Scénario d'utilisation n°2
Évaluation de l'indexation

Protocole d'évaluation & Exemplier

Auteur : Sabine Barreaux

Date : 12/06/2015

Version : 1

I. Protocole d'évaluation

Le scénario 2 de Termith a pour objectif d'évaluer la performance des différentes méthodes d'indexation proposées par le LINA, par rapport aux pratiques d'indexation en vigueur à l'INIST.

Ce scénario se déroule en 4 phases :

- Phase 1 : évaluation de l'indexation produite par 2 méthodes à partir du résumé dans un domaine (linguistique)
- Phase 2 : évaluation de l'indexation produite par 2 méthodes à partir du texte intégral dans un domaine (linguistique)
- Phase 3 : évaluation de l'indexation produite par 4 à 6 méthodes à partir du résumé dans 5 domaines (linguistique, archéologie, sciences de l'information, chimie)
- Phase 4 : évaluation de l'indexation produite par 6+ méthodes à partir du texte intégral dans 5 domaines (linguistique, archéologie, sciences de l'information, chimie)

Pour chacune des phases, il s'agit d'évaluer :

- la pertinence de chaque mot-clé proposé
- le silence de la méthode, en se basant sur l'indexation INIST

Consignes générales :

- On ne consulte pas l'article en texte intégral quand on évalue l'indexation faite à partir du résumé
- On ne modifie pas le score Termith quand on voit l'indexation INIST pour évaluer le silence de la méthode

1. Évaluation de la pertinence

Cette tâche consiste à déterminer si un mot-clé est pertinent pour représenter la problématique de l'article ou du résumé.

Cette évaluation est formalisée par l'attribution d'un score de 0 à 2 :

- Score 0 : mot-clé non pertinent
- Score 1 : mot-clé pertinent, mais pas dans la forme proposée
- Score 2 : mot-clé pertinent

Cas particuliers :

- Le mot-clé n'est pas pertinent
  ⇨ score 0 + indiquer éventuellement la raison en commentaire (ex : "trop générique", "trop spécifique", etc ...)
- Le mot-clé est pertinent, mais pas dans la forme proposée

    - Si la forme préférentielle est présente également dans l'indexation
      ⇨ score 1 + lien vers forme préférée dans l'indexation
      + score 2 à la forme préférée
    - Si la forme préférentielle n'est pas présente dans l'indexation mais est présente dans le texte
      ⇨ score 1 + indiquer en commentaire : « forme préférée dans le texte : xxxx »
    - Si la forme préférentielle n'est pas dans le texte
      ⇨ score 2 + indiquer en commentaire : « forme canonique absente du texte : xxxx»
- Le mot-clé est pertinent dans la forme proposée (que la forme soit présente ou non dans le texte)
      ⇨ score 2

Synthèse des commentaires les plus fréquents utilisés pour la pertinence :

score	**commentaire**
0	trop générique
	trop spécifique
1	forme préférée dans le texte : xxx
	segmentation erronée
2	forme canonique absente du texte : xxx

2. Évaluation du silence

Une fois l'évaluation de la pertinence terminée pour une méthode, il s'agit de repérer les mots-clés pouvant manquer à l'indexation proposée par cette méthode.
Ces mots-clés manquants sont recherchés dans l'indexation INIST, qui représente l'indexation de référence.
Cette tâche consiste donc à évaluer chaque mot-clé INIST en lui attribuant un score de 0 à 2 :

- Score 0 : le mot-clé ne manque pas à l'indexation Termith
- Score 1 : le mot-clé manque moyennement
- Score 2 : le mot-clé manque absolument

Cas particuliers :

- Le mot-clé ne manque pas :
    - parce qu'il est déjà présent dans l'indexation Termith

⇨ score 0 + lien vers mot-clé Termith correspondant (correspondance exacte ou approximative)

- parce qu'il provient d'une erreur d'indexation, ou qu'il est trop générique, ou que la notion n'est pas abordée dans le texte (donc on ne peut pas juger), ou que l'on aurait mis un autre mot-clé

  ⇨ score 0 + commentaire : « erreur d'indexation », « trop générique », « notion absente du résumé », « préférence pour : xxx»

- Le mot-clé manque moyennement à l'indexation Termith :
  - s'il est présent dans le texte

    ⇨ score 1 + commentaire : « secondaire »
  - s'il n'est pas présent dans le texte

    ⇨ score 1 + commentaire : « implicite » ou « générique » en fonction des cas
- Le mot-clé manque à l'indexation Termith :
  - s'il est présent dans le texte

    ⇨ score 2
  - s'il n'est présent dans le texte

    ⇨ score 1 + commentaire « implicite »
  - s'il n'est pas présent dans le texte sous cette forme

    ⇨ score 2 + commentaire : « forme variante dans le texte : xxx»

Synthèse des commentaires les plus fréquents utilisés pour le silence :

score	**commentaire**
0	erreur d'indexation
	trop générique
	notion absente du résumé
	préférence pour : xxx
1	secondaire
	implicite
	générique
2	forme variante dans le texte : xxx

II. Exemplier

A. Pertinence

- **Le mot-clé est pertinent mais pas dans la forme proposée :**

Le mot-clé est alors considéré comme une variante. Il peut s'agir d'un synonyme, d'une variante orthographique, d'une variante morphosyntaxique, d'un sigle, d'une abréviation ou d'un fragment de texte mal segmenté.

Type de variante	**Mot-clé Termith**	**Forme préférée**
synonyme	*sémiologie*	*sémiotique*
variante orthographique	*irakien* *tsigane*	*iraquien* *tzigane*
variante morphosyntaxique	*culturel* *bakhtiniens* *française*	*culture* *Bakhtine* *France*
sigle	*lsf*	*langue des signes française*
abréviation	*pub*	*publicité*
fragment de texte mal segmenté	*nouvel étudiant* *appelée métaphonie*	*étudiant* *métaphonie*

Lorsque la variante est un fragment de texte mal segmenté, on ajoutera le commentaire "segmentation erronée" en plus du lien vers la forme canonique présente dans l'indexation.
Lorsque la variante est un synonyme ou une variante orthographique, le choix de la forme préférée n'a pas d'importance. Ce qui est important, c'est de ne pas attribuer un score 2 à chaque forme car on perd dans ce cas l'information sur la redondance des mots-clés fournis.

- **Le mot-clé est un nom propre de personne :**

La forme canonique est "nom + prénom" ou "prénom + nom" (le prénom pouvant être en entier ou sous forme d'initiale).
Si le mot-clé est pertinent pour l'indexation, en fonction des cas :

➔ il sera considéré comme une variante, si le nom est proposé sans le prénom et si le prénom est précisé dans le texte (score 1).

➔ il sera accepté tel quel, si le nom est proposé sans le prénom et si le prénom n'est pas précisé dans le texte (score 2).

- **Le mot-clé est au pluriel :**

Si le mot-clé est pertinent pour l'indexation, il n'est pas considéré comme une variante et peut recevoir le score 2.
Si l'indexation propose à la fois la forme au pluriel et la forme au singulier, on donnera par contre le score 1 au pluriel (avec lien vers la forme au singulier) et le score 2 au singulier.

- **Plusieurs mots-clés sont proposés pour la notion à indexer :**

notion à indexer	**MC**	**score**
genre du discours	*genre*	1 + forme préférée dans le texte : "genre du discours"
	discours	2
genre grammatical	*genre*	1 + forme préférée dans le texte : "genre grammatical"
	grammatical	1 + forme préférée dans le texte : "genre grammatical"
domaine spécialisé	*domaine*	1 + forme préférée dans le texte : "domaine spécialisé"
	spécialisé	1 + forme préférée dans le texte : "domaine spécialisé"
quantificateur existentiel	*quantificateur*	1 + forme préférée dans le texte : "quantificateur existentiel"
	existentiel	2

Remarques :

- ❖ Les adjectifs "grammatical" et "spécialisé" ne sont pas suffisamment précis sans le nom qu'ils modifient.
- ❖ L'adjectif « existentiel » est accepté comme mot-clé car apporte une information précise dans l'indexation.
- ❖ Le nom "discours" est suffisamment générique pour constituer un mot-clé autonome

- **Plusieurs formes préférées sont possible pour un même mot-clé :**

Si le mot-clé est pertinent pour l'indexation, mais pas assez précis et qu'il peut être relié à plusieurs formes préférées dans le texte, lesquelles renvoient à des concepts différents.
On lui attribuera le score 0 en précisant en commentaire le motif de rejet et les formes potentielles.

MC	Score	Commentaire
analyse	0	trop générique : renvoie à analyse tracéologique et analyse technologique

B. Silence

- **Le mot-clé INIST est plus spécifique que le mot-clé Termith**

MC Termith	score P	MC INIST	score S	lien vers MC termith
personnages	1 + forme préférée dans le texte : “personnage littéraire”	*personnage littéraire*	0	oui
nom	1 + forme préférée dans le texte: “nom propre”	*nom propre*	0	oui
article	1 + forme préférée dans le texte : “article défini”	*article défini*	0	oui

- **Le mot-clé INIST est plus générique que le mot-clé Termith**

MC Termith	score P	MC INIST	score S	lien vers MC termith
fonctionnements sociolinguistiques	2	*sociolinguistique*	0	oui
perspectives lexicographiques	2	*lexicographie*	0	oui

- **Plusieurs mots-clés INIST correspondent à un mot-clé Termith**

MC Termith	score P	MC INIST	score S	lien vers MC termith
dénomination toponymique	2	*dénomination*	0	oui
		toponyme	0	oui
pratique enseignante	2	*enseignant*	0	oui
		pratique professionnelle	1 + implicite	non
écriture scolaire	2	*milieu scolaire*	1 + implicite	non
		écriture	0	oui
enfant sourd	2	*enfant*	0	oui
		surdité	0	oui

- **Le mot-clé INIST correspond à plusieurs mots-clés Termith**

MC Termith	score P	MC INIST	score S	lien vers MC termith
genre	1	*genre du discours*	0	genre
discours	2			
genre	1	*genre grammatical*	0	genre
grammatical	1			
quantificateur	1	*quantificateur existentiel*	0	quantificateur
existentiel	2			

- **Le mot-clé INIST ne correspond à aucun mot-clé Termith et n'a pas d'occurrence dans le texte**

Le mot-clé est plus ou moins important et provient d'informations déduites par l'indexeur à partir d'éléments présents dans le texte ; le texte pouvant être soit le résumé, soit le texte intégral que l'on ne voit pas dans certaines phases d'évaluation (phases 1 et 3).
On lui attribuera un score 1 avec commentaire "implicite".

Le mot-clé est plus ou moins important, provient d'informations implicites et est en plus très générique. Il y a donc de fortes chances pour qu'il ait été généré automatiquement par le système d'indexation de l'INIST qui a pour règle de générer des mots-clés génériques à partir de mots-clés proposés par l'indexeur.
On lui attribuera un score 1 avec commentaire "générique".

Le mot-clé est important, mais est présent dans le texte sous une forme plus spécifique.
On lui attribuera un score 2 avec commentaire "forme variante dans le texte : xxx".

MC INIST	**contexte**	**score**	**commentaire**
anthroponyme	nom des personnages dans des textes littéraires	1	implicite
ancien français	français du 12e et du 13e siècles	1	implicite
linguistique appliquée	traitement automatique des langues	1	générique
sociolinguistique	politique linguistique	1	générique
lexicométrie	approche lexicométrique	2	forme variante dans le texte : approche lexicométrique

- **Le mot-clé INIST correspond à un mot-clé Termith de manière plus ou moins proche**

La correspondance entre un mot-clé INIST et un mot-clé Termith peut être faite pour valider tout type de lien entre les deux, pas uniquement une correspondance exacte.
Le score de silence est toujours 0 quand on fait une correspondance.

MC Termith	**MC INIST**
logiciels	*logiciel*

texte délibératif	*délibération*
échange exolingue	*communication exolingue*
étudiants	*milieu universitaire*
corpus	*analyse de corpus*