socle sera développé dans la structure de cette partie afin de prendre conscience de l'utilité de considérer chacune de ces étapes avant de pouvoir les éluder. Cependant, il incombe au chef de projet d'effectuer des ajustements à cette méthode en fonction des spécificités de l'environnement, de la mission qui lui est confiée, ainsi que des contraintes pratiques liées à la chronologie du projet. Il est très vite apparu que le contexte original de la fusion des vocabulaires du CNC et de la cinémathèque française apporterait des ajustements dans l'organisation de ma mission, conduisant souvent à privilégier certaines étapes au détriment d'autres.

4.2.1. Un projet au sein du projet

La première incertitude venait du calendrier, et plus spécifiquement de la place du projet de thesaurus au sein du projet plateforme en cours. Au CNC comme à la

cinémathèque française, jeter les premières bases de l'architecture du futur portail était une priorité durant cet été 2016. Si les demandes d'entretiens et de précisions ont pu être correctement satisfaites, les cheffes de projet et les équipes impliquées dans la plateforme étaient occupées à régler d'autres questions plus opérationnelles, telles que la fusion des autorités ou l'implémentation des collections dans le portail. Ces priorités faisaient sens en ce qu'une vision globale sur les données et leur intégration est nécessaire avant de penser à leur accès. L'implémentation d'un thesaurus pouvait attendre le troisième jalon du projet, courant 2018, voire s'opérer plus tard, en marge du projet.

De ce fait, peu de discussions avaient eu lieu sur le sujet et les documents disponibles et diffusables étaient peu nombreux. Pour les mêmes raisons et pour cause de limitations de nombre dans les groupes de travail, je n'ai pas pu être intégré aux réunions sur l'élaboration du portail, qui auraient pu fournir plus d'éléments sur la manière dont le thesaurus pourrait y'être intégré. Les entretiens menés avec les cheffes de projet et avec le responsable du service AGDC du CNC me furent d'autant plus précieux pour acquérir la vue globale nécessaire à tout chef de projet embarqué dans une étude de faisabilité. Les équipes du projet attendaient mon arrivée pour poser les bases des discussions sur le thesaurus. Je me trouvais dans une position de défricheur qui allait fournir les premières pierres à l'édifice. Ainsi je devais intégrer à mes scénarios de faisabilité le plus possible de routes à emprunter, tout en prenant en compte les obstacles de chacune.

4.2.2 Harmoniser les vocabulaires

Lors de la phase préalable au projet, il est important de clarifier la demande, en distinguant le besoin exprimé du besoin supposé.

La mission qui m'était confiée consistait à *« étudier la faisabilité de la refonte et/ou de la fusion des vocabulaires contrôlés respectifs »* de la cinémathèque et du CNC. Cette absence de détermination entre une refonte et une fusion de vocabulaire était suffisamment vague pour intégrer toutes les possibilités d'adaptation des mots clés de Lise, du thesaurus *Cinédoc* et éventuellement du thesaurus Garnier. A ce niveau les différences de terminologie deviennent importantes. Il ne s'agissait donc pas d'une étude de faisabilité sur la construction d'un thesaurus, mais d'une étude de faisabilité sur le rapprochement de vocabulaires dans un contexte précis, qui comportait des questions sous-jacentes.

- Etudier les possibilités de fusion des termes des vocabulaires films et non films. Jusqu'à quel niveau et sur quels champs sémantiques cette fusion était-elle faisable ?

- Etudier les possibilités d'intégration au sein d'*Adlib* de deux vocabulaires séparés pour les films et les non-films, qui pourraient éventuellement être allignés conformément à la nouvelle norme.

- Dans le cas où cette option serait effective, étudier les possibilités de transformer les mots clés libres utilisés pour les films en un thesaurus / étudier l'évolution possible du thesaurus *Cinédoc* afin d'augmenter la faisabilité et la praticité de ces alignements.

Ces considérations allaient impacter l'allure du livrable final. Contrairement à une ingénierie de thesaurus qui se serait focalisée sur des préconisations liées à la faisabilité et au déploiement de l'outil, je devais orienter mon approche vers la faisabilité de différents scénarios et sur les possibilités d'interopérabilité des vocabulaires.

4.2.3 Une position d'expert extérieur

Mon positionnement à mi-chemin du CNC et de la cinémathèque française est particulier. Je travaillai dans les locaux de la bibliothèque de la direction du patrimoine du CNC et j'étais rattaché au service AGDC de cette même direction, à Bois d'Arcy. Ce fait s'explique par la mission de coordination qu'opère le CNC sur des institutions privées telles que la cinémathèque française et la cinémathèque de Toulouse[57], mais surtout par l'affectation au CNC des crédits financiers liés à la plateforme et au portail. Je me retrouvai donc en hiérarchie directe avec le CNC. Mais cela n'enlevait rien à ma position médiane. Les deux institutions attendaient du stagiaire chef de projet un avis « dépassionné » et objectif à même de faire progresser les débats. Cette fonction d'expert extérieur emportait plusieurs conséquences :

- Attribuer la même attention au CNC et à la cinémathèque française pour chaque phase de l'étude. Je fus donc amené à effectuer des déplacements à la cinémathèque française, dans les locaux de la rue de Bercy. La responsable du traitement des collections, ma référente au sein de l'institution, a permis de coordonner ces déplacements.

[57] Cette mission s'explique par les larges subventions accordées par l'Etat à ces institutions privées.

- Développer une autonomie dans ma méthode de travail, condition *sine qua non* pour conserver l'indépendance requise, mais aussi la confiance des deux institutions lors des entretiens. Cette position typique de l'assistant à maîtrise d'ouvrage fut facilitée par les équipes et par une note de mission que je transmettais à chacune des deux institutions dès le début de stage, dans laquelle je développais les raisons de ma présence, les étapes de ma réflexion et les moyens nécessaires pour la remplir (Annexe 1, p.109).

- M'assurer que le livrable rendu prenne en compte l'ensemble des vocabulaires et des collections des institutions partenaires dans les scénarios proposés.

- Ma position transverse faisait peu à peu une personne ressource tributaire d'une vue globale sur le projet qui devait être claire dans ma tête et je devais pouvoir la communiquer aux personnes qui me posaient des questions sur le thesaurus.

Un schéma préliminaire complété à mesure de l'avancement de ma mission, me fut très utile pour construire ma réflexion (Annexe 2, p.112).

4.3 Analyser le besoin

4.3.1 Pourquoi analyser le besoin ?

Le besoin des utilisateurs est toujours à l'origine de la volonté de concevoir et réaliser un produit, et ce fait ne souffre pas d'exception lorsqu'il s'agit de la constitution d'un thesaurus. Etudier le besoin des utilisateurs d'un thesaurus dans le cadre d'une structure, c'est connaître leurs caractéristiques, leurs missions quotidiennes, le type de recherches thématiques auxquelles ils sont confrontés lors de ces missions, mais aussi savoir recenser les difficultés rencontrées avec les outils existants dans le but de les corriger. Ainsi l'analyse de besoins effectuée dans le cadre de ma mission s'est très vite rendue complémentaire de l'étude des portails existants. Etudier le besoin, c'est aussi déjà se projeter dans le contexte dans lequel le thesaurus sera utilisé : En concevant un outil proche des besoins de ses utilisateurs, on facilite son appropriation, et dans une certaine mesure l'étape ultérieure de conduite de changement.

4.3.2 Définir une typologie d'utilisateurs

Afin d'optimiser cette analyse des besoins, il me fut nécessaire d'identifier des catégories d'acteurs, de créer des profils standards et d'évaluer les besoins en termes de recherche de chacune de ces catégories. Dans l'optique de cette analyse de besoins et à l'aide du schéma « enjeux » développé précédemment, j'ai établi une typologie des utilisateurs futurs du thesaurus, qu'ils soient liés à la plateforme, ou au portail (Annexe 1, p.109).

- Les professionnels de la documentation des institutions partenaires qui utiliseront le/les futurs thesaurus pour indexer les films ou non-films, et pour effectuer des recherches en interne, que l'on appellera **documentalistes**.

- Les équipes des institutions partenaires accédant au backoffice de la future plateforme, et qui pourront effectuer des recherches via l'outil pour des travaux internes, que l'on appellera **usagers internes**. Cette catégorie comprend les documentalistes qui effectuent des recherches pour des usagers internes.

- Les usagers des portails Lise et Cinéressources, du centre de documentation du CNC, ainsi que les autres usagers potentiels (grand public, chercheurs, étudiants en cinéma…), qui n'auront accès à la recherche et aux notices que par le portail. Nous les appellerons **usagers externes**.

4.3.2 Elaboration, conduite et synthèse d'entretiens

Les contraintes de durée de ma mission m'ont conduit à centrer la phase d'analyse des besoins sur une série d'entretiens individuels semi-directifs auprès des documentalistes et des usagers internes. Au nombre de trente et un, les entretiens se répartirent comme tels au niveau de chacune des institutions. Un entretien fut mené par ailleurs auprès de la directrice des collections de la cinémathèque de Toulouse :

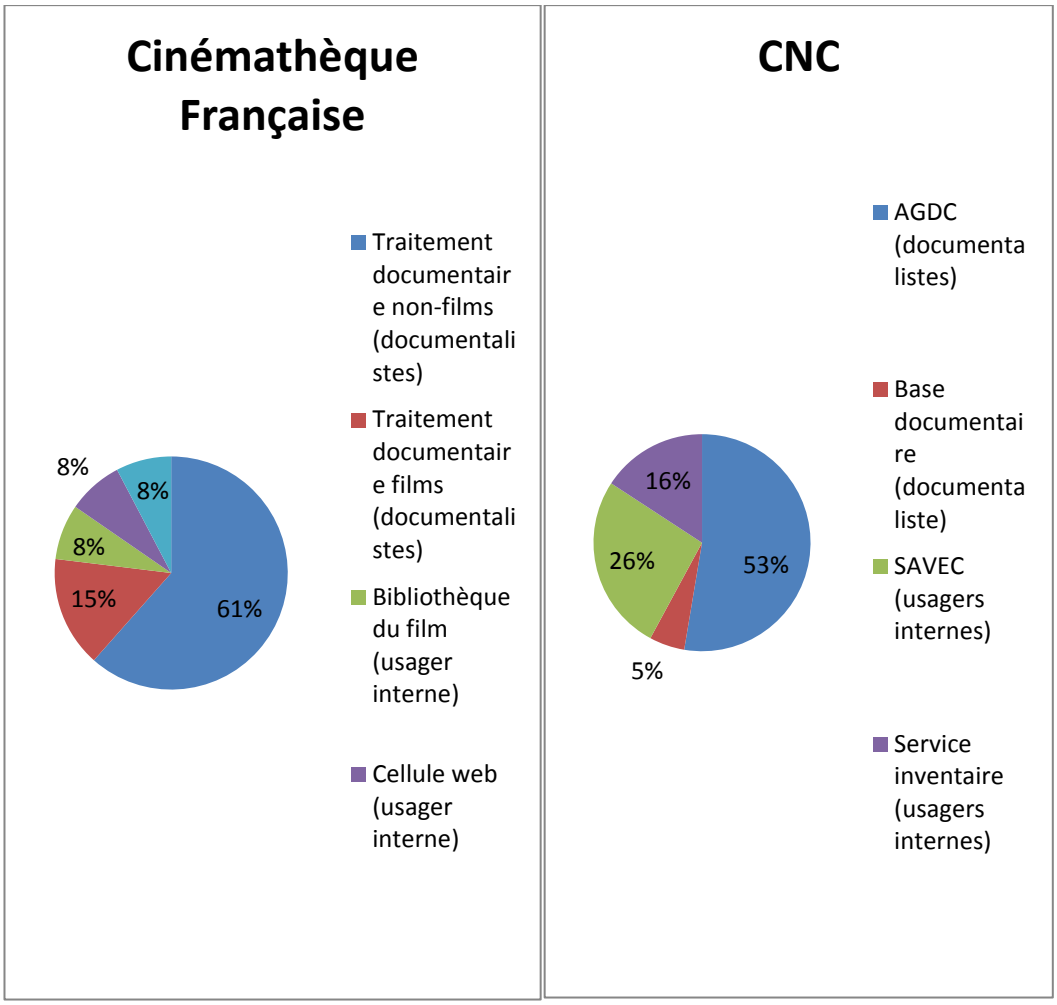


Figure 7 - Proportion des entretiens menés dans les institutions partenaires

Ces entretiens se déroulaient sur le lieu de travail de chacune des personnes interrogées. J'ai ainsi pu évaluer in situ la manière dont les documentalistes s'étaient appropriés Lise ou Ciné-ressources en leur suggérant de me faire une démonstration d'indexation de film ou de non-film le cas échéant. Les questionnaires étaient ciblés en fonction des profils établis (des exemples de questions se trouvent dans l'Annexe 1, p.109). Ils avaient pour but d':

- Obtenir des compléments sur l'historique des systèmes documentaires de la direction du patrimoine du CNC et de la cinémathèque française et l'impact de leur évolution sur les équipes. Ces premiers constats issus des entretiens m'ont permis de dégager des points de convergence entre le CNC et la cinémathèque.

- Evaluer la connaissance des outils actuels, l'appropriation des règles d'indexation par les documentalistes et cerner en particulier les besoins en

terme de présentation de langage contrôlé (nécessité de listes, d'arborescences).

- Evaluer le besoin de contrôle du langage et le niveau de structuration à donner au thesaurus.

- Approfondir les défauts constatés sur les outils actuels dans les recherches effectués par les documentalistes et les usagers extérieurs dans le cadre de leurs missions quotidiennes, identifier ces recherches. Dans une approche plus prospective, intégrer les contraintes de ces missions dans le futur outil permettrait de faciliter son appropriation.

- Identifier les stratégies des usagers extérieurs et des documentalistes pour pallier à ces manques, et éventuellement les prendre en compte.

Une synthèse des entretiens menés fut ensuite constituée, qui serait incorporé au livrable pour enrichir les scénarios de faisabilité. La durée de la mission étant trop courte, il n'a pas été possible d'effectuer une enquête quantitative sur les pratiques des usagers. Cette enquête aurait permis de dégager des profils plus précis du grand public potentiel susceptible de fréquenter ces institutions, et qui est le public cible du portail. Une analyse des requêtes effectuées sur les portails existants (type de requêtes, descripteurs utilisés, champs thématiques utilisés etc…) à l'aide de statistiques aurait également pu compléter cette étude en mettant en exergue les recherches récurrentes comme autant de hiérarchies à creuser. Elle aurait pu permettre de mieux ajuster le vocabulaire aux usagers externes. Certains modules statistiques liés aux portails permettent de sortir ces renseignements, mais il n'a pas été possible d'obtenir ces informations dans *Lise* et *Ciné-ressources*.

4.4 L'importance d'établir un état de l'art des langages documentaires

« Le responsable de l'étude de faisabilité doit consulter de nombreuses sources pour établir une liste de langages documentaires compatibles ou potentiellement utiles comme base terminologique » [30, HUDON]. Cet état de l'art peut être très instructif pour l'équipe qui sera amené à composer le thesaurus et peut même le cas échéant dispenser de la création d'un thesaurus tout en remplissant le besoin. Il est par ailleurs une constante dans les normes sur la conception de thesauri de s'inspirer de l'existant.

Même si un thesaurus a été choisi, il faut élargir la recherche pour inclure les schémas de classification et toutes les listes organisées de termes, vedettes matières ou autres qui couvrent au moins en partie le domaine à décrire. Ce recensement peut aller jusqu'aux schémas de données disponibles sur le domaine sur le site « schéma.org », qui peuvent aider à structurer l'outil[58]. Les classifications serviront non seulement de base terminologique, mais également d'appui structurel pour la construction d'un thesaurus. Elles peuvent entre autres appuyer les choix pour établir les champs sémantiques. Les bases documentaires existantes peuvent également utiliser des classifications comme une base de construction de thesaurus, et ceci est particulièrement vrai lorsque le thesaurus est encyclopédique[59]. Le thesaurus que le *British Film Institute (BFI)* a intégré au logiciel *Adlib* (voir 5.2.4) fut construit à partir d'une adaptation des termes de la classification décimale universelle.

Les vedettes matières et thesauri existant dans le domaine qui nous intéresse pourront servir de réservoir de mots clés simples et composés, qui pourraient être inclus dans le nouveau thesaurus. Après évaluation, ces mêmes thesauri existants peuvent s'emboîter au thesaurus existant. Cette pratique d'utilisation de branches de différents thesaurus est déjà institutionnalisée. D'après Sylvie Dalbin qui a expertisé le thesaurus *Cinédoc* de la cinémathèque française, *« cela fait déjà plusieurs années que des centres exploitent des thésaurus propres à leur centre construits "officiellement" partir de plusieurs thésaurus. Il pourrait donc y'avoir des pistes pour une extension de Cinédoc à partir de branches/domaines d'autres thésaurus pour lequel il y aurait des manques. Même sans web de données cela était possible et réalisable. Mais sachant que les thésaurus vont se déployer (pour les plus importants) sous peu avec des URI pérennes, c'est une éventualité tout à fait jouable»*[60].

Enfin, il est tout à fait dispensable de *« réinventer la roue »* si un thesaurus existant et parfaitement fonctionnel peut décrire les collections. Il pourra éventuellement faire l'objet d'adaptations à mesure des indexations de l'équipe de documentalistes. Il est d'ailleurs globalement conseillé d'utiliser des thesauri existants pour des spécialités communes, ne serait-ce que pour aller dans le sens de l'interopérabilité en harmonisant les vocabulaires des différentes structures.

[58] https://schema.org/Movie

[59] Un thesaurus encyclopédique ne reflète pas une spécialité, mais couvre de manière générale l'entièreté des connaissances.

[60] Partie d'un échange de mails effectués avec Sylvie Dalbin.

4.4.1 Collecter les langages existants

Deux sites m'ont été particulièrement utiles pour effectuer cet état des lieux des langages documentaires :

- Le *« thesauro-annuaire »* du site tard.bourrichon propose une liste de thesauri mis en ligne classés par domaines de spécialité : http://www.tard-bourrichon.fr/documents/THESAURUSonline/tab/Thesauro-annuaire.html

- L'annuaire dmoz qui répertorie les thesauri et listes d'autorité francophones. http://www.dmoz.org/World/Fran%C3%A7ais/R%C3%A9f%C3%A9rences/Th%C3%A9saurus/

La liste de thesauri établie par Lorraine Keller en annexe de son mémoire sur la réingénierie de thesaurus compléta ces ressources [32, KELLER, p.147].

Je me suis également aidé des suggestions des documentalistes experts du cinéma. Le blog *« descripteurs »* de Sylvie Dalbin a permis d'établir un état des lieux exhaustif sur l'évolution récente des langages documentaires (31, DALBIN).

Sur la base du besoin établi, le périmètre d'évaluation s'étendait aux :

- Les thesauri spécialisés dans le cinéma pour les collections non-films.

- Les thesauri encyclopédiques, pour élargir aux thématiques hors du cinéma.

- Les thesauri géographiques et historiques.

4.4.2 Evaluer les langages collectés

Une fois les langages existants collectés dans le domaine de spécialité, il faut évaluer si ces langages correspondent au besoin. Michèle Hudon suggère d'établir cette évaluation d'après les critères suivants [30, HUDON et Al, p.76] :

- La capacité de l'outil à décrire tous les concepts qui constituent l'environnement sémantique que l'on désire représenter.

- Sa richesse lexicale. Dans le cas d'un système documentaire possédant déjà un vocabulaire d'indexation, le chef de projet en charge de l'étude de faisabilité pourra s'appuyer sur une étude des termes les plus employés à la recherche et des occurrences d'indexation pour les mots clés employés.

- La structure relationnelle. Le réseau créé par les relations sémantiques est-il complet et représentatif des points de vue acceptés dans l'environnement d'utilisation du thesaurus ?

- La conformité du lexique thésaural et de la structure relationnelle aux normes.

Les outils présentant un intérêt eut égard à ces critères d'évaluation doivent ensuite passer le cap d'une évaluation pratique qui consiste à indexer un échantillon représentatif de films et de non-films afin de vérifier l'utilité des outils dans l'environnement prévu.

Au niveau des thesauri spécialisés dans le cinéma, le thesaurus *Cinédoc* fait autorité faute de combattants. Il existe néanmoins le thesaurus de la FIAF, utilisé à la fois par la bibliothécaire de la direction du patrimoine du CNC et les documentalistes des cinémathèques (cinémathèque française, Toulouse etc…) pour indexer des périodiques. Il s'agit donc d'un thesaurus spécialisé, non francophone et utilisable dans un environnement très limité[61].

Les thesauri encyclopédiques, qui sont plus à même de décrire les films, sont déjà plus présents en ligne.

4.5 Analyser l'existant et l'environnement documentaire

4.5.1 Utilité de l'analyse de l'existant

L'analyse de l'existant d'une structure est le préalable à tout projet documentaire. Elle sert à qualifier les documents qui seront intégrés au nouveau système d'information et à établir quels ont été les gains et les manques du système d'information précédant. Dans le cadre d'un projet d'élaboration de thesaurus, cette étape concerne à la fois les bases de données existantes et les collections que le thesaurus devra décrire. Elle permettra au chef de projet :

- De réunir des données statistiques qui seront utiles à la décision.

- D'établir le périmètre du thesaurus, les domaines qu'il devra couvrir.

[61] Le thesaurus de la FIAF est disponible sur le web : http://www.fiafnet.org/pages/E-Resources/PIP-Subject-Headings.html?PHPSESSID=ps5nudk09b8frhua0oqu0l7503

- D'évaluer le nombre et le type des collections qui seront intégrées dans le futur système, ainsi que leur accroissement prévisionnel et d'ajuster le thesaurus en conséquence. Le résultat de cette évaluation est en annexe 4 du mémoire (p.131).

- Le cas échéant, de prendre en compte les langages documentaires existants en réfléchissant à ce qui sera conservé dans le nouveau thesaurus.

- De ne pas renouveler les défauts liés à ces langages documentaires.

Inscrire cette étape après l'état de l'art ou de manière concomitante permet de ne pas se laisser influencer par les langages documentaires existants au sein de la structure. Il peut-être tentant de conclure que les langages existants ont déjà fait leur preuve et qu'ils pourraient constituer l'unique base de travail pour le futur thesaurus. Cependant, leur analyse à l'aune de ce qui existe à l'extérieur de la structure peut révéler des défauts qui n'auraient pas été relevés ou permettre d'appliquer des alternatives.

Ainsi la question de savoir si les mots clés libres décrivant les films dans la base *Lise* est-elle intervenue très vite au cours de ma mission. Ces 8200 mots clés avaient précédemment fait l'objet d'une liste ordonnée hiérarchique, établie par la documentaliste du centre de documentation du CNC. Cette liste ordonnée était parvenue à fusionner des termes proches et à dégager les champs sémantiques suivants : *Découverte du monde, Histoire et Politique, Industrie et Artisanat, Nature, Santé, Sciences et technologies, Société, Sports et Loisirs, Personnages*.

Lors d'un entretien, la documentaliste me confirma que cette classification des mots clés provenait en grande partie des catégories prédéfinies par les champs « thème principal » et « thème secondaire » accessibles au public pour filtrer les recherches sur le portail actuel de *Lise* (voir 3.1.2.2). Par ailleurs, les mots clés ordonnés avaient été créés au fur et à mesure de l'indexation des collections. Ils semblaient donc représentatifs des collections films du CNC et des cinémathèques partenaires de *Lise*. Ainsi il était tentant de s'appuyer sur cette liste hiérarchique et les mots clés existants afin de construire un thesaurus spécifique pour les films. L'état de l'art révéla par la suite qu'un thesaurus encyclopédique existant pouvait être une solution plus adéquate pour indexer les films. Bien qu'elle ne s'appuie pas sur une base préexistante, cette solution paraissait plus avantageuse en ce qu'elle dispensait du coût de construction d'un nouveau thesaurus et que des thesauri

encyclopédiques disponibles en ligne couvraient très bien les domaines thématiques de la liste ordonnée. Cette possibilité ne devait donc pas être éludée par les travaux déjà effectués en interne.

4.5.2 Des outils pour auditer les langages contrôlés « actuels »

4.5.2.1 Les occurrences d'indexation des termes

Il s'agit du nombre de documents indexés pour chaque terme présent dans le langage d'indexation. Des modules documentaires connexes aux systèmes de gestion de bases de données permettent de sortir ces chiffres. Ils seront très utiles pour effectuer des fusions de terme, mais aussi pour évaluer le niveau de sous-emploi de certains termes comme de la facture d'ensemble des mots clés.

- 2911 mots clés, soit 35% du total, n'ont qu'une seule occurrence.
- 1537 mots clés ont plus de 10 occurrences, soit 18,7% du total.
- 5 mots clés ont plus de 500 occurrences. Le plus utilisé est le terme « tourisme », avec 657 occurrences.

L'étude des occurrences des mots clés de *Lise* permit de constater qu'il y'avait une trop grande dispersion de l'indexation et que dans certains domaines, les termes étaient trop spécialisés. Paradoxalement, il se dégageait une poignée de termes « passe-partout » très employés, mais qui ne permettaient pas de faire une recherche satisfaisante en ce qu'ils étaient trop généraux.

4.5.2.2 Des outils qualitatifs

- Les entretiens. Les entretiens établis pour l'analyse des besoins des équipes intégrèrent des questions sur l'existant. Ces questions dégagèrent les points de vue des équipes indexeurs sur leurs outils actuels et bien souvent, les critiques formulées par les documentalistes permirent de constater qu'un langage structuré et contrôlé était la solution. La cinémathèque française avait par ailleurs résolu ces problèmes dans la description de leurs fonds non-films en utilisant le thesaurus *Cinédoc*.

- Des audits établis sur les langages existants par des experts.

- L'indexation d'une sélection de documents. Lors du rapprochement des langages documentaires de deux institutions, il convient d'évaluer les possibilités de rapprochement. Ces rapprochements peuvent être faits à l'aide d'indexations croisées de la première collection avec la seconde

(exemple : indexer une trentaine de documents films avec le thesaurus non-film *Cinédoc*).

4.6 Etablir un benchmark, difficultés et atouts

Dans la ligne de l'état de l'art des langages documentaires de la discipline, il a paru utile d'établir un benchmark de l'existant en termes d'intégration de langages documentaires. Il s'agissait de chercher sur quelle base les institutions cinématographiques d'autres pays qui possédaient des portails fédérés géraient leur recherche thématique ? Avaient-elles intégrés des mots clés thématiques à leur portail ? L'enjeu de cette question était la possibilité de découvrir des thesaurus incorporant films et non-films existant dans d'autres langues, mais surtout de pouvoir s'appuyer sur des expériences déjà menées qui avaient prouvé les forces et les faiblesses d'une ou de l'autre des méthodes.

Le premier niveau de ces recherches était lié aux portails fédérant plusieurs institutions. Pour cela, je m'aidais du répertoire des membres de la FIAF[62]. renvoyant aux portails web des cinémathèques et organes de gestion patrimoniale des films dans les différents pays. Au terme de mes recherches, je ne trouvais pas de portail fédérant les collections de plusieurs institutions. Le président de la FIAF, également chef du service du SAVEC, me confirma qu'il n'en existait pas. La fusion en cours de *Lise* et de *Cinéressources* était un cas unique au monde. Il me restait donc à explorer la piste d'une seule et même institution possédant un portail qui donne un accès thématique à la fois aux films et aux documents liés aux films.

Le portail d'accès aux collections du *British Film Institute (BFI)* me fournit de nombreux indices qu'il correspondait à ma recherche, mais il m'était impossible de déterminer si un thesaurus se dissimulait derrière les possibilités de recherche thématiques qu'il offrait sur toutes les collections. Le chef du département Data du BFI put me donner des précisions, ainsi que d'autres informations très utiles à la mission :

- La confirmation qu'un thesaurus était bien au centre de cette indexation. Ce thesaurus avait été de surcroit construit via la base Thesaurus d'*Adlib*, le même outil qui était utilisé par le CNC et la cinémathèque française.

[62] Le répertoire est disponible à cette adresse : http://www.fiafnet.org/pages/Community/Members.html?PHPSESSID=b8i23o5v54cftgouvlnb9sjn76

- Les équipes du *BFI* travaillaient déjà depuis plusieurs années à une unification de la recherche thématique sur toutes les collections via la base Thesaurus d'*Adlib*. Cela faisait donc une expérience sur laquelle le CNC et la cinémathèque française pouvaient se baser pour construire un vocabulaire commun.

- En tant que co-développeur sur la version cinéma d'*Adlib* que j'étais chargé d'étudier, il fournit des renseignements supplémentaires sur les fonctionnalités thesaurus de l'outil.

4.7 Analyser les contraintes

Les contraintes internes pouvant impacter la gestion d'un projet de thesaurus doivent être listées au niveau de l'étude de faisabilité. Les critères suivant seront essentiels à l'évaluation des scénarios finaux qui permettront une prise de décision et dans certains cas, une subvention du projet. La Faisabilité de la construction d'un thesaurus en termes de ressources humaines évalue la disponibilité du personnel, les compétences requises et le nombre de personnes pour boucler le projet. Elle peut se concrétiser dans organigrammes de tâches (OT). Les personnels sélectionnés pour le groupe de travail devront disposer de temps libre pour pouvoir mener à bien les différentes phases de construction du thesaurus. La faisabilité en termes de temps peut être définie à l'aide d'un échéancier, si possible précisée par un diagramme de GANTT qui établira les différentes étapes et les dates d'échéance de ces étapes dans la construction d'un thesaurus. Bien qu'envisagées au début de la mission, ces référentiels très utiles ne furent pas développés en raison de l'abandon progressif de la solution d'un thesaurus unique pour le CNC et la cinémathèque française. Adapter ces outils à l'organisation d'alignements aurait demandé une durée de mission plus importante.

La Faisabilité technique évalue la productibilité. Il s'agit de recenser les équipements informatiques, infrastructures et compétences logicielles disponibles pour mener à bien le projet de thesaurus. Des logiciels pourront très bien épauler l'équipe qui va construire le thesaurus. Il faut aussi déterminer à quel niveau les services informatiques pourront être impliqués dans le projet.

La Faisabilité économique a pour but de calculer le coût provisionnel attribué au projet, de le budgétiser. Il s'agit non seulement d'évaluer les coûts de réalisation du thesaurus, mais aussi les coûts de gestion et de maintenance.

5. Des perspectives pour la construction et la gestion d'un thesaurus films et non films

5.1 Le livrable de faisabilité, concrétisation de l'étude

Il convient désormais de porter un regard sur les préconisations délivrées pour ce cas. Il sera également nécessaire de donner des clés d'action au chef de projet qui poursuivra la mission à travers des perspectives de gestion et de maintenance du ou des thesauri constitué.

5.1.1 Objectifs et structure du livrable de faisabilité

Le livrable rendu à l'issue de cette étude était un outil pour orienter les discussions qui auraient lieu entre les partenaires et il devait être adapté en conséquence. Les choix sur sa présentation et son contenu se fondèrent sur des lignes directrices établies dans le tableau ci-dessous, chacun des objectifs ramenant à une consigne de présentation du livrable afin de le rendre le plus opérationnel possible. Ces lignes directrices se remplirent à mesure des entretiens formels ou informels qui survinrent et des ajustements liés à la mission.

Objectifs	Présentation du livrable
Laisser ouvertes les discussions sur chaque voie de rapprochement possible, quitte à les écarter ensuite faute de faisabilité (ce fut le cas du scénario de fusion).	• Présentation sous forme de scénarios distincts • Développement suffisamment important de chacun des scénarios • Si certaines interrogations n'ont pas obtenu des réponses, les inclure dans l'étude pour alimenter les débats.
Du fait de mon absence pour la suite du projet, donner une base de réalisation des scénarios pour les équipes opérationnelles.	• Développer des recommandations pour les rapprochements de vocabulaires : champs sémantiques communs, pertinence des rapprochements, association à des champs particuliers dans la future base. • Donner une base pour la réalisation d'un thesaurus, dans le cas où cette option serait choisie. • Etablir une bibliographie des documents nécessaires pour la suite du projet.
Etre suffisamment concis et compréhensible pour être compris des instances décisionnelles.	• Etablir un glossaire des termes • Séparer les explications techniques des scénarios des outils de prise de décision.

Permettre de prendre des décisions rapides et éclairées	• Exposer en préambule les enjeux opérationnels d'un rapprochement des indexations thématiques films et non-films • Lister et développer les coûts et les avantages des différents scénarios, ainsi que les conséquences sur la gestion du thesaurus.
Etablir une opinion claire sur la faisabilité des scénarios	• Donner une évaluation de la faisabilité de chacun des scénarios via un indice de faisabilité. *Un SWOT aurait été nettement plus efficace.*

Tableau 3 - Lignes directrices sur l'orientation du livrable

5.2 Vue sur les préconisations

A l'issue de cette étude de faisabilité, trois scénarios pour le rapprochement des vocabulaires films et non films de *Lise* et *Ciné-ressources* ont été dégagés et évalués. Nous exposerons les deux premiers scénarios dans leurs grandes lignes, tout en veillant à mettre en évidence les possibilités qu'ils offrent pour un accès thématique croisé aux collections. Puis nous nous attarderons sur le troisième scénario en ce qu'il propose un exemple opérationnel d'alignement via un logiciel documentaire.

5.2.1 Scénario 1 : Fusion des vocabulaires films et non-films

Le premier scénario développé écarte l'hypothèse d'une fusion des vocabulaires films et non-films. Cette solution n'a pas été retenue pour les raisons suivantes :

- L'indexation thématique des non-films nécessite un thesaurus spécialisé dans le domaine du cinéma du type de *Cinédoc*.

- Les films, qu'ils soient fictions, non-fictions ou documentaires, s'attachent à décrire le réel. Ils devront emprunter idéalement un thesaurus encyclopédique qui serait adapté à l'étendue des sujets des films indexés.

- Seuls les films documentaires sur le cinéma, très peu nombreux dans les collections, pourront admettre des descripteurs proches de ceux utilisés pour les collections non-films.

Pour le reste des collections, un méta-thesaurus hybride pourrait être constitué à des fins utiles à condition qu'il établisse des passerelles entre la description de films

et de non-film. En d'autres termes, que des champs spécifiques d'indexation puissent utiliser une partie de ce thesaurus pour décrire les deux familles de documents. Ainsi, il pourrait être parfois utile qu'une requête d'un utilisateur renvoie à la fois à des films et des non-films sur un thème précis. Voici un exemple concret de passerelle :

« Histoire du cinéma /mouvement cinématographique » est un champ sémantique du thesaurus non-film Cinédoc. Si l'on veut évaluer l'utilité d'une passerelle avec les films, il faut se poser la question : En quoi une période de l'*histoire du cinéma* ou un *mouvement cinématographique* peut-elle être le sujet d'un film non documentaire ?

Les films relèvent bien de mouvements et de courants cinématographiques, mais on ne peut pas indexer ces mouvements en sujet du film. La nouvelle vague n'est pas le sujet de « A bout de souffle ».

Il serait néanmoins utile dans le cadre d'une recherche sur un courant d'illustrer par des résultats films des recherches portant sur de la documentation sur *la nouvelle vague*, sur le *cinéma vérité* etc…

Ainsi il est possible de recommander :

- Un champ sujet pour les documentaires sur le cinéma et pour les non-films.

- Un champ thème de valorisation pour rattacher les films de fiction qui ont été décisifs à la constitution d'un mouvement cinématographique ou qui sont emblématiques d'un courant (exemple : *Festen* de Thomas Winterberg pour 'le Dogme'). L'indexation pourrait être effectuée par les documentalistes, ou a posteriori par les organisateurs d'évènements.

Greffer le vocabulaire cinéma utilisé par les non-films sur un thesaurus général serait néanmoins un gonflement artificiel d'une efficacité relative. Il en résulterait des branches disproportionnées en termes de descripteurs et de relations. De plus, il serait d'une grande complexité à manier pour les documentalistes indexeurs. Toutes les étapes de production du thesaurus devraient être mutualisées par les deux institutions. Enfin, aboutir à un tel thesaurus revient à détruire l'intégrité du thesaurus Cinédoc pour le recomposer dans un outil moins pratique et il déposséderait la cinémathèque française et les cinémathèques partenaires de la gestion exclusive des documents non-films.

5.2.2 Scénario 2 : Alignement des vocabulaires films et non-films via un outil externe

Contrairement au premier scénario, nous nous trouvons ici dans une configuration où chacun des langages conserve son intégrité. Il n'est pas question de fusionner les termes en un méta-thesaurus, mais d'effectuer des alignements de termes afin de favoriser l'interopérabilité entre les vocabulaires.

Chacun des vocabulaires d'indexation sera donc entré dans un outil de gestion de thesaurus qui effectuera le rapprochement des termes, créant des liens d'équivalence inter-thesaurus. Ces liens d'équivalence pourront s'établir à l'indexation ou à la recherche. Bien que les alignements s'opèrent via des programmes spécifiques, une médiation humaine est toujours nécessaire, en particulier pour tous les termes qui ne sont pas d'équivalences exactes (voir 2.2.3.4).

Pour ce scénario, il est nécessaire d'avoir recours à un outil de création de thesaurus qui possède les fonctions d'alignement. Certains nouveaux outils, tel que Ginco (voir 2.3.2) ou OpenTheso[63] ont été créés sur la base de la norme ISO 25964-2. Ils intègrent donc les fonctions d'alignement et permettent l'interopérabilité avec d'autres langages documentaires. Il est également nécessaire de mobiliser des équipes pour piloter les alignements de termes qui n'ont pas d'équivalence exacte.

La pertinence de ce scénario est à étudier à l'aune de l'utilité qu'apporterait aux usagers des résultats incorporant des films et des non films dont les référentiels d'indexation sont différents (voir 3.3). Les alignements ne sauraient se borner à l'étendue du thesaurus car il est peu pertinent en terme d'indexation et de recherche de rapprocher de manière globale les termes films et non-films. Afin de maximiser leur pertinence, les passerelles établies dans le premier scénario peuvent constituer une base de champs sémantiques au sein desquels les alignements seront possibles. Il serait ensuite nécessaire d'examiner les termes et les concepts de ces champs plus en détail afin de les rapprocher. En se basant sur ces passerelles et l'étude de chacun des vocabulaires existants, voici quelques exemples de champs sémantiques de Cinédoc pouvant accueillir les alignements de vocabulaire

[63] https://masa.hypotheses.org/99

1. Vocabulaire Histoire du cinéma, mouvements cinématographiques (cinédoc) / thesaurus film
2. Vocabulaire Typologie et genre (cinédoc) / thesaurus film
3. Thésaurus film / Thème cinématographique (cinédoc) / Thesaurus iconographique
4. Thésaurus film / Héros et personnages (cinédoc) / thesaurus iconographique
5. Technique (cinédoc) / thesaurus film
6. Liste des périodes et événements (cinédoc)/ thesaurus film / thesaurus iconographique
7. Liste des pays (cinédoc) thesaurus géographique sujet / thesaurus film / thesaurus iconographique

Afin de faciliter l'intégration des termes à un outil de construction et de gestion de thesaurus, il est nécessaire de se placer dans le cadre de la norme ISO 25964-2 : 2013 [25, NISO]. Un thesaurus film qui suive les recommandations de la première partie de la norme serait préférable afin de s'en servir comme référent pour effectuer les alignements en « hub » (voir 2.2.3.2). Ce thesaurus film pourrait être un thesaurus général existant adapté aux fonds films ou un thesaurus hybride. Un thesaurus *Cinédoc* retravaillé selon la norme peut également servir de référent.

Bien que permettant une ouverture à d'autres domaines dans le cadre de l'ouverture des données culturelles (voir 2.3.2), Cette solution se révèle lourde en termes de gestion de projet, en termes de serveurs et de compétences logicielles mobilisée. Elle est aussi inédite dans le domaine, même si elle peut recevoir l'expertise du ministère de la culture, qui pourrait lui apporter son soutien dans le cadre du programme Hadoc.

5.2.4 Scénario 3 : Alignement des vocabulaires films et non-films via l'outil Adlib

On complétera ce scénario par le guide d'utilisation du logiciel, en annexe 3 du présent mémoire (p.114).

La base thésaurus d'*Adlib* est construite comme un thesaurus global qui :

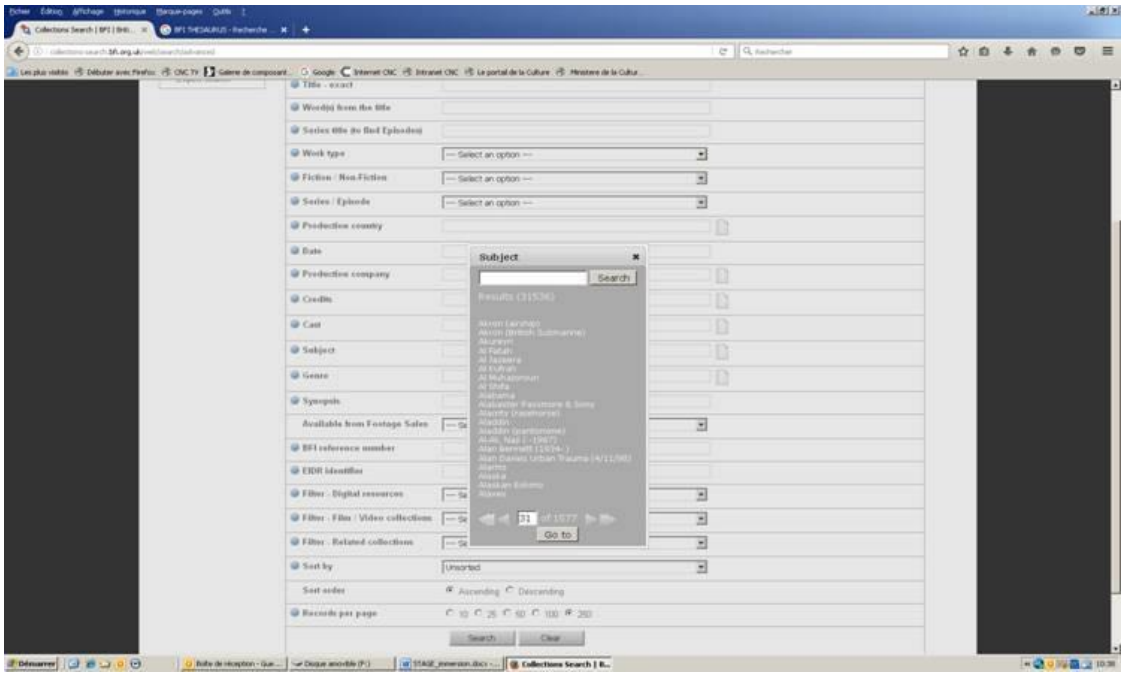
- Accueille plusieurs domaines, comme autant de listes ouvertes.
- Propose des liens d'équivalence, de hiérarchie et d'association entre tous les termes dont la fiche est créée dans la base, y compris des termes de listes différentes.

- Il inclut la polyhiérarchie, ce qui offre la possibilité de créer des facettes indépendantes des relations hiérarchiques du thésaurus.

- Par l'intermédiaire d'une API, permet de créer une interface vers les collections vers un champ de recherche thématique. Ce champ est transverse. Il pourrait permettre d'effectuer des recherches à la fois sur les films et les non-films.

Dans ce scénario, nous nous basons sur l'expérience des équipes du *British Film Institute* qui furent les premiers à intégrer des collections cinéma au logiciel.

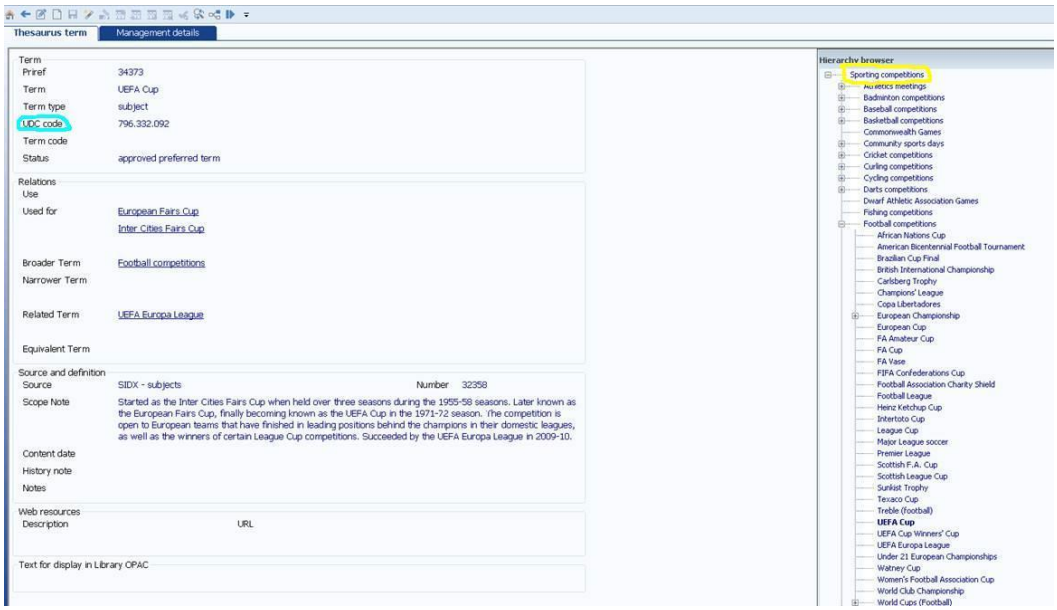
5.2.4.1 Intégration de thesaurus dans Adlib : l'exemple du British Film Institute



Champ sujet de la recherche avancée films de l'interface Adlib du British Film Institute [64]

Les mots clés présents dans le champ sujet de la recherche avancée ont été entrés dans la base thesaurus *Adlib* du *British Film Institute* et complétés des relations entre les termes. Chacun dispose donc d'une fiche au sein de la base Thesaurus. Cette fiche accueille l'identifiant du terme, son domaine (exemple : sujet), son statut (descripteur, candidat…), les relations du terme avec les autres termes, une note d'application à l'attention des indexeurs, mais aussi la source du vocabulaire, s'il appartient à un thesaurus ou une classification externe.

[64] http://collections-search.bfi.org.uk/web/search/advanced



Exemple de fiche descripteur dans la base Thesaurus Adlib du BFI

Aidés par l'arborescence affichée à droite de l'image, nous pouvons constater que le top term « Sporting Competitions » est lié au domaine « subject ». Ce domaine accueille trois niveaux de termes spécifiques, mais il pourrait en accueillir plus. « Sujet » peut-être assimilé à un champ sémantique.

5.2.4.2 Application aux vocabulaires de *Lise* et *Cinéressources*

A travers cet exemple, nous voyons qu'il est possible de créer un meta-thesaurus de toute pièce au sein d'Adlib, après en avoir établi les termes et les hiérarchies. Ce thesaurus comprendrait toutes les listes ouvertes ainsi que les thesaurus intégrés dans l'outil. Il est également possible d'importer des thesauri entiers au sein de la base Thesaurus. De ce fait, un thesaurus externe tel que le thésaurus Garnier pourrait être importé si un lien est créé dans Adlib.

Ces possibilités offrent la voie à une intégration à Adlib des termes non-films de *Cinédoc* et de leurs relations ainsi qu'à une intégration des termes de *Lise*. La construction d'un thesaurus film est néanmoins préconisée afin de pouvoir utiliser toutes les possibilités de l'outil. Comme le montre l'image ci-dessus (champ Source UDC Code de la notice), le BFI a construit son thesaurus à l'aide de la classification décimale universelle en l'adaptant à mesure des indexations. La piste de la CDU est recommandable pour une base de travail, mais il s'agit à la base d'une classification, qui doit de ce fait être fortement adaptée (seules les relations

hiérarchiques sont véritablement exploitables). Il faut donc tendre vers un langage documentaire spécifique pour les films. Il faudra adapter ce thesaurus aux fonds films existants. Dans le cas du BFI, ce travail d'adaptation s'est fait sur plusieurs années et il n'est pas encore terminé.

Le thesaurus iconographique Garnier peut être intégré en le liant à la base thesaurus *Adlib* comme thésaurus externe (Sur la capture, source=Garnier à la place de UDC code). Il est ensuite possible de créer des alignements entre les termes de ces thesauri.

5.2.4.3 Exemple d'alignement

Ce scénario nécessitera de créer de multiples domaines et d'établir un méta-thesaurus[65] au sein de la base thesaurus d'Adlib. Il comprendrait d'un côté les domaines non-films éclatés, de l'autre les domaines film, et ensuite lier le tout par des relations. Ainsi les alignements entre les différents langages seraient effectués à travers Adlib via des équivalences de termes. Ils pourraient se baser sur les alignements préconisés dans le scénario 2.

Exemple de creation d'alignement via la relation d'équivalence:

Perception sensorielle (domain=thème cinématographique, source = Cinédoc) <-> equivalent_term <-> Film representation (domain= fonction sensorielle, source = Motbis)

Cette relation d'équivalence peut être utilisée pour la recherche afin d'être sûr que des enregistrements soient retournés à la requête de l'usager, qu'il utilise le terme fonction sensorielle ou perception sensorielle.

Est-il possible de constituer à terme un thesaurus pour les films et non-films au sein d'*Adlib* ?

Il est à noter que dans le cas du *BFI*, les termes de sources différentes ont d'abord été importés dans les domaines spécifiques des collections (posters, books…), avant de s'attaquer à un domaine général sujet. Ce domaine est encore en cours d'élaboration.

[65] Terme utilisé pour le incorporer les vocabulaires des différents thesauri et les liens entre termes de ces vocabulaires.

5.3 Perspectives de gestion et de maintenance pour les services documentaires des partenaires

5.3.1 Intégrer le thesaurus à un outil de gestion

Au cours de sa vie, un thesaurus est amené à évoluer. Il accueillera de nouveaux descripteurs, abandonnera certains autres, se ramifiera de nouvelles branches et parfois de nouvelles relations (c'est notamment le cas lors d'un fort taux de renouvellement documentaire). Il convient de s'assurer, dès l'étude de faisabilité, que cette évolution pourra être assurée dans les meilleures conditions. Les logiciels de gestion et de maintenance de thesaurus permettent de faciliter l'utilisation et la gestion sur le long terme du thesaurus. Ils peuvent être autonomes ou constituer un module d'un logiciel existant [32, DALBIN]. A la construction du thesaurus, ces outils permettent des validations automatiques de termes et de relations. Comme nous l'avons vu pour les scénarios, ils peuvent aussi permettre, pour certains d'entre eux, d'intégrer plusieurs thesauri et de constituer des relations entre eux, à l'indexation ou à la recherche.

Une étude de faisabilité sur un thesaurus intégré à un portail ne peut donc faire abstraction d'une évaluation du module thesaurus de l'outil de gestion qui accueillera les données accessibles depuis le portail. Si des insuffisances eut égard aux besoins de maintenance du thesaurus ont été constatées au terme de cet audit, il convient de préconiser un logiciel autonome ou le développement particulier d'un logiciel. Ce tableau établit les fonctions principales auditées sur le logiciel *Adlib*, tout en les reliant aux scénarios. A partir de cette analyse, il a été décidé qu'un logiciel de thesaurus autonome n'était pas nécessaire. Il serait néanmoins requis dans le cas où le scénario 2 serait préféré.

Besoin constaté	Fonction Testée	Réponse d'Adlib
Familiarisation nécessaire des documentalistes indexeurs avec le thesaurus.	Présentation ergonomique du thesaurus à l'indexation	OK. Présentation arborescente claire à partir du module. Il y'a une interaction directe entre le thesaurus et les opérations d'indexation.
	Production de divers types d'affichage des données, sous forme imprimée ou numérique.	OK. *Adlib* offre la possibilité de créer des fichiers d'export pour imprimante ou fichier. Dans le standard, il y a un export vers une liste alphabétique et une vue hiérarchique. Il est possible d'ajouter des formats d'export. Il est aussi possible de configurer l'outil *Office Connect* pour que les indexeurs peuvent

		avoir les listes et recherches directement dans Word / Excel.
Evolution du	Faciliter la gestion des candidats descripteurs Scénarios 1,2,3	OK. Gestion des statuts des descripteurs avec six statuts possibles dont un statut rejeté pour éviter la réutilisation du descripteur.
	Intégrer les termes et les concepts	Evoluer vers un logiciel autonome intégrant la norme ISO 25964-1
Evolution des liens sémantiques	Faire évoluer les relations dans les limites de la norme AFNOR Z 47-100. Scénario 1, 2,3	OK. L'administrateur système dispose de toutes les relations, qu'il peut utiliser ou ne pas utiliser. Il peut également créer des facettes en dehors des relations à l'aide de la polyhiérarchie.
	Développer des relations plus fines. Scénario 2.	Evoluer vers un logiciel autonome intégrant la norme ISO 25964-1
Apprécier la nécessité d'une évolution	Faciliter la compilation de statistiques relatives à l'indexation et à la fréquence d'utilisation des termes.	Il existe un compteur d'occurrences d'indexation. Pour les statistiques de recherche, il convient d'ajouter d'autres modules.

Tableau 4 - Test des fonctions de gestion et maintenance de thesaurus

5.3.2 Une évolution des langages documentaires de Lise et Cinéressources

L'étude de faisabilité a permis de faire le point sur les possibilités de fusion et d'alignement des vocabulaires des institutions partenaires. Quel que soit le scénario choisi, des ajustements devront être établis pour les vocabulaires décrivant les films et les non-films dans la perspective du futur portail.

5.3.2.1 Un thesaurus pour décrire les films

Les mots clés libres actuellement utilisés dans *Lise* ne permettront pas en l'état d'effectuer la transition vers le portail commun, quel que soit le scénario envisagé. Il sera donc nécessaire de créer un thesaurus pour les films ou d'adopter un thesaurus existant. C'est cette deuxième proposition qui semble la plus efficiente, compte tenu du vocabulaire existant et des moyens disponibles. Les équipes du *CNC* pourront s'orienter vers le thesaurus encyclopédique *MotBis* ou le thesaurus de l'Institut National de l'Audiovisuel (INA).

Suite à une étude des mots clés possédant le plus d'occurrences au sein de Lise, il apparaît que *MotBis* comporte tous ces termes. Les mots clés les plus généralistes peuvent être spécifiés par le thesaurus et des relations d'association, comme d'équivalence sont établies entre les termes. De plus, *MotBis* possède son interface web de recherche comprenant les relations de tous les termes, ce qui facilite à la fois son appropriation par les usagers internes et le public externe[66]. En plus des champs thèmes, *Motbis* intègre également des thesauri géographiques et historiques sujet. Il n'y aurait donc pas besoin d'avoir recours à des thesauri extérieurs pour les événements historiques et les pays en sujet. Enfin, de nombreuses institutions, même hors du domaine de l'éducation, ont déjà recours à ce thesaurus. Or, le recours à un même thesaurus permet de rendre plus facilement interopérable les collections.

A terme, l'*INA* va rendre accessible son thesaurus en ligne. Il offre pour l'heure une description globale des champs sémantiques couverts[67]. Spécifiquement construit pour les films, il pourrait offrir via une adaptation progressive, des possibilités intéressantes pour la description des fonds du *CNC*.

Les pratiques différentes des cinémathèques partenaires et des archives du film ont occasionné des évolutions parallèles des pratiques d'indexation concernant les films, bien que *Lise* soit un outil commun. Les référentiels existants (guide de catalogage et pratique du résumé et des mots-clés) sont peu mis en avant. Les entretiens ont de plus établi l'absence de clarté des procédures d'indexation. Le thesaurus évacue les ambigüités liées aux mots clés (gestion de la synonymie, de l'homonymie…), il unifie le langage, mais il n'organise pas les pratiques. Les quelques grands travaux accompagnant la création ou la reprise d'un thesaurus film seraient donc :

- Le choix de termes sans équivoques.
- Des notes d'applications claires attachées à ces termes.
- Des versions papiers du thesaurus avec plusieurs entrées sur les vocabulaires pour familiariser les équipes avec l'outil.
- Le paramétrage d'une arborescence claire et proposant des termes à l'indexation.

[66] http://www.cndp.fr/thesaurus-motbis/site/
[67] http://www.inatheque.fr/publications-evenements/ina-stat/ina-stat-methodologie.html

- Une formation des documentalistes à la recherche avancée et aux spécificités du logiciel *Adlib* afin de s'assurer que chacun a acquis la base de fonctionnalité de l'outil et ne se pose plus de questions à même de causer des indexations inadéquates. Cette formation pourrait permettre de dégager des responsables de l'évolution du thesaurus au sein du service AGDC de la direction du Patrimoine du *CNC*.

5.3.2.2 Une adaptation des vocabulaires non-films

L'étude du thesaurus *Cinédoc* fait apparaître quelques ajustements dans les relations relevés lors des entretiens d'étude de besoins, mais il est globalement adapté pour effectuer la transition vers le nouveau portail. Cinédoc pourra être chargé dans la base thesaurus comme thesaurus externe (scénario 3). Son vocabulaire est également assez riche pour permettre des alignements avec le thesaurus film dans le cadre des scénarios 2 et 3.

Le principal chantier pour la cinémathèque française sera de déterminer si le thesaurus iconographique Garnier sera utilisé de manière permanente pour les photos, affiches, dessins et le matériel publicitaire, et à quel niveau il sera adapté. Comme nous l'avons vu, ce thesaurus pourrait constituer un thesaurus externe à la base Thesaurus d'Adlib. Cependant, il y'a peu de possibilités d'alignements du vocabulaire Garnier avec les autres vocabulaires. L'indexation de ces documents devrait donc être, dans un premier temps, liée à un champ spécifique avant d'envisager les alignements décrits dans le scénario 2 (voir 5.2.3).

CONCLUSION

Le thesaurus employé pour décrire les collections films et non-films au sein du futur portail du patrimoine cinématographique suivra donc la voie de l'interopérabilité. Cette voie, constituée de liens entre les vocabulaires, dégage des opportunités pour les recherches sur les différents types de documents liés au cinéma. A terme, elle permettra d'interagir avec d'autres collections liées à la culture, et éventuellement d'autres domaines. Le chef de projet chargé d'étudier la faisabilité d'un thesaurus mutualisé doit dès le départ envisager ces horizons et prendre en considération l'évolution des pratiques de recherche sur le web. Il doit néanmoins ne pas perdre de vue que plus un thesaurus est spécialisé, plus les alignements avec d'autres vocabulaires se révéleront artificiels. Les difficultés liées à de hautes ambitions d'interopérabilité seront également imputables aux équipes opérationnelles. Mutualiser les moyens demandera aux documentalistes du CNC et de la cinémathèque française une grande flexibilité dans un nouveau mode d'organisation de l'indexation des collections : des autorités œuvres et personnes communes, un meta-thesaurus commun et une adaptation à un nouvel outil. A ces nouveautés s'ajoutera une nécessité de rapprocher les pratiques d'indexation en dépit de missions qui demeurent différentes entre les institutions. Bien que progressant vers la valorisation de ces collections, le CNC demeure une institution de dépôt et d'archivage. Les cinémathèques s'identifieront quant à elles durablement à des musées. Il convient donc d'accompagner les équipes dans cette transition avant de penser plus grand.

Le livrable rendu n'est qu'un outil pour amorcer cette coopération. Les préconisations établies constitueront un appui à la réflexion et aux discussions des partenaires. Cette étude comporte les défauts d'un trop grand nombre d'hypothèses pour peu de certitudes, ce qui comble sa nature d'aide à la décision, mais laisse quelque peu en suspens les modalités pratiques de l'élaboration de l'outil qui auraient demandé un peu plus de temps pour être développées. Véritablement orientée utilisateurs, elle constituera dans tous les cas un document commun au *CNC* et aux cinémathèques partenaires qui permettra de cerner les besoins documentaires des équipes et des publics respectifs ainsi qu'une meilleure appréhension du logiciel qui accueillera les collections. Et si un des scénarios vient à être adopté, ces axes donneront des outils au chef de projet qui prendra la suite.

BIBLIOGRAPHIE

La bibliographie analytique suivante a été arrêtée le 19 octobre 2016. Elle comprend 32 références. Elle n'a pas pour vocation l'exhaustivité sur le sujet. Seules les sources les plus pertinentes ont été sélectionnées pour figurer dans cette bibliographie.

Les notices ont été regroupées sous les quatre rubriques thématiques suivantes, suivant la progression logique de ce mémoire :

- Thesauri et langages documentaires
- Thesauri et recherche d'information
- L'interopérabilité sémantique
- Etude de faisabilité de thesaurus

La rédaction des références bibliographiques est conforme aux normes suivantes :

- Z44-005. décembre 1987. Documentation. Références bibliographiques : contenu, forme et structure.
- NF ISO 690-2 février 1998 Information et documentation. Références bibliographiques, documents électroniques, documents complets et parties de documents.

Thesauri et langages documentaires

[1] Menon Bruno, « Les langages documentaires. Un panorama, quelques remarques critiques et un essai de bilan», *Documentaliste-Sciences de l'Information* 1/2007 (Vol. 44) , p. 18-28 URL : www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-18.htm. DOI : 10.3917/docsi.441.0018.

> *L'auteur dresse une présentation et une Histoire des langages documentaires de leur apparition à leur emploi comme « systèmes d'organisation des connaissances ». En filigrane, il aborde les grandes phases qui ont mené à l'évolution des pratiques de recherche d'information et dégage un bilan pour l'utilisation future de ces langages.*

[2] Thesauro-glossaire des langages documentaires, Danièle Dégez et Dominique Ménillet, Collection : Sciences de l'information - série Recherches et documents, 2001, 185 page(s), ISBN 2-84365-051-8

> *Dans son glossaire, cet ouvrage présente une série de définitions liées aux langages documentaires qui me furent très utiles pour préciser certaines notions et me faire comprendre de mes interlocuteurs durant la mission.*

[3] L'avenir des langages documentaires dans le cadre du Web sémantique : conception d'un thésaurus iconographique pour le Petit Palais, Béatrice Pierre, 2010, Mémoire INTD- Titre I

> *Dans sa première partie, ce mémoire de l'INTD développe les différentes théories de l'indexation, notamment l'indexation des images fixes. Il pourra servir de complément à cette étude.*

[4] Les classifications, theorie et pratique, l'avenir de la classification décimale de Dewey dans une bibliothèque universitaire : L'exemple du SCD de Lille, Frédéric Watrelot, 1995. 54 p. Mémoire ENSSIB.

> *Dans le cadre d'utilisation des bibliothèques universitaires, l'auteur compare la classification Dewey et la classification décimale universelle en expliquant les atouts de chacune. Il explique ainsi le succès de la première auprès du public, du fait de sa simplicité et sa praticité et l'abandon progressive de la seconde, bien plus complexe dans sa formulation.*

[5] Des classifications aux thesaurus, du bon usage des facettes, Jacques Maniez, Documentalistes sciences de l'information, Volume 36, N° 4-5, paru le 1 juillet 1999, page(s) 249-262

Cet article ancien est toujours d'actualité. Exposant les différents courants théoriques liés à la classification à facettes, il permet autant d'aborder des bases théoriques de l'indexation que de mettre en évidence l'utilité des facettes dans les classifications, et par extension pour les thesauri.

[6] Chaumier Jacques, « Les ontologies. Antécédents, aspects techniques et limites», *Documentaliste-Sciences de l'Information* 1/2007 (Vol. 44) , p. 81-83 URL : www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-81.htm. DOI : 10.3917/docsi.441.0081.

L'auteur apporte des précisions sur les ontologies et leur fonction en les rapprochant de leur cadre d'origine, l'intelligence artificielle. Il établit des distinctions claires entre thesauri et ontologies.

[7] Francis Élie, Quesnel Odile, « Indexation collaborative et folksonomies », *Documentaliste-Sciences de l'Information* 1/2007 (Vol. 44) , p. 58-63 URL : www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-58.htm. DOI : 10.3917/docsi.441.0058.

Publié en plein essor de l'indexation collaborative, cet article analyse les raisons du succès des folksonomies et démontre leurs faiblesses au regard des langages documentaires. Il demeure d'actualité en ce qu'il démontre un type d'appropriation de l'espace du web en réaction aux langages documentaires classiques.

Thesauri et recherche d'information

[8] Le Marec, Joëlle. « Les oPACs sont-ils opaques ? ». Bulletin des bibliothèques de France (BBF), 1989, n° 1, p. 78-85. Disponible en ligne : <http://bbf.enssib.fr/consulter/bbf-1989-01-0078-007>. ISSN 1292-8399.

Cet article est riche d'enseignement sur le rapport de l'usager à un outil de médiation documentaire informatisé. L'auteure se place du point de vue de l'usager de la BPI lors de la mise à disposition des premiers OPAC et analyse le fossé entre les intentions de ceux qui ont créé les systèmes et le ressenti des usagers qui y sont confrontés.

[9] Thesaurus et informatique documentaire : Partenaires de toujours? ", Sylvie Dalbin, In Documentalistes, science de l'information, 2007, vol.44, n°1, p. 42-55 http://www.cairn.info/article_p.php?ID_ARTICLE=DOCSI_441_0042

Spécialiste des langages documentaire, Sylvie Dalbin explique dans cet article la division entre le thesaurus à la recherche et le thesaurus à l'indexation. Elle développe ensuite des cas où un thesaurus peut être utilisé par les systèmes documentaires dans une perspective de recherche.

[10] Optimiser l'accès à l'information, une opportunité pour les langages documentaires ? Brunon Menon, Journée d'étude ADBS, Documentaliste, sciences de l'information 2007/6 (vol. 44), p.385-388.

Datant de 2007, Cette synthèse permet de faire le point sur plusieurs années d'exposition au web et de constater comment les moteurs de recherche ont façonné les habitudes de recherche du grand public. Les opportunités d'utiliser les langages documentaires en appui aux systèmes de recherche sont étudiées par différents intervenants.

[11] GREENBERG, Jane. « User Comprehension and Application of Information Retrieval Thesauri ». Cataloging & Classification Quarterly, 2004, vol. 37, n° 3-4, p. 103-120

L'article décrit une expérience menée sur des publics étudiants exposés à des recherches avec ou sans thesaurus. Ses résultats vont dans le sens d'une utilisation du thesaurus à la recherche, pourvu qu'il soit adapté et accompagné de formations.

[12] BOURGEAUX Laure, FRESNEAU Amélie. Moteur ou labyrinthe ? Le portail documentaire de la Bibliothèque publique d'information évalué par ses utilisateurs. BBF, 2009, n°6, p.73-77.

Cet article se place également du point de vue de l'utilisateur de la BPI, mais 20 ans après. Il est cette fois confronté au portail documentaire. Il liste les différents écueils du portail qui rendent son utilisation confuse.

[13] Thesaurus et informatique documentaire : les noces d'or, Sylvie Dalbin, In Documentalistes, science de l'information, 2007, vol.44, n°1, p. 76-80, <http://www.cairn.info.proxybib.cnam.fr/revue-documentaliste-sciences-de-l-information-2007-1.htm>

L'auteure aborde les liens qui existent entre thesauri systèmes informatiques. Il établit une Histoire de ce rapport à travers l'évolution de l'utilisation du langage documentaire pour les serveurs de banques de données, les centres documentaires, l'archivage et le web.

L'interopérabilité sémantique

[14] GIUSTI Aurélia. La recherche fédérée des portails patrimoniaux : quelles solutions documentaires ? L'exemple du MuCEM. Mémoire pour l'obtention du titre professionnel « Chef de projet ingénierie documentaire ». Paris : INTD, 2009,110 p.

Ce mémoire de l'INTD offre des exemples de recherches fédérés sur plusieurs portails. Il permet de cerner avec précision les enjeux de l'accès simultané à différents fonds.

[15] FEYLER François. De la compatibilité à l'interopérabilité en matière de repérage d'information pertinente : la problématique et l'exemple d'OTAREN. Documentaliste-Sciences de l'information, 2007, vol.44, n°1, p.84-92.

Cet article est plus intéressant dans son explication du contexte qui préside à l'interopérabilité des langages documentaires que pour sa description de l'expérience OTAREN, qui en est alors à ses débuts.

[16] RAVET Brice. De l'usage des outils du Web de données pour une recherche efficace sur des ressources disséminées et hétérogènes : la mise en place d'un portail de recherche fédérée pour le Musée National du Sport. Mémoire pour obtenir le titre professionnel « Chef de projet en ingénierie documentaire », Paris : INTD, 2011, 187 p.

Ce mémoire INTD particulièrement complet permet de faire un point à la fois sur la fonction des portails documentaires, sur la recherche fédérée et sur l'utilisation des thesauri dans l'interopérabilité des systèmes.

[17] FEYLER François, OTAREN, conditions de la migration de G3I à GTMT3. Savoirs Cdi, publié en décembre 2011, consulté le 9 août 2016.< https://www.reseau-canope.fr/savoirscdi/centre-de-ressources/fonds-documentaire-acquisition-traitement/le-traitement-documentaire/levolution-des-normes-et-les-outils-de-gestion-des-vocabulaires-controles.html >

Ce second article sur l'expérience OTAREN explique les différentes phases d'alignement et il est assorti d'exemples qui introduisent très bien le contexte de normalisation sur l'interopérabilité des langages.

[18] Le thesaurus W dans le web de données, Ministère de la culture, data.culture <http://data.culture.fr/thesaurus/static/thesaurus-w-web-de-donnees>

Parmi les nombreux textes vulgarisant le web de données liées sur le web, cette introduction à partir de l'exemple du thesaurus W est la plus accessible et la mieux illustrée. Elle sera un bon complément à la partie de ce mémoire traitant des outils du web sémantique.

[19] Michèle Lénart, « SKOS, un langage de représentation de schémas de concepts », Documentaliste-sciences de l'information 2007/1 (vol.44), p.75.

L'auteur de cet article explique le fonctionnement global du schéma SKOS ainsi que ses particularités eut égard aux autres schémas de données.

[20] L'interopérabilité sémantique : Une révolution? Les normes SKOS (W3C, 2009) et ISO 25964-1, Isabelle Boydens, 10/04/2012 - consulté le 06/07/2016, Smals Resaerch, 2012, http://blogresearch.smalsrech.be/?p=4091

Cet article de blog développe les enjeux de SKOS et de la norme ISO 25964-1 pour les futurs concepteurs de thesauri, autant dans une perspective de gestion de thesaurus que dans une perspective managériale.

[21] Reengineering thesauri for new applications : the AGROVOC Example, Soergel Dagobert et al., JoDI, vol 4, n°4, 20014. https://core.ac.uk/download/pdf/11887854.pdf

A travers l'exemple d'AGROVOC, ce rapport permet de comprendre les défauts des thesauri classiques et comment une modélisation dans le type de SKOS peut les corriger. L'auteur nous délivre ensuite une méthode opérationnelle pour transformer un thesaurus en ontologie SKOS.

[22] Norme ISO 25964-1 : 2011. Information and documentation – Thesauri and interoperability with other vocabularies – Part 1 : Thesauri for information retrieval. 1ère édition. Genève, International Organization for Standards, Août 2011. 152 p. (anglais)

Cette norme peut être considéré comme la norme actuellement en vigueur pour la construction d'un thesaurus. Elle est illustrée de nombreux exemples et de schémas permettant de faciliter l'intégration de thesauri dans les systèmes d'information. Il n'existe pas à ce jour de version française de la norme.

[23] Livre blanc sur la norme ISO 25964-1 « Thésaurus pour la recherche documentaire » [PDF en ligne]. DALBIN Sylvie, YAKOVLEFF Nathalie, ZYSMAN Hélène, et al., 29 janv. 2013

En l'absence de traduction française de la norme, ce livre blanc demeure la ressource francophone la plus complète sur ISO 25964-1. Les auteures, qui ont participé à son élaboration mettent en avant les points saillants de la norme dans ce qui les différencient avec les normes françaises.

[24] ISO 25964 : de la distinction formelle concept/terme préconisée par la norme pour la création et la gestion de thesaurus, Laurence Maroye, Consulté le 06/07/2016, I2D, P.72. n°1 2015 – CAIRN

L'auteure de cet article se penche sur la norme ISO 25964-1 et discute l'attribution d'un identifiant unique aux concepts en s'appuyant sur des références linguistiques.

[25] Norme ISO 25964-2 :2013 – Information and documentation – Thesauri and Interoperability with other vocabularies.1e edition. Suisse. ISO, Mars 2013. 99p (Anglais)

Cette seconde partie de la norme se focalise sur l'interopérabilité des thesauri avec les autres langages documentaires et introduit la notion d'alignements. Elle est autant applicable à des contextes de recherches sur plusieurs fonds que dans le contexte du web sémantique. Aucune version française de cette norme n'est à ce jour disponible.

[26] De la documentation à la médiation, Marie Després Lionnet. Documentaliste - science de l'information, n°2, 2014 vol.51, p. 62.

Issu d'un dossier sur les nouvelles médiations culturelles dans le cadre du web de données, cet article d'une professionnelle de la documentation expose très bien la manière dont le documentaire parvient désormais à

occuper une place dans la valorisation de collections patrimoniales, notamment muséales.

[27] Le développement du web des données culturelles. Les enjeux pour le ministère de la culture et de communication, Bertrands Sajus, Marie-Véronique Leroi, I2D - 2016/2 (volume 53), p 46-47, p.65.

Cet article décrit les différentes initiatives du ministère de la culture en lien avec le web de données et qui tendent à valoriser les collections auprès des publics : Hadoc, Semanticpedia, JocondeLab.

[28] Hadoc, un programme pour harmoniser les données culturelles, Katell Briate, Documentaliste - science de l'information, n°2, 2014 vol.51. p.65.

L'initiatrice du projet Hadoc explique les raisons et le contexte qui ont mené à construire l'alignement des vocabulaires du Ministère de la culture.

Etude de faisabilité de thesaurus

[29] Construire un thesaurus. DEGEZ, Danièle, Archimag, SERDA, 2009, n°222, p.44-45. ISSN 0769-0975

Concis et explicite, cet article permet d'avoir une vue générale sur les étapes de l'ingénierie d'un thesaurus. Il pourra être utile au chef de projet conduisant une étude de faisabilité afin de planifier et d'organiser ces différentes phases.

[30] Guide pratique pour l'élaboration d'un thesaurus documentaire, Michèle Hudon, avec la collaboration de Danièle Dégez et Dominique Ménillet. – Montréal : Les Éditions ASTED (diff. en France : ADBS), 2009. – 274 p. – ISBN 978-2-923563-17-6

Cet ouvrage de référence sur la création de thesaurus me fut particulièrement utile durant ma mission. Il aborde non seulement les étapes de création du langage, mais aussi de manière exhaustive l'étude de faisabilité préalable. Il présente également des critères d'évaluation des logiciels de gestion de thesauri.

[31] Encadrer la réingénierie d'un ***thesaurus*** *: méthode, enjeux et impacts pour l'équipe d'un service de* veille et documentation en entreprise, Loraine Keller, 2013. Mémoire INTD - Titre I ; http://memsic.ccsd.cnrs.fr/mem_00945542/document

Ce mémoire de l'INTD décrit avec brio les méthodes mises en place pour une réingénierie de thesaurus dans un contexte bancaire. Il met en évidence la nécessité d'un chef de projet et d'une équipe de professionnels formés afin de bien mener cette mission.

[32] Logiciel de création et de maintenance de thesaurus, Sylvie Dalbin, Descripteurs, Création : le 5 janvier 2006 - Dernière mise à jour : Mars 2016 (en cours)- vu le 02/08/2016, http://dossierdoc.typepad.com/descripteurs/2011/12/logiciel_de_thesaurus.html

Ce billet du blog Descripteurs, spécialisé dans les langages documentaires, dégage une typologie des logiciels et modules de création, gestion et maintenance de thesauri et met à jour une liste de ces différents outils. Il constitue une bonne entrée en matière pour réaliser un benchmark de ces outils.

Le Blog Descripteurs a constitué une très bonne entrée en matière pour suivre l'état de l'art des langages documentaires.

ANNEXES

ANNEXE 1 – NOTE DE MISSION

DEROULEMENT DE LA MISSION

Cadre

Dans le cadre du projet de fusion des portails documentaires LISE et cinéressources, la mission consiste en étude d'opportunité et de faisabilité de la refonte et/ou de la fusion des vocabulaires contrôlés respectifs. Elle implique :

- Une analyse comparative des pratiques utilisateurs des institutions partenaires du projet en terme d'indexation matières / mots du sujet (outils et méthodes).
- Une étude de l'existant.
- La participation à l'évaluation des fonctionnalités thesaurus du progiciel ADLIB, choisi pour devenir l'outil intégré de gestion de l'ensemble des collections.
- Le cas échéant, un accompagnement de la conception et de la réalisation du/des langages contrôlés unifiés.

Différentes phases

1. Etude des pratiques utilisateurs / opportunités d'un thesaurus thématique

Quels utilisateurs ?

La première phase correspond à l'étude des besoins et des pratiques des utilisateurs du futur thesaurus. Le terme utilisateur regroupe à la fois :

- Les professionnels de la documentation des institutions partenaires qui utiliseront le/les futurs thesaurus pour indexer les films ou non-films, et pour effectuer des recherches en interne, que l'on appellera **documentalistes**.
- Les équipes des institutions partenaires accédant au backoffice de la future plateforme, et qui pourront effectuer des recherches via l'outil pour des travaux internes, que l'on appellera **usagers internes**. Cette catégorie comprend les documentalistes qui effectuent des recherches pour des usagers internes.
- Les usagers des portails Lise et Cinéressources, du centre de documentation du CNC, ainsi que les autres usagers potentiels (grand public, chercheurs, étudiants en cinéma…), qui n'auront accès à la recherche et aux notices que par le portail. Nous les appellerons **usagers externes**.

But de la démarche

Cette première phase permettra d'évaluer quels sont les avantages découlant d'une indexation thématique des films et non films sur le futur portail, et ce pour chacun des types d'usagers. Elle répondra entre autres aux questions suivantes :

- Quels sont les points satisfaisants et les inconvénients de l'existant ?

- Dans quelle mesure les usagers du portail auront-ils besoin d'interroger par thème sur le thesaurus ?
- Quels avantages peuvent procurer pour les publics la mutualisation d'une indexation thématique des collections des partenaires ?
- Quels sont les types de requêtes menées par ces usagers ? Quels seraient les gains significatifs d'un thesaurus commun ?
- Elle me permettra également de connaître les pratiques des documentalistes, ainsi que les difficultés rencontrées avec les thesaurus et listes de mots clés actuellement utilisés.

Moyens et outils

Cette première phase de l'étude pourrait être couverte par :

- Des entretiens individuels menés auprès des documentalistes sur chacun des fonds qui intégreront la future plateforme. Ces entretiens permettront également de développer les recherches des usagers externes et les besoins des usagers internes.

- Des entretiens individuels menés auprès d'usagers internes dont les besoins de recherches thématiques seront identifiés comme les plus importants. En fonction de l'institution, ils peuvent comprendre :

- Les documentalistes, pour leurs missions de recherche.
- Les programmateurs dans le cadre d'expositions thématiques ou de valorisation d'une « collection » particulière.
- Les équipes travaillant sur l'identification et la restauration des films.

Cette liste n'est pas fermée. Elle est limitée par ma connaissance actuelle des équipes et admet vos suggestions.

- La consultation de documents (enquêtes de satisfaction, document statistiques…) permettant d'identifier les usagers externes et la nature des requêtes effectuées sur les portails Lise et Ciné ressources ainsi qu'au centre de documentation du CNC.

Exemples de questions pour les usagers internes

- Quelles types de recherche êtes-vous amené(e) à faire dans le cadre de vos missions (dossiers documentaires, sélection pour expositions, filmographies…) ?
- Pour quel(s) public(s) ?
- Etes-vous autonomes dans vos recherches ou passez-vous par un intermédiaire ?
- S'agit-il de recherches par thème ?
- Effectuez-vous des recherches sur LISE ? Ciné ressources ? Si oui, lesquelles ? A partir du front/du back office ?
- Comment procédez-vous pour les effectuer ? Les résultats sont-ils pertinents ? (bruit/silence ?)
- Quelles seraient les apports d'un thesaurus thématique pour vos recherches ?
- Quelles seraient les apports à votre travail et d'une manière générale d'une fusion FILM/NON FILMS sur les recherches thématiques ?

La courte durée du stage ainsi que le temps à dégager pour les autres phases nécessitent de privilégier l'information qualitative, d'où le choix des entretiens et la limitation aux interlocuteurs les plus représentatifs. Les personnes interrogées pourront donc se limiter à une personne par service concernée (sous réserve de la disponibilité et de l'intérêt pour le projet d'autres interlocuteurs).

Si des études de besoin ont déjà été menées dans le passé concernant chacun de ces types d'utilisateurs et les requêtes effectués, merci de m'en informer.

Durée estimée

Etant donnée son importance pour la suite, cette phase pourrait courir jusqu'à début août.

2. Etude de l'existant

Cette phase est déjà amorcée dans sa partie recensement, mais elle sera principalement menée suite aux entretiens.

Analyse sur la forme

- Recensement des thesaurus et listes de mots clés utilisés.
- Etude de ces langages à l'aune des besoins recensés et des normes en vigueur.
- Réflexion autour des normes à utiliser en fonction des différents scénarios.

Analyse sur le fond

- Etude des rapprochements possibles des vocabulaires, **des points d'alignement**, des rapprochements de synonymes, des hiérarchisations et associations possibles dans le cas d'un thesaurus unique.
- Etude de la possibilité de champs sémantiques communs.
- Etude des rapprochements avec les autres listes.
- Etude des possibilités d'intégration/utilisation de thesaurus, d'interopérabilité des langages et d'exportation au regard des fonctionnalités de l'outil ADLIB.

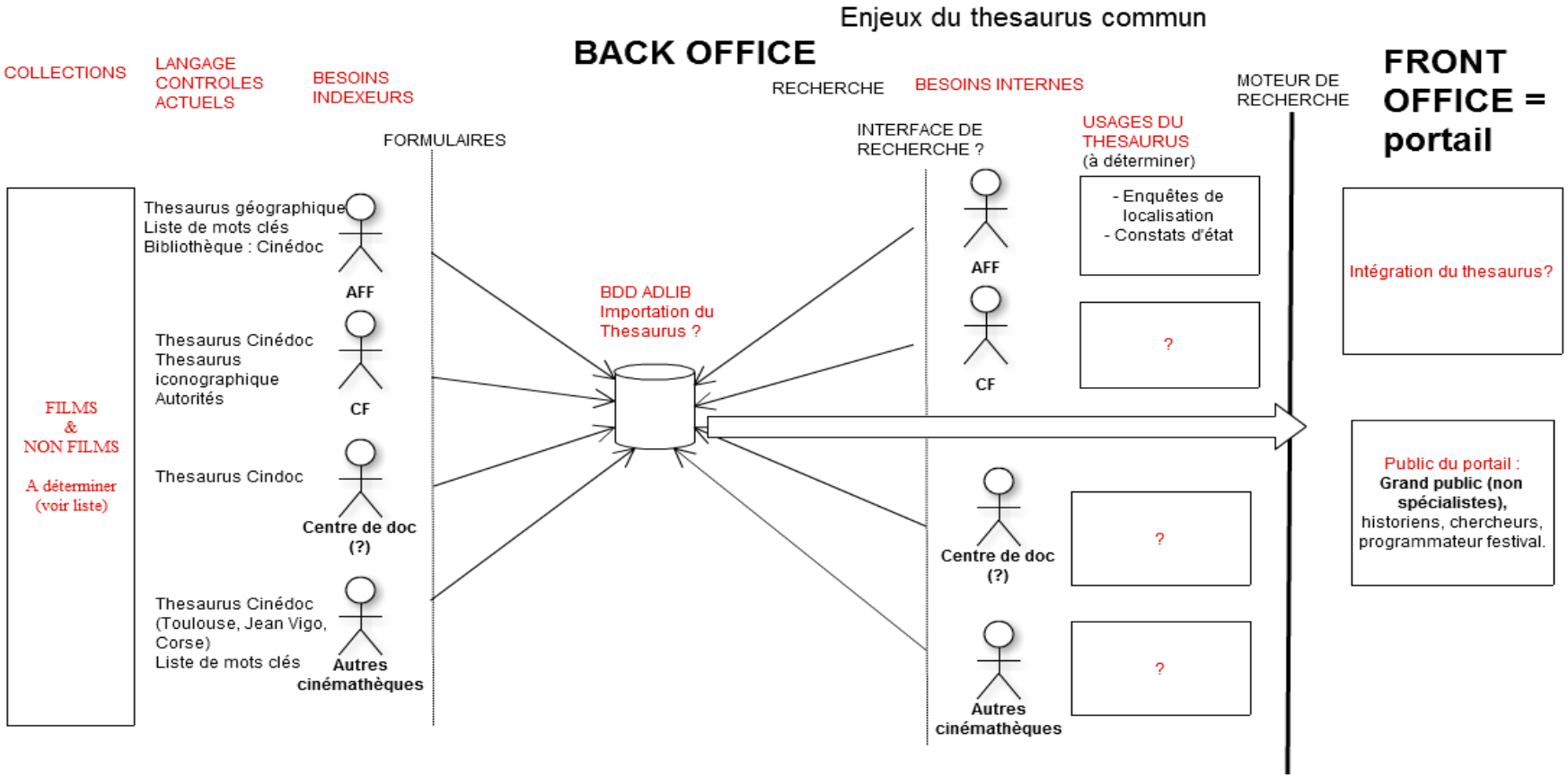
3. Etablissement des éléments à développer à partir de l'existant sur chacun des scénarios (alignement, fusion, autre voie)

- en fonction des objectifs définis par les besoins des utilisateurs
- en fonction de l'existant
- en fonction des possibilités d'ADLIB

Cette étape mènera à des **préconisations**.

Un livrable de préconisations sur la conception, la réalisation du thesaurus et l'accompagnement au changement des équipes pourra, le cas échéant, être rédigé.

ANNEXE 2 – SCHEMA « ENJEUX DU THESAURUS COMMUN »


Enjeux du thesaurus commun
BACK OFFICE
FRONT OFFICE = portail
COLLECTIONS
LANGAGE CONTROLES ACTUELS
BESOINS INDEXEURS
FORMULAIRES
RECHERCHE
BESOINS INTERNES
INTERFACE DE RECHERCHE ?
MOTEUR DE RECHERCHE
USAGES DU THESAURUS (à déterminer)
FILMS & NON FILMS
A déterminer (voir liste)
Thesaurus géographique
Liste de mots clés
Bibliothèque : Cinédoc
AFF
Thesaurus Cinédoc
Thesaurus iconographique
Autorités
CF
Thesaurus Cindoc
Centre de doc (?)
Thesaurus Cinédoc (Toulouse, Jean Vigo, Corse)
Liste de mots clés
Autres cinémathèques
BDD ADLIB
Importation du Thesaurus ?
AFF
CF
Centre de doc (?)
Autres cinémathèques
- Enquêtes de localisation
- Constats d'état
?
?
?
Intégration du thesaurus?
Public du portail : Grand public (non spécialistes), historiens, chercheurs, programmateur festival.

Vue sur les fonctionnalités thesaurus d'Adlib for Windows

Intégration d'un ou plusieurs thesaurus matière

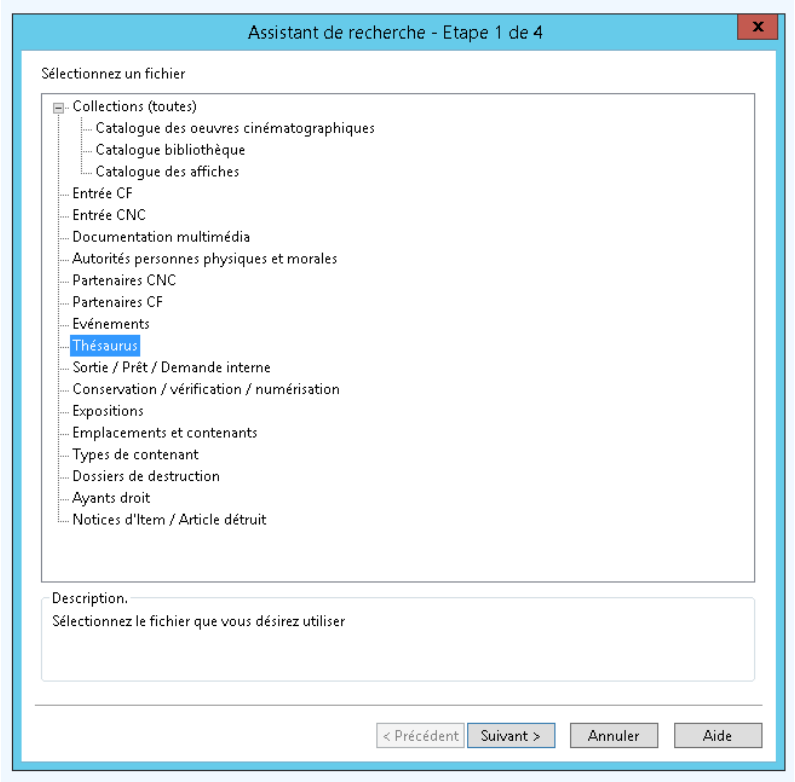
Eveno Guénaël

30/08/2016

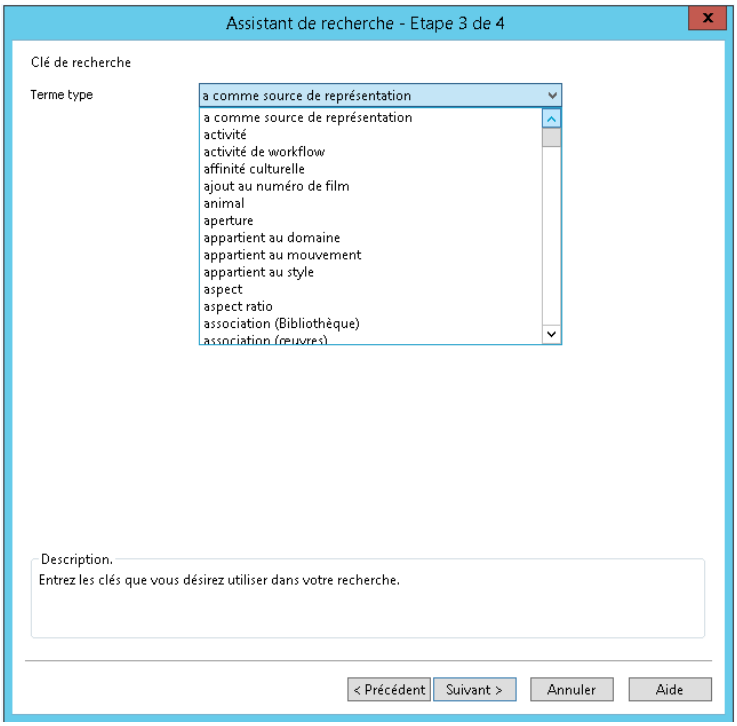
Maj : 21/09/2016

La base Thesaurus

C'est une base de données d'ADLIB for Windows qui intègre les différentes listes contrôlées ouvertes. Elle est distincte des autorités personnes physiques et morales, qui se trouvent dans une autre base.



Ces listes sont nommées « domaines ». Le domaine est également désigné par l'expression terme type (type term). Chaque domaine comprend des descripteurs (termes). Chaque descripteur admet des relations. Un domaine peut donc être considéré comme une facette d'un Thesaurus plus important.



En faisant une recherche sur le type de terme en Etape 2 de l'*Assistant de recherche*, on peut voir défiler en Etape 3 une liste des domaines déjà présents dans l'application.

Parmi ces domaines, on peut actuellement compter des listes appartenant au CNC ou la cinémathèque Française, mais il s'y trouve aussi des listes déjà intégrées à Adlib. Selon la réponse d'Axiell et bien qu'il ait été demandé explicitement de les retirer

pendant la phase de tests, ces domaines proviennent de l'éditeur. Ils sont intégrés à l'application et ne peuvent être supprimés. Ils comportent entre autres un domaine thématique. Cependant, ces domaines sont vides. Ils ne sont reliés à aucun terme, et ne constitueront pas de parasitage pour les listes du CNC et de la cinémathèque française :

Axiell 2016-07-29 :

« (…) Le thésaurus est intégré dans l'application complète. Les domaines qui y se retrouvent sont donc un ensemble de tous les domaines qui sont utilisés dans l'application complèt(e) d'Adlib. Comme il y a des parties qui ne pas encore activés pour le projet, il y a des domaines qui semblent vide, mais qui vont être alimentés au fur et à mesure du projet. »

Parmi les scénarios possibles, un thesaurus sujet partagé par le CNC et la Cinémathèque Française pourrait être intégré en tant que domaine dans la base Thesaurus. Il serait alors considéré comme une des facettes ou un des sous-groupes de cette base. De cette manière, il est possible de constituer un thesaurus autonome au sein de la base Thesaurus. L'intégration du Thesaurus géographique de *Lise* en est la démonstration.

Il est également possible de décomposer un thesaurus en différents champs sémantiques qui seraient des domaines et d'intégrer des listes autonomes.

Enfin, Adlib inclut la polyhiérarchie au sein du thesaurus. Il est donc possible de créer une facette comme un domaine et de lier chacun des termes à ce nouveau domaine. Le terme aurait alors deux domaines : le domaine de la liste d'origine (thesaurus créé) et le domaine de la nouvelle facette. Cette solution peut être employée pour créer une facette commune à plusieurs domaines dans le but d'améliorer la recherche en front office.



Fonctions d'alignement d'Adlib

Le thesaurus proposé par la base *Thesaurus* étant global, des termes d'un domaine peuvent avoir des relations avec ceux de domaines différents. Cette configuration permet entre autres de créer dans les faits des relations d'alignement entre plusieurs thesaurus.

On peut faire valoir des relations d'équivalence entre termes de différents thesaurus en utilisant la fonction d'équivalence (« equivalent term ») dans la base Thesaurus.

Exemple d'alignement :

Alpinisme (domain=thème cinématographique, source = Cinédoc) <-> equivalent_term <-> Montagne (domain= Montagne, source = thesaurus Film)

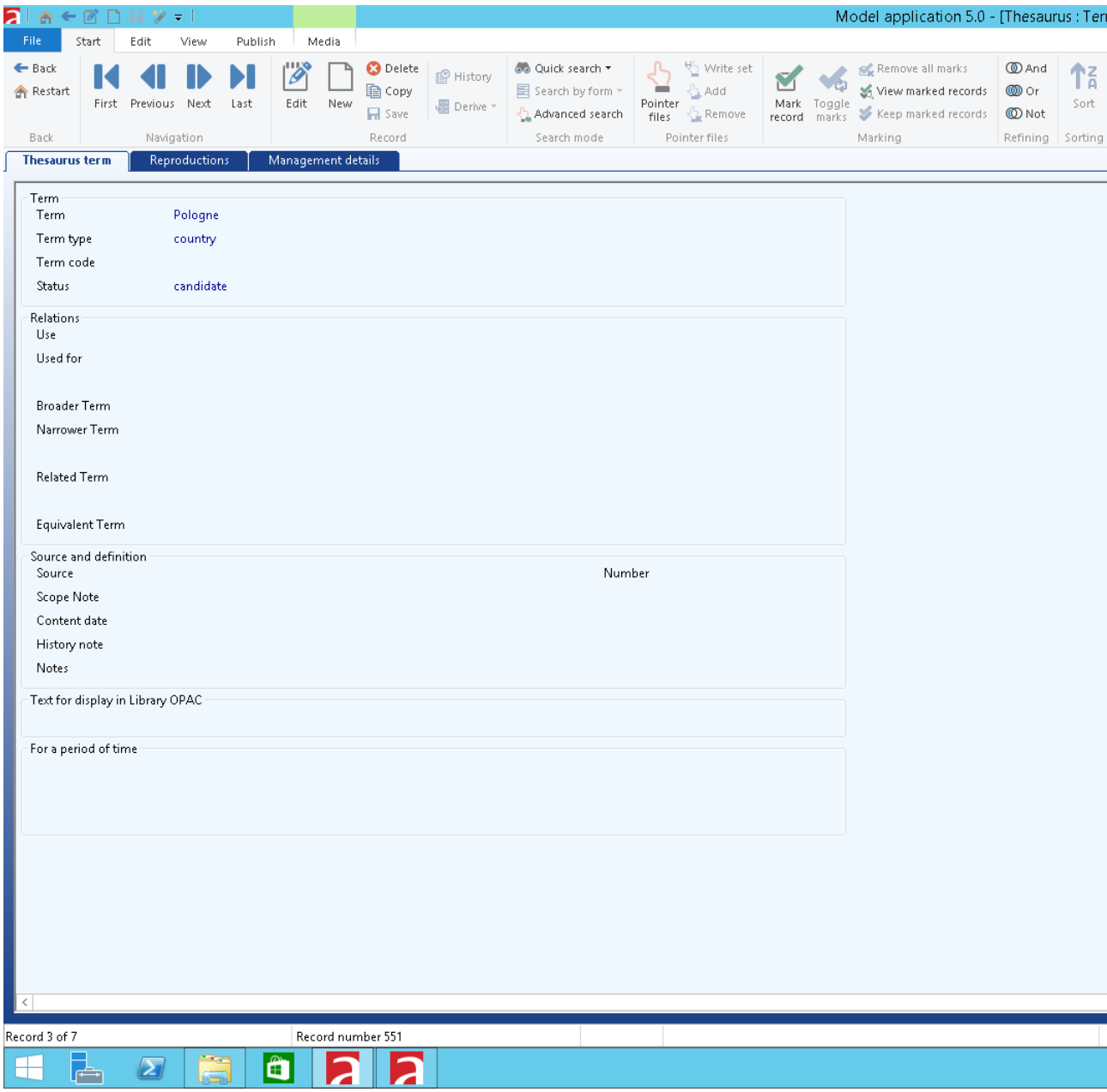
Cette relation d'équivalence peut être utilisée pour la recherche pour être sûr que des enregistrements soient retournés qu'ils utilisent le terme d'un thesaurus ou celui d'un autre, à chaque fois que le chercheur cherche sur chacune d'elles. Elle permet la recherche sur plusieurs collections.

Par ailleurs, Il est possible de paramétrer une fonctionnalité d'alignement au sein de l'application, à la demande du client. Mais cette fonctionnalité contourne le principe de thésaurus global d'Adlib.

Relations

Fiche descripteur

Il existe une fiche descripteur pour chaque terme. La fiche d'un descripteur se présente actuellement comme telle :



- Le terme Pologne appartient au domaine country. Si ce terme était présent dans d'autres domaines, il aurait plusieurs « Term type ».
- Le terme comporte un code de désignation dans ADLIB.
- Son statut est celui d'un candidat.
- La section Relations présente les termes qui seront effectivement en relation avec celui-ci, par type de relations (hiérarchique, équivalence…).
- Il est possible d'ajouter une note sur chaque terme.
- Il est possible d'ajouter une catégorie facteur sémantique.

Relations possibles entre les termes

Vocabulaire ADLIB (notice)	**Utilisation courante**	**Vocabulaire normé**
Utiliser/Utilisé pour	Préférence liée à un usage de la langue dans la discipline / par le public visé. **Ex :** **Voiture ---------- automobile**	Terme préféré (TP descripteur) / Terme non préféré (TNP non descripteur)
Terme spécifique/Terme générique	Se réfère à des concepts plus globaux ou plus fins par rapport à un terme (concept) donné. Un terme peut avoir plusieurs termes génériques et/ou plusieurs termes spécifiques. **ADLIB gère autant de sous-spécifiques que l'on souhaite**.	Terme spécifique (TS) / terme générique (TG)
Terme associé	Renvoi d'un descripteur à un autre descripteur pour compléter la recherche sur un sujet connexe. **Transport en commun ----- moyen de transport collectif / transports urbains**	Terme associé (TA) *« Voir aussi »*
Terme equivalent	Ce sont les termes considérés dans l'application comme des synonymes ou comme signifiant exactement la même chose (par exemple, les traductions d'un même terme dans une langue différente). **Vélo de course ---- Bicyclette/VTT**	Employé (EM) /Employé pour (EP) Permet de lier différents thesaurus.
Facteurs sémantiques	Une combinaison de termes crée un concept nouveau **Ex : « radeau pneumatique de**	Précoordination des termes intégrée au logiciel

	sauvetage » = radeau pneumatique ET sauvetage. Evite l'utilisation de termes trop spécifiques et la multiplication des descripteurs. Nécessité de paramétrage par l'administrateur fonctionnel – Non disponible dans notre version	

Application pratique du facteur sémantique

Si le terme « radeau pneumatique de sauvetage » vient à être tapé dans une notice du catalogue, il est immédiatement remplacé par ses facteurs sémantiques. Lors d'une recherche, si l'utilisateur tape les deux termes précoordonnés, les deux résultats sur « radeau pneumatique de sauvetage » seront affichés.

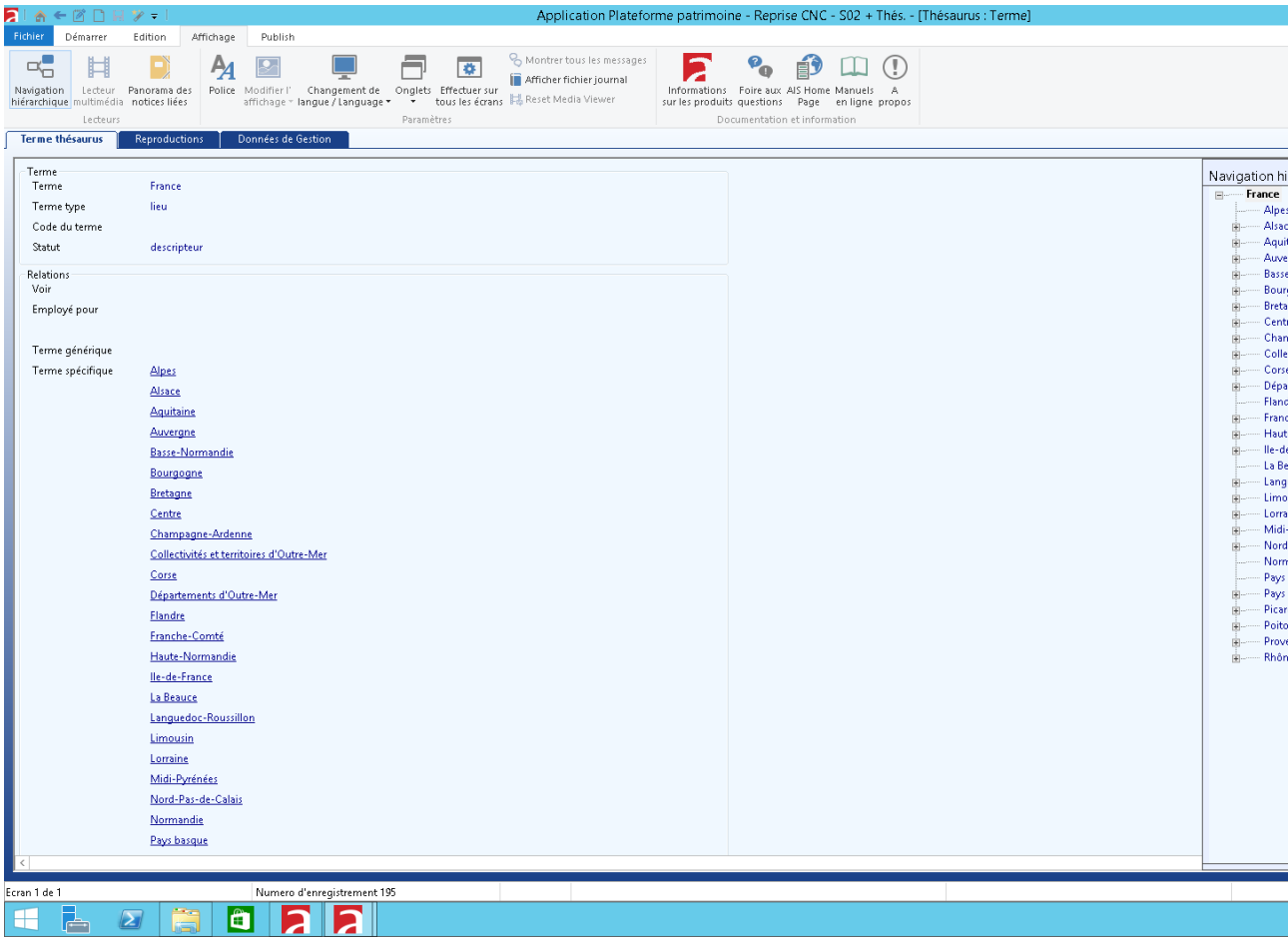
Si l'administrateur de la base choisit cette option, des champs sémantique de et Facteurs sémantiques sont ajoutés à la base de données Thesaurus dans la notice des descripteurs.

Il est alors possible de coordonner les termes comme pour chacune des relations. Le concept ainsi crée est défini dans *Terme*, et les facteurs sémantiques dans les occurrences du champ *Facteurs sémantiques.* Dans la notice d'un facteur sémantique, le champ *Facteurs sémantiques de* résume les concepts (par le biais d'un champ répétable) qui font de ce terme un facteur sémantique.

Visualisation des relations

En choisissant l'option Affichage / Navigation Hiérarchique, il est possible d'obtenir à gauche les relations hiérarchiques du thesaurus.

Exemple du thesaurus géographique du CNC.



Fonctionnalités de recherches liées au thesaurus

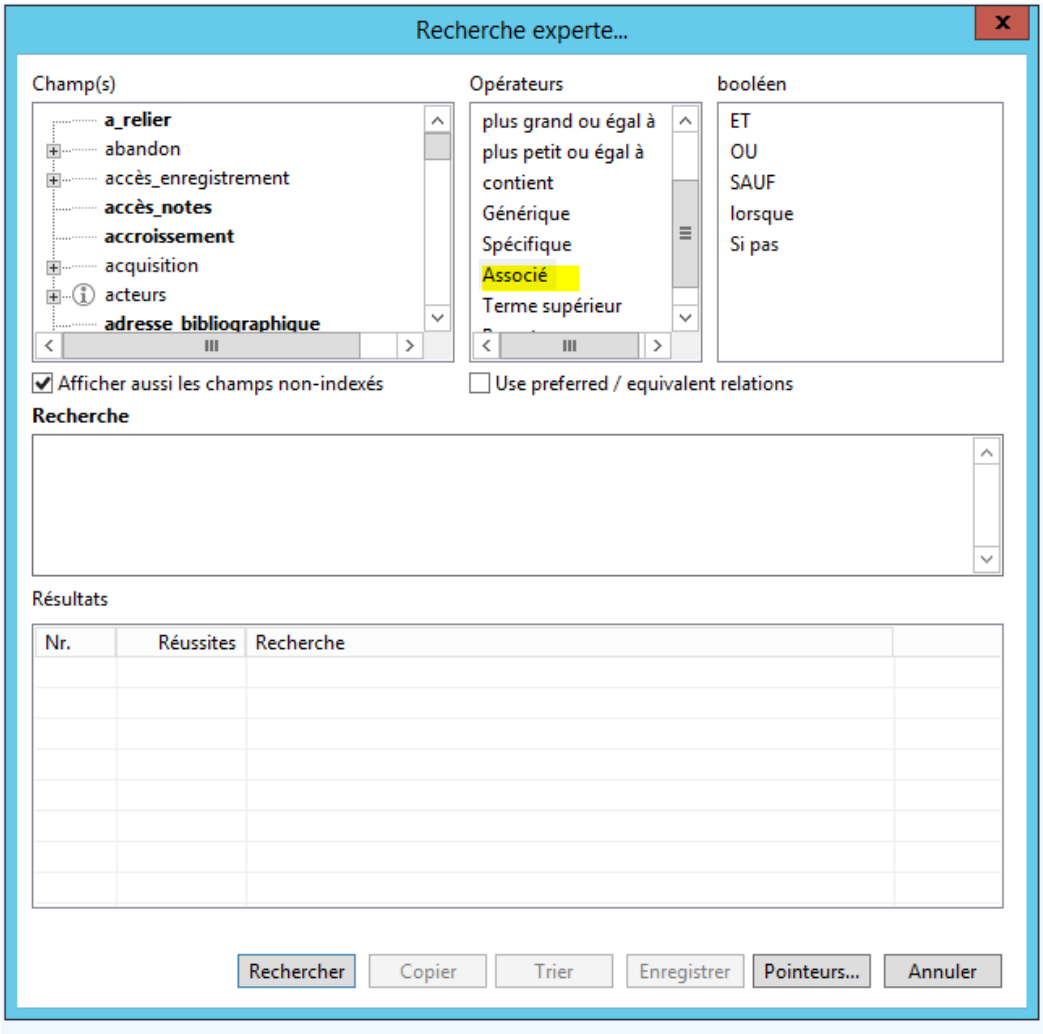
Utiliser/Utilisé pour S'appliquent après avoir coché l'option *Inclure les relations* à l'*Etape 3* de l'*Assistant de recherche*.	**Remplacement de terme** - Je tape à l'intérieur du champ *Terme sujet* un terme non-préféré, celui-ci sera automatiquement remplacé par le terme préféré approprié dès que je quitte la zone de saisie. Par exemple, si je remplis le champ bike, il est remplacé instantanément par bicycle **Extension de la recherche** – Je tape « bike » et j'ai défini ce terme comme étant un terme non-préféré subordonné à bicycle. Adlib récupère alors les notices dans lesquelles se trouve le terme *bicycle*. Je trouverai donc ces notices même en cherchant sur bike. Ce remplacement du terme non-préféré par le préféré ne se produit pas si vous faites votre recherche dans le *Thésaurus*

	lui-même (ou toute autre base de données d'autorité)
Terme spécifique/Terme générique en cochant l'option *Inclure les spécifiques* avant de cliquer sur le bouton *Toutes les clés*. en cliquant sur le bouton *Afficher* après avoir coché l'option *Inclure les spécifiques*	**Extension de la recherche** - Pour élargir la recherche, je peux la faire porter non seulement sur les **termes principaux** affichés dans la liste, mais aussi sur tous leurs **termes génériques** (que la liste de ceux-ci soit déployée ou non) On peut aussi faire porter la recherche sur le **terme sélectionné** ainsi que sur tous ses **termes spécifiques** (que la liste de ceux-ci soit déployée ou non). Un des termes spécifiques peut également être sélectionné pour faire porter la recherche uniquement sur lui et ses propres spécifiques.
Terme associé	Disponible seulement en recherche experte
Terme équivalent Il faut cocher l'option *Inclure les relations* à l'*Etape 3* de l'*Assistant de recherche*.	**Extension de la recherche** - Au moment de la récupération des notices répondant aux clés de recherche tapées, les notices qui ne contiennent pas la clé elle-même, mais contiennent un *Terme équivalent* seront elles-mêmes récupérées. Par exemple, si je cherche bicyclette, les notices qui contiennent l'équivalent anglais *bicycle* seront aussi sélectionnées.
Facteurs sémantiques	Nécessité de paramétrage par l'administrateur fonctionnel – Non disponible dans notre version

Recherche incluant les termes associés

Réponse d'Axiell 2016-07-29 :

Les termes associés ne sont pas intégrés dans la recherche comme les termes hiérarchiques et équivalents. Ce qui s'explique de leur typage. Il y'a la possibilité de recherche sur les termes associés via la Recherche experte (Adlib for Windows et Axiell Collections):



Extrait du Manuel sur la recherche associée :

« L'opérateur *related*, dans le langage d'interrogation, est utilisé en cas de fichier autorité, ou de catalogue lié à une base de données autorité par le champ dans lequel vous faites une recherche. La base de données autorité dont nous parlons est de la sorte de celles dont les termes sont organisés hiérarchiquement, à l'aide de termes reliés entre eux; c'est souvent le signe qu'il y a un thésaurus. Quand vous cherchez en utilisant *related*, vous cherchez sur le terme lui-même et sur tous les termes qui lui sont liés au même niveau, conformément à ce qui est spécifié dans la zone de saisie *Terme lié* (dans le thésaurus). C'est pourquoi vous ne pouvez pas chercher avec *related* sur le champ *titre*, mais c'est tout à fait possible avec, par exemple, *keyword.contents* (Adlib Library), *object_name* (Adlib Museum), *content.subject* (Adlib Archive), ou *term* (champ de Thesaurus):

```
object_name related "orange juice"
```

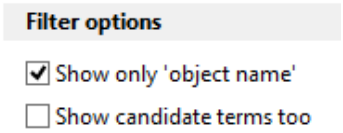
Dans cet exemple, seront récupérées toutes les notices qui sont liées au terme 'orange juice', comme les notices des termes : *apple juice* ou *pineapple juice*. »

Il est à noter que pour toute recherche, la troncature est installée par défaut. Il faut inclure les guillemets '' '' pour une recherche sur terme exact. Etant dans une base de données SQL, on peut faire une troncature à droite / à gauche / au milieu avec précision d'un *.

Gestions des statuts

Au sein d'un catalogue, les termes qui ont pu être forcés pour intégrer la base Thesaurus en cliquant sur le bouton Forcer notice se voient assignés le statut de candidat. Pour cela un champ doit avoir été paramétré dans la base de données par l'administrateur système.

Une option *Afficher aussi les termes candidats* apparaîtra dans la fenêtre R*echerche des termes pour le champ…* Ces termes forcés ne seront ajoutés à la liste affiché des clés (de recherche) trouvées qu'à condition d'avoir coché cette option. Si l'option n'est pas cochée, la liste n'affiche que la liste des clés de recherche non forcées, c'est-à-dire qui ont été créées à partir de la base Thesaurus.



La gestion des termes s'effectue dans la base de données Thesaurus.

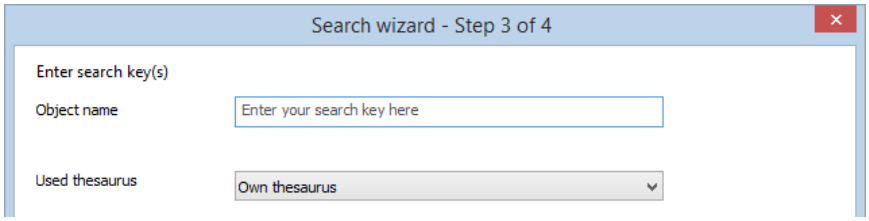
Un champ statut apparaît à côté de la definition du terme, qui indique son statut en cours (voir fiche descripteur).

Statuts possibles :

0. **Indéfini –** Aucun statut encore assigné
1. **Terme préféré approuvé**
2. **Terme non-préféré approuvé**
3. **Candidat** – Statut que reçoit tout terme force vers le Thesaurus à partir d'une autre BDD.
4. **Obsolète** – Possibilité de réutilisation
5. **Rejeté** – l'nformation du rejet mémorisée empêche de saisir à nouveau le terme.

Utilisation de thesaurus externes

Il est possible d'utiliser des thesaurus externes. **L'administrateur fonctionnel peut prévoir d'intégrer au thésaurus des champs liés spécifiques dans lesquels il est possible chercher des termes dans un autre thésaurus ou parmi plusieurs thesaurus différents**, avec lesquels on peut procéder à la validation de champs ou à la mise à jour de bases de données complètes. **Cette option n'est pas encore paramétrée dans notre version.**



Exemple de recherche sur plusieurs thesauri

Provenance des thesauri ajoutés (= thesauri externes)

Ces thesauri standard « peuvent être utilisés à l'aide d'un CD-ROM, depuis internet ou un intranet, ou bien en utilisant votre propre disque dur. De ce fait, il n'y a pas d'obligation de convertir les données pour faire une mise à jour, il est possible de se servir simplement d'un autre CD-ROM, ou de faire la validation dans un autre thésaurus situé dans un autre endroit de votre système ».

Condition de réalisation : Le thésaurus ajouté au champ devra être accessible en tant que base de données Adlib indexée. C'est à dire que le champ thésaurus avec lequel un lien est établi doit faire partie du thésaurus externe Adlib et avoir été indexé.

Dans la pratique, il y'a deux possibilités pour ajouter un thesaurus externe.

- Créer une nouvelle base de données Adlib. Les données du Thésaurus externe sont importées dans cette base de données. Il y a une connexion des champs de catalogue vers aussi cette base de données.
- Faire une connexion vers un webservice (SRU Gateway). L'import est direct et nécessite un fichier de mapping entre les champs webservice et les champs Adlib.

Le résultat de l'import est que la source du nouveau thesaurus est créée au sein d'Adlib. Les fiches des termes du thesaurus sont également créées. Les relations entre les termes peuvent faire partie de l'import.

Formats d'imports de thesaurus externe

Question : Quels sont les formats d'import compatibles avec Adlib ?

Réponse Axiell :

Il y a beaucoup de format standard compatible pour Adlib. Pour un SRU Gateway, il faut que la source ait un webservice SRU. Pour import dans une base de données Adlib, il y par exemple csv (fichier Excel), Access DB, SQL DB, Oracle… L'important est que les données ont une structure lisible par un ordinateur. Un fichier Word avec hiérarchie en couleur et tab est illisible pour un import.

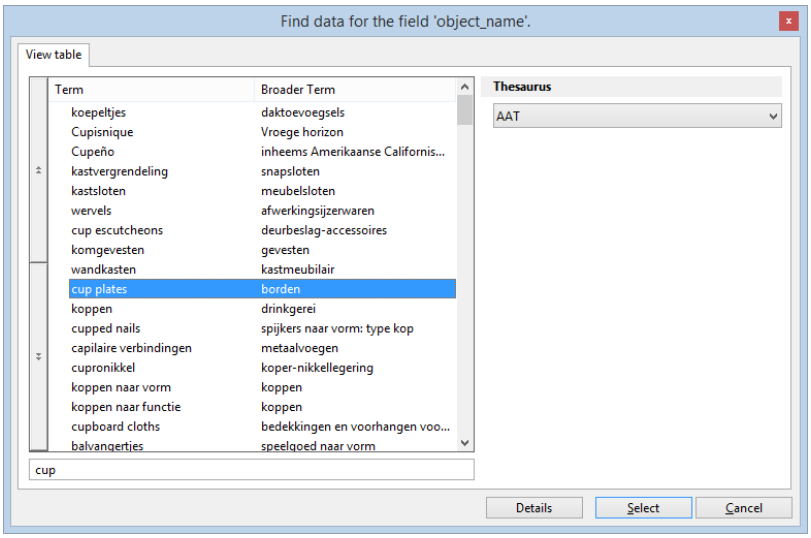
Maintenance du thesaurus externe

Question : La maintenance d'un thesaurus externe doit-elle se faire au sein de l'outil ?
Réponse Axiell : Il est possible de la faire au sein de la base de données Adlib en passant par la source (le thesaurus/domain en question), ou bien il est possible de vider la base de données, puis de re-importer une nouvelle version du fichier de source.

Cas d'utilisation des thesaurus ajoutés

Au moment de la saisie des notices

Si un champ est lié au thésaurus standard, comme *Nom d'objet* dans Adlib Museum, il est possible d'utiliser le bouton *Consulter la liste* pour ouvrir la fenêtre *Recherche des termes pour le champ...* et faire la recherche d'un terme, mais il se peut aussi que cette fenêtre s'ouvre automatiquement, quand le terme ne figure pas dans ce thésaurus.



Dans le coin supérieur droit de cette fenêtre, une liste déroulante *Thésaurus* peut être présente (seulement si elle a été paramétrée pour le champ en question par votre administrateur fonctionnel). Cliquez sur cette liste déroulante et choisissez la base de données thésaurus souhaitée. La liste des termes proposés est immédiatement ajustée.

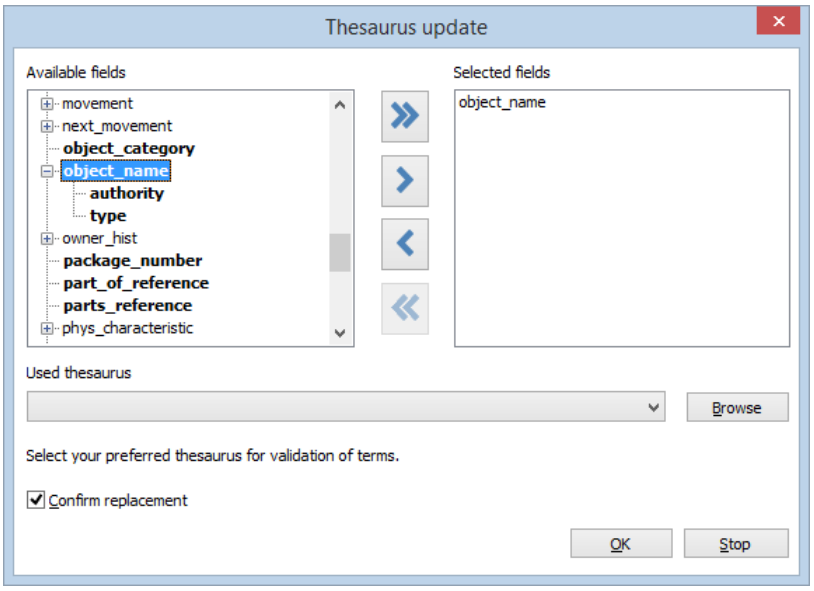
Notez bien que pendant votre recherche en ligne dans le thesaurus externe, la liste pourra contenir des termes paraissant ne pas suivre l'ordre alphabétique, mais cela est dû au fait que la clé de recherche est lancée dans plusieurs champs à la fois tandis que (sur l'onglet *Voir hiérarchie*) seul le champ *Terme* est affiché.

Pour une mise à jour globale

Une mise à jour globale permet de mettre à jour les champs spécifiés dans toutes les notices marquées de la base de données en cours d'utilisation, et peut concerner un thésaurus différent du thésaurus principal. Cela signifie que chaque terme dans les champs sélectionnés étant défini comme un terme "utilisé pour" (non-préféré), sera remplacé par le terme préféré approprié.
Ex : Le terme "motorbike" apparaît dans des notices du catalogue. Si vous marquez ces notices et sélectionnez les champs dans lesquels ce terme apparaît, alors "motorbike" sera remplacé si le nouveau thésaurus définit le terme "motorbike" comme un terme non-préféré. Le terme préféré, c'est-à-dire. "motorcycle", remplacera "motorbike" dans les notices marquées. Un tel remplacement n'aura pas lieu si "motorcycle" a été défini comme un terme non-préféré dans votre thésaurus.

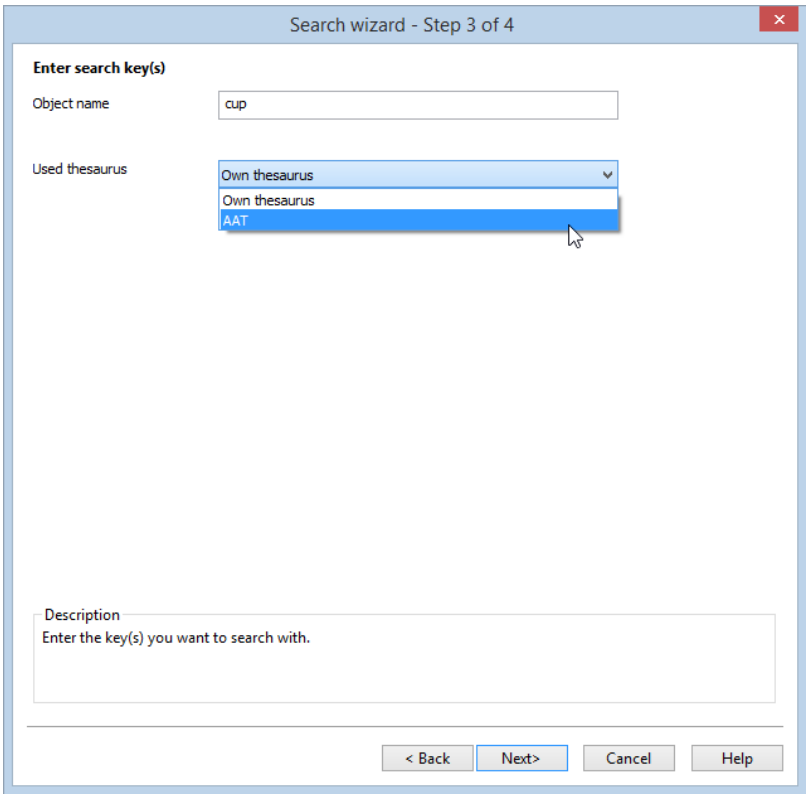
Ainsi, une nouvelle notice de thésaurus est ajoutée à votre propre thésaurus, et l'ancienne notice de thésaurus reste intacte, mais aucune relation n'a été établie entre l'ancien et le nouveau terme. Aucune notice-miroir n'est créée pour la nouvelle notice.

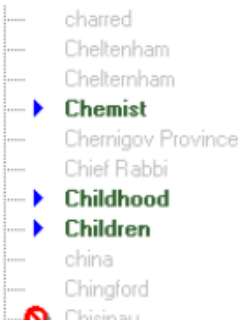
- Cliquer sur le bouton *Mise à jour du thésaurus* pour ouvrir la fenêtre *Mise à jour du thésaurus* ,qui se trouve dans la barre d'outils de l'affichage simplifié et de l'affichage détaillé des notices. Cette fenêtre vous permettra de procéder à la *Mise à jour du thésaurus*.
- Cliquer sur le bouton *Recherche* pour trouver le thésaurus souhaité sur votre système, ou choisissez-en un dans la liste déroulante *Thésaurus utilisé*. Si vous voulez faire une mise à jour automatique, cela revient à décocher l'option *Confirmer remplacement* (ce qui peut bien sûr comporter des risques).



Pour une recherche générique

Il est également possible d'utiliser un thesaurus externe pour une recherche avec l'*Assistant de recherche*, si le champ en question a été paramétré par l'administrateur fonctionnel dans cette optique. Une liste déroulante est alors disponible à l'*Etape 3* de l'*Assistant de recherche*, vous permettant de choisir un thésaurus. Si votre clé de recherche comprend plusieurs termes, une liste de ces termes s'ouvrira comme auparavant, mais les clés dans le thésaurus sélectionné qui proviennent de votre propre base de données seront en surbrillance. Les clés grisées ne vous permettront pas de trouver de notice de catalogue, aussi vous faut-il choisir la clé en surbrillance.





Liste des clés trouvées dans l'Assistant de recherche avec le thésaurus choisi

Création, Gestion et maintenance du thesaurus

Adlib n'est pas un système de gestion et de maintenance de thesaurus, mais il possède un module qui intègre des fonctionnalités de gestion et de maintenance très avancées, d'un niveau proche des logiciels autonomes. La question des systèmes de gestion de thesaurus spécialisés (Thesaurus Management System) est abordée dans le document « étude de faisabilité ». Il n'est ici question que des fonctions de construction et de gestion de thesaurus d'Adlib.

Construction de thesaurus

En tant qu'outil de construction de thesaurus, le module d'Adlib permet d'effectuer des opérations de vérification automatique de vocabulaire :

- Il gère les relations entre les termes et permet de corriger des relations qui seraient incohérentes (ex : un terme non préféré indiqué comme descripteur). Il n'autorise pas deux statuts différents à un seul et même terme. Il permet que le même descripteur ne soit pas à la fois générique et spécifique d'un autre descripteur. Le logiciel bloque l'établissement de plusieurs types de relations entre les deux mêmes termes au moment de la saisie.
- Il permet la création d'une notice miroir à la création de nouveau descripteur (création de la réciproque à l'enregistrement des données). L'existence d'un terme non-préféré dans une notice fait obligatoirement référence à un terme préféré. A la création du terme préféré, si le terme préféré n'existe pas, la fiche sera créée automatiquement par Adlib. S'il existe, la relation sera inscrite automatiquement dans la fiche du descripteur sans avoir à l'ouvrir pour saisir la relation réciproque.
- Il établit une fusion des résultats sur les listes présentes dans la base thesaurus dans le cas où un descripteur se retrouve dans plusieurs domaines.
- Il gère les arborescences qui seront disponibles à la recherche. Il permet également de modifier les relations hiérarchiques par un glisser-déposer dans l'arborescence.
- Il permet de vérifier l'intégrité des termes.
- Il vérifie qu'aucun terme ne soit relié à lui-même. Le logiciel bloque automatiquement la création de relations circulaires lors de la saisie.

- Il permet que le même descripteur ne soit pas lié de manière hiérarchique et associative à un autre descripteur. Le logiciel bloque l'établissement de plusieurs types de relations entre les deux mêmes termes au moment de la saisie.
- Il vérifie que les non-descripteurs ne sont accompagnés d'aucune relation sémantique. Le logiciel bloque les relations hiérarchiques et associatives dans la notice d'un terme ayant le statut de non-descripteur.

Adlib ne permet pas de rapprocher des termes libres qui auraient été saisis au préalable dans le but de les fusionner et de les ordonner pour constituer un thesaurus. Il sera donc nécessaire de prévoir une étape d'ingénierie de thesaurus au préalable (intervention humaine). Elle impliquerait :

- d'opérer des choix dans les langages préexistants, des rapprochements et des fusions de concepts et de termes.
- d'opérer des choix en termes de hiérarchisation et de construction des facettes et microthesauri.
- De définir les types de relations utilisées, les niveaux de relation et les termes qui constitueront des descripteurs.
- De saisir ces relations dans l'outil.

Présentation de thesaurus

Question : *Adlib peut-il faciliter la production imprimée ou numérique de différents types de présentation des termes (à destination des indexeurs) ? Exemple : Présentation alphabétique des termes du thesaurus, présentation hiérarchique, présentation thématique, liste permutée.*

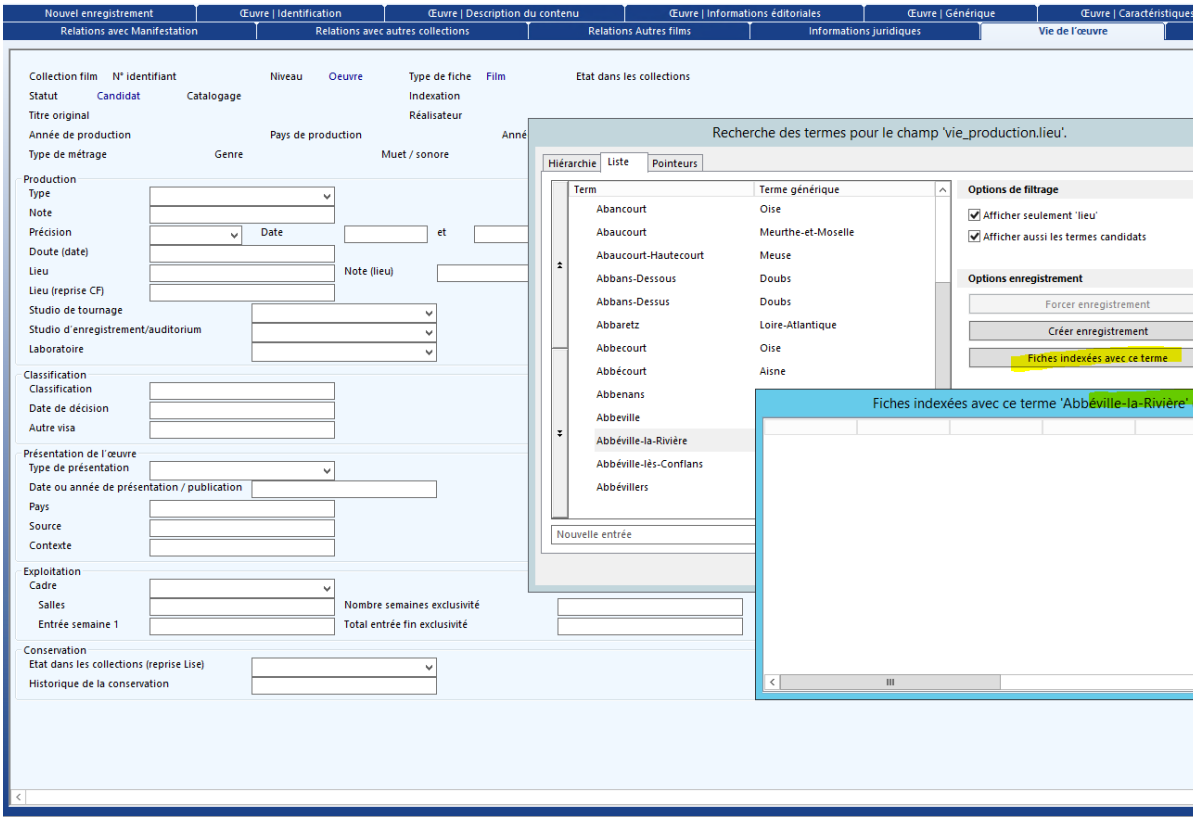
Réponse Axiell :

Il y a la possibilité de créer des fichiers d'export pour imprimante ou fichier. Dans le standard, il y a un export vers une liste alphabétique et une vue hiérarchique. Vous pouvez tester sur le serveur. Il est possible d'ajouter des formats d'export. Il est aussi possible de configurer l'outil Office Connect pour que les indexeurs peuvent avoir les listes et recherches directement dans Word / Excel.
(http://www.adlibsoft.com/support/downloads/adlib-office-connect).

Compteur d'occurences d'indexation

Question : Proposez-vous des compteurs permettant de visualiser les occurrences d'indexation relatives à un terme lié à un domaine ?

Axiell 2016-07-29 :
Il y a un compteur dans la liste d'index :



Les résultats de recherche contiennent un marquage pour les termes non-utilisé :

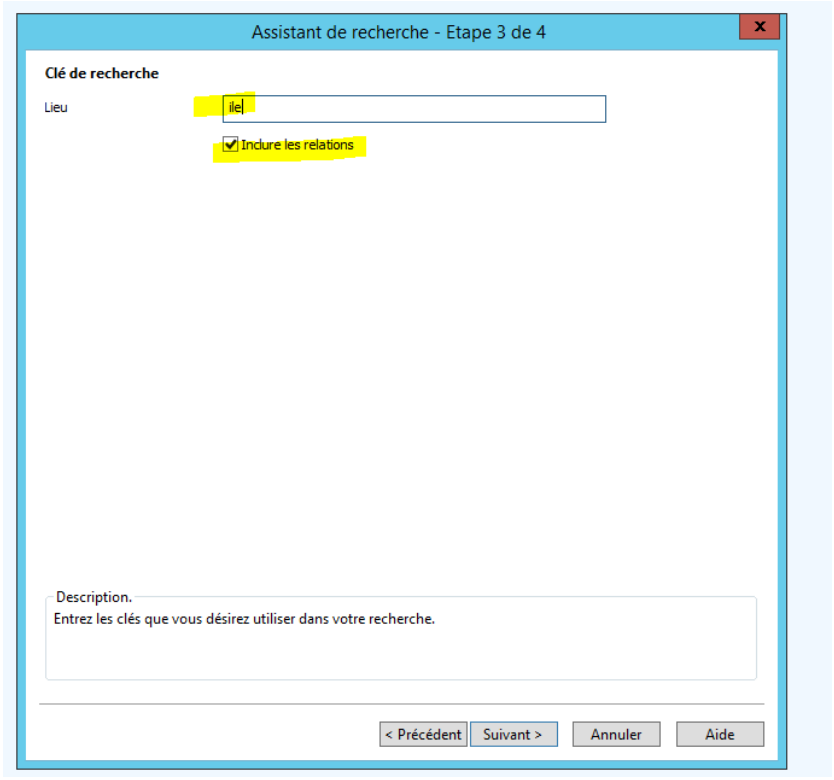
Maintenance du thesaurus : Prise en compte des requêtes vides

Question : Proposez-vous des compteurs / outils statistiques permettant d'enregistrer les requêtes qui n'aboutissent pas lors d'une recherche, dans le but de les prendre en compte ultérieurement comme termes pour la mise à jour du thesaurus ?

Axiell 2016-07-29 :
Pas dans Adlib for windows et Axiell Collections.
Nous recommandons d'utiliser la recherche avec la hiérarchie. De cette manière le risque de tomber sur aucun résultat, car recherche sur un terme non-utilisé est réduit, à la condition que la hiérarchie est construite dans le thésaurus.

Exemple :
Une exposition à lieu à Antony. L'utilisateur ne se rappelle plus du lieu concret, mais il sait que l'exposition a eu lieu quelque part 'Ile-de-France'. Il peut donc rechercher sur Ile-de-France avec la hiérarchie et il trouvera l'exposition d'antony, même si le terme 'Ile-de-France' n'est pas inscrit explicitement dans la notice de l'exposition :


Données exposition
Related exhibitions
Documentation
Objets liés
Prêts
Reproductions
Données de Gestion
Exposition
Titre
Exposition
Numéro
Titre alt.
Type
Date d'entrée
2017-07-29
Date de fin
Start date action
End date action
Action type
Person responsible
Role
Particularités
Pouvoir organisateur
Nom
Adresse
Code postal
Pays
Téléphone
Agenda
Date d'entrée
Date de fin
Arrivée
Lieu
Particularités
2017-07-29
Antony


Assistant de recherche - Etape 3 de 4
Clé de recherche
Lieu
ile
Inclure les relations
Description.
Entrez les clés que vous désirez utiliser dans votre recherche.
< Précédent
Suivant >
Annuler
Aide

ANNEXE 4 – EVALUATION DES COLLECTIONS FILMS ET NON FILMS DES PARTENAIRES

	Eléments films	Eléments non-films
CNC	- 110000, dont 7410 numérisés	DOCUMENT ECRIT - matériel publicitaire de film. Nombre indéterminé - 12400 Monographies - 420 titres de Périodiques DOCUMENT ICONOGRAPHIQUE - Affiches. Nombre indéterminé. - 500 Plaques de verres
Cinémathèque Française	40000 films 7256 DVD, 1238 VHS, 83 Bluray Indéterminé.	DOCUMENTS ECRITS - 21000 monographies - 490 titres de périodiques - 30000 dossiers d'archives - 870 matériels publicitaires - Brevets d'invention, dessins techniques, archives. Indéterminé. DOCUMENT ICONOGRAPHIQUE - 23000 affiches - 500000 photos - 14500 dessins - 25000 Plaques de lanterne OBJETS 3D - 42000 appareils - 21000 costumes - 2300 objets et éléments de décor
Cinémathèque de	45000 films	DOCUMENTS

Toulouse	2500 films en DVD 3500 films en VOD	ICONOGRAPHIQUE - 15000 ouvrages 1500 titres périodiques DOCUMENTS ECRITS - 75000 affiches - 500000 photographies

Soit près de **200 000 documents films** et **693 640 documents non-films recensés** (ne comptant pas le nombre de périodiques effectivement indexés et les fonds au nombre indéterminé).