

Indexation de documents

Ingénierie Documentaire
<http://doc.crzt.fr>



Stéphane Crozat

1 septembre 2016

Table des matières



I - Introduction à l'indexation	3
1. Définition de l'indexation	3
2. Indexation de fragments documentaires numériques	4
3. Principes de l'indexation	5
II - Introduction aux techniques d'indexation et de recherche	6
1. Techniques d'indexation	6
2. Thésaurus	8
3. Ontologies	10
4. Techniques de recherche	10
III - Exercice :	13
Bibliographie	14

Introduction à l'indexation

I



L'indexation a pour principal objectif de rendre accessibles des informations, que l'on repère pour cela au moyen d'index. L'indexation est le processus selon lequel le contenu d'un document est analysé pour être ensuite reformulé dans une forme permettant d'accéder au contenu et de le manipuler. Le terme d'indexation qualifie à la fois le processus et son résultat. Une indexation est par conséquent la description d'un document effectuée dans la perspective d'une utilisation et exploitation données.

L'indexation repose traditionnellement sur deux étapes clairement distinguées :

- une étape d'analyse conceptuelle : le contenu est analysé et interprété par un documentaliste pour définir les principaux concepts permettant de le caractériser ;
- une étape de reformulation documentaire : l'analyse conceptuelle permet au documentaliste de reformuler le contenu dans une forme permettant sa manipulation.

(Bachimont, 2007 *)



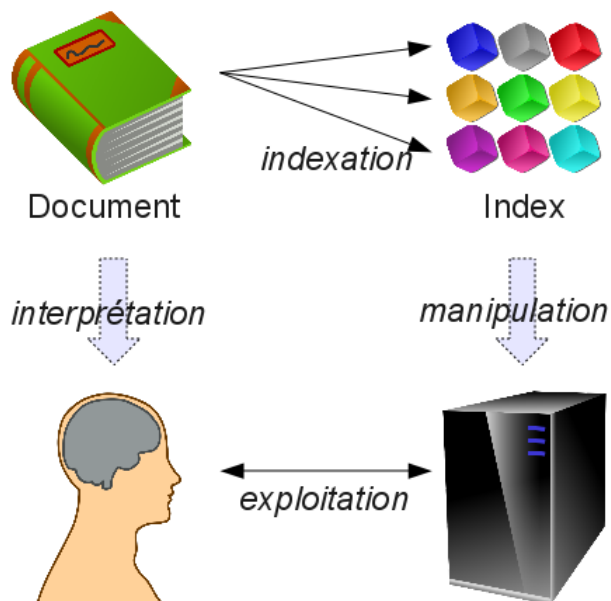
1. Définition de l'indexation

Définition

L'indexation est un processus consistant à reformuler le contenu d'un document sous une forme plus adaptée à son exploitation dans une application donnée (par exemple la recherche).

Dans le contexte du numérique :

- Les documents sont destinés à être interprétés par des êtres humains.
- Les index sont des représentations des documents destinés à être manipulés par des ordinateurs



Principe de l'indexation

👉 Exemple

- Exemple d'exploitation de l'indexation : chercher, réutiliser, recomposer, etc.
- Exemple de formes d'index : liste de termes, langue naturelle, formalisme logique, etc.

2. Indexation de fragments documentaires numériques

De l'analogique au numérique

En contexte analogique l'unité d'indexation (ce que l'on référence) correspond exactement au document physique complet :

- on ne sait pas manipuler de sous-ensemble plus fin
- on doit gérer chaque support physique séparément

Le numérique permet :

- d'*intégrer* l'indexation (une indexation unique pour des contenus multiples)
- d'indexer l'*intérieur* du document (tout fragment de contenu devient adressable).

Intégration

Les contenus étant unifiés par la numérisation, il devient possible de gérer une indexation également unifiée y compris pour des contenus hétérogènes.

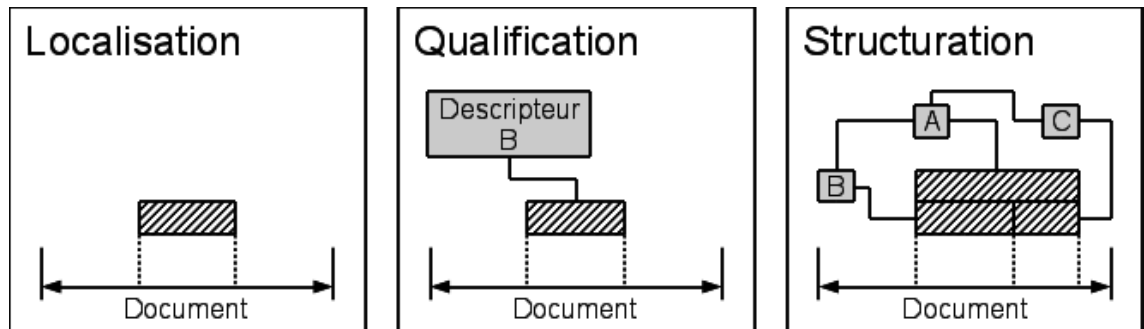
Indexation fine du contenu

Toute information possédant une adresse exploitable par le processus d'indexation devient accessible, donc tout segment arbitraire du contenu (fragment) est potentiellement indexable.

3. Principes de l'indexation

L'indexation de documents numériques repose sur 3 principes :

- La localisation : qu'est ce qui est indexé ?
- La qualification : par quelle valeur ?
- La structuration : dans quelle organisation ?



Processus d'indexation

La localisation

Le problème est de repérer les unités signifiantes : Il n'existe pas *a priori* d'unités signifiantes, les unités signifiantes résultent d'une interprétation.

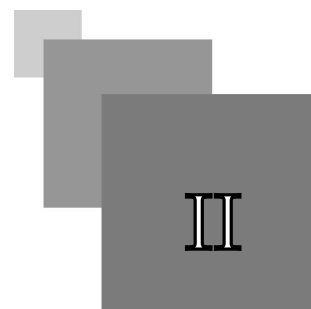
La qualification

La qualification est la condition sous laquelle on est capable d'attribuer une valeur, une utilité à une unité d'indexation, elle justifie son exploitation.

Structuration

Il s'agit de rassembler les différentes qualifications effectuées en un système organisé.

Introduction aux techniques d'indexation et de recherche



1. Techniques d'indexation

Différentes stratégies d'indexation

- Indexation plein-texte : des mots clés sont extraits automatiquement du contenu
- Méta-données documentaires : des propriétés documentaires sont renseignées manuellement
- Référentiels : les documents sont associés à des entrées de référentiel (taxinomie, thésaurus)
- Ontologies : Le contenu des documents est représenté logiquement
- Folksonomies : Les documents sont associés à des tags (mots clés) posés par les utilisateurs finaux
- ...

Exemple : Méta-données Dublin Core

Le Dublin Core est une norme ISO proposant un jeu de méta-données très générales pour la description des ressources numériques. Il comporte 15 descripteurs de base (titre, créateur, mots-clés, description, date, langue, ...).

Propriétés	
* Nom:	nf29cas.pdf
Type de contenu:	Adobe PDF Document
Codage:	UTF-8
Titre:	Étude de cas : base de contenus pédagogiques
Description:	
Auteur:	Stéphane Crozat
* Éditeur:	UTC
* Contributeur:	Manuel Majada
* Type:	Document Scenari/Opale
* Identifiant:	nf29cas.pdf
* Source:	http://www4.utc.fr/~nf29/2008p/src
* Couverture:	France
* Droits d'accès:	CC BY-NC-SA (http://creativecommons.org/licenses/by-nc-sa/4.0/)
* Sujet:	Ingénierie documentaire

Interface d'indexation Dublin Core (Alfresco)

☞ Exemple : Référentiels

Les valeurs des méta-données peuvent être contrôlées par des référentiels, ici mots-clés et couverture.

Titre*

Description

Mots-clés

Vie quotidienne	Sélectionnez une valeur.	Ajouter
Sciences Humaines	Astronautique	
Société	Électronique	
Technologie	Energie	

Sélection:

Technologie/Energie

Droits

Source

Couverture

Créé le 24/02/10 15:03

Indexation selon des référentiels géographique et thématique (Nuxéo-DM)

2. Thésaurus

Définition : Glossaire

Un glossaire est une liste de termes associés à leurs définitions, en relation avec un *domaine* particulier.

Les termes sont généralement présentés par ordre alphabétique, c'est une sorte de dictionnaire contextuel, et limité à des mots spécifiques ou complexes.

Définition : Taxinomie

Une taxinomie (parfois taxonomie, *taxonomy* en anglais) est un système de classification *hiérarchique* de l'information (le terme est issu à l'origine de la classification des espèces vivantes.)

La relation de structuration est de type *est-un (is-a)* : épagneul *est-un* (chien *est-un* (canin *est-un* (mammifère))).

Définition : Thésaurus

Un thésaurus est un *vocabulaire normalisé* concernant un domaine particulier, en général pour un objectif particulier, par exemple l'indexation.

Il se présente sous la forme d'un *réseau* de termes reliés entre eux par des associations sémantiques : synonymie, hyper/hyponymie, ...

Les termes peuvent également être décrits par une définition en texte libre, permettant d'en faciliter la compréhension, il a alors également valeur de glossaire.

✂ *Méthode : Élaboration d'un thésaurus*

1. Déterminer les termes qui peuvent être pris en compte :
 - Descripteurs utilisables
 - Non descripteurs ne pouvant pas l'être
2. Associer des relations sémantiques entre ces groupes :
 - Hyperonymie ou hyponymie
 - Synonymie
 - Association

✂ *Méthode : Synonymie*

Choisir un terme préféré parmi les termes synonymes, et déclarer les autres comme non-descripteurs :

- Logiciel, descripteur, employé-pour (Software, Application)
- Software, non-descripteur, employer (Logiciel)
- Application, non-descripteur, employer (Logiciel)

Utiliser la relation de synonymie pour gérer les polysémies :

- Bateau, non descripteur, employer (Bateau pavé, Navire)

✂ *Méthode : Hyper/hyponymie*

Ces relations permettent de déclarer les relations de généralité/spécificité entre les descripteurs.

- Europe, descripteur, hyperonyme-de (France, Allemagne)
- France, descripteur, hyponyme-de (Europe)
- Allemagne, descripteur, hyponyme-de (Europe)
- Oiseau, descripteur, hyperonyme-de (Canard)
- Canard, descripteur, hyponyme-de (Oiseau)

Ces relations peuvent être ontologiques (sorte-de) ou méréologiques (partie-de). Ces deux types peuvent être distingués pour affiner la description :

- France, descripteur, partie-de (Europe)
- Canard, descripteur, sorte-de (Oiseau)

✂ *Méthode : Association*

L'association est peu précisément définie dans les thésaurus, c'est en fait toute relation qui n'est pas l'une des précédentes :

- Relation entre agent et action
- Relation entre action et objet
- Relation entre des termes co-occurents
- ...

3. Ontologies

Définition : Ontologie

Une ontologie (en informatique, ne pas confondre avec le concept philosophique) est :

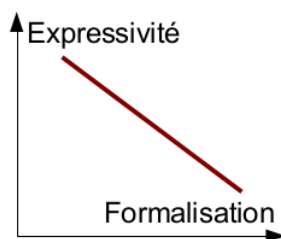
- une taxinomie ou un thésaurus
- *plus* une représentation dans un langage formalisé doté d'une sémantique formelle, permettant de faire des inférences sur les termes

Une ontologie est donc une *conceptualisation calculable* d'un domaine, qui poursuit deux objectifs :

- Expliciter une *compréhension* commune de notions ou concepts d'un domaine par les hommes (expression d'un consensus dans une communauté, ce sur quoi tout le monde est d'accord)
- Permettre une *opérationnalisation* des concepts par les machines (formalisation permettant des inférences conformes à la sémantique des concepts)

Remarque

Plus on peut exprimer de choses, moins on peut les formaliser et calculer sur elles ; plus c'est formalisé et calculable, moins c'est expressif.



Dualité expressivité/formalisation

Fonction des ontologies

L'objectif des ontologies est d'améliorer l'indexation :

- en augmentant les taux de réponse, typiquement pour trouver un document qui ne contient pas un terme recherché, mais un synonyme ou un hyperonyme.
- en diminuant les taux de bruit, typiquement pour éliminer un document qui contient un homonyme.

Complément : Schémas de classification : thésaurus, taxonomie, ontologie...

<http://www.dia-logos.net/ressources/2009-11/schemas-de-classification-thesaurus-taxonomie-ontologie>

4. Techniques de recherche

- Requêtes par la formulation de mots clés libres qui sont appariés avec tous les index

- Extension avec utilisation de connecteurs logiques (AND / OR)
- Extension avec des modèles probabilistes (indice de pertinence)
- Requêtes par la formulation de mots clés libres ou contrôlés index par index
 - auteur = X AND/OR titre=Y ...
- Recherche par la navigation
 - dans des référentiels (ontologies, thésaurus, etc.)
 - dans des "nuages" de mots-clés ayant une proximité statistique
- ...

☞ Exemple : Recherche dans des référentiels

Un document peut être indexé selon un ou plusieurs référentiels, qui peuvent ensuite être navigués pour retrouver les documents associés. Un référentiel se comporte comme une organisation virtuelle de documents.


The screenshot shows the Nuxeo-DM interface. The top navigation bar includes the Nuxeo logo, user identity (ICI55-01), and search options. The main content area is titled "Navigation virtuelle par thème" and displays a table of documents under the "Technologie > Energie" category.

	Titre	Dernière modification	Auteur	Cycle de vie
<input type="checkbox"/>	Albert Einstein	25/02/10 22:58	ICI55-01	En projet
<input type="checkbox"/>	La théorie de la relativité	25/02/10 22:59	ICI55-01	En projet

Below the table are buttons for "Coller", "Supprimer", "Copier", and "Ajouter à mon lot de documents". The sidebar on the left shows a hierarchical tree of subjects, with "Technologie" expanded to show sub-categories like "Astronautique", "Électronique", "Énergie", "Industrie", "Informatique", "Robotique", and "Transport".

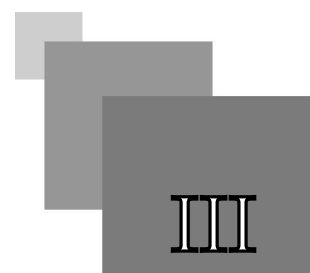
Recherche dans un référentiel thématique (Nuxéo-DM)

The screenshot shows the Nuxeo web interface. At the top, the Nuxeo logo is on the left, and the user's identity 'ICI55-01' and navigation links are in the center. A search bar with the text 'Recherche' and a 'Recherche avancée' link is on the right. Below the header, a left sidebar shows a tree view under 'Navigation par couverture' with 'Europe' selected. The main content area is titled 'Navigation virtuelle par couverture' and displays a table with one search result. Below the table are buttons for 'Coller', 'Supprimer', 'Copier', and 'Ajouter à mon lot de documents'.

<input type="checkbox"/>	Titre	Dernière modification	Auteur	Cycle de vie
<input type="checkbox"/>	 La théorie de la relativité	25/02/10 22:59	ICI55-01	En projet

Recherche dans un référentiel géographique (Nuxéo-DM)

Exercice :



[15 minutes]

En quoi peut-on rapprocher l'indexation de l'opération de balisage que l'on peut effectuer sur un document ? Ont-elles la même finalité ? Utilisent-elles les mêmes outils ? Peut-on dire que l'une est une partie de l'autre ? Un moyen pour l'autre ?

Bibliographie



Bruno Bachimont, *Ingénierie des connaissances et des contenus : le numérique entre ontologies et documents*, Lavoisier, Hermès, 2007