

Yahoo travaille pour nous

Grâce à RDFa, la technologie la plus abordable du web sémantique, faites de Yahoo votre interface de recherche simple, votre affichage court (voire détaillé) et votre web service. Bientôt disponible en version Google.

On connaît la fable du catalogue et du moteur de recherche. Sa morale ne joue pas en faveur des catalogues de bibliothèque, réputés périmés depuis l'avènement du moteur de recherche comme voie royale.

Mais l'Histoire est une saga, pas une fable. Certains catalogues ont réagi à cette concurrence, d'une part en faisant en sorte d'être efficacement indexés par les moteurs de recherche¹, d'autre part en concevant des interfaces web plus riches, plus réactives et surtout plus faciles à utiliser. Ces OPAC de nouvelle génération ont compris que la survie des catalogues ne résidait pas dans l'imitation servile du sobre Google, mais dans la mise en valeur de leur potentiel, à savoir les métadonnées finement structurées par les catalogueurs. La saga rebondit, non sans ironie. Depuis quelques années, les moteurs font eux-mêmes l'effort de mieux exploiter les métadonnées qui tombent dans l'escarcelle de leurs robots-crawlers. Si l'on ignore toujours comment les algorithmes de calcul de la pertinence prennent en compte ces métadonnées, on comprend que celles-ci jouent un rôle de plus en plus important dans la présentation et l'organisation des résultats. On peut ignorer cette tendance, l'observer passivement ou encore essayer d'en tirer profit. Cet article suggère quelques opportunités à saisir face aux dernières nouveautés de Yahoo.

Yahoo SearchMonkey, vitrine pour nos métadonnées

En 2008, Yahoo a lancé un nouveau service nommé SearchMonkey.

Ce service donne au gestionnaire d'un site les moyens de maîtriser la manière dont s'affichent les résultats d'une recherche dans Yahoo.

Au lieu de cela :

[Calames : MS 36 Job Coat. "Bue Jeanne Thibaut anvet Jeanne Darc ...](#)
Catalogue en ligne des archives et des manuscrits de l'enseignement supérieur.
... 2, acte de vente, sur papier timbré, des manuscrits 36, 37 et 38, signé de ...
[www.calames.abes.fr/pub/ms/UN160047 - En cache](#)

voici ce que le lecteur peut trouver au terme de sa recherche dans Yahoo :

[Calames : MS 36 Job Coat. "Bue Jeanne Thibaut anvet Jeanne Darc ...](#)
[Catalogue Calames](#)
• Collection: Manuscrits de la bibliothèque de l'Université de Rennes
• Date: 1853
• Auteur: Coat (Job)
• Format: Cartonné
[www.calames.abes.fr/pub/ms/UN160047 - En cache](#)

Aujourd'hui, un tel affichage est possible pour **Calames, le catalogue des archives et des manuscrits de l'enseignement supérieur développé par l'ABES²**.

Demain, imaginons un affichage analogue pour le futur portail des thèses :

[Thèse - Nouvelles méthodologies pour l'accès aux ions sulfénates et dérivés](#)

[Résumé](#) [Sommaire](#) [Lire en ligne](#) [Emprunter en bibliothèque](#)

• Auteur : Florian Foucoin

• Directeur : Stéphane Perrio

• Université : Université de Caen

• Année : 2008

TH <http://www.theses.fr/2008CAEN2039> - En cache

À la place d'un affichage par défaut souvent confus et minimaliste, SearchMonkey nous donne les moyens d'élaborer nous-mêmes l'affichage de nos sites, en exploitant les métadonnées structurées que nous avons nous-mêmes incrustées dans nos pages web. On peut ainsi sélectionner les informations à mettre en avant ou proposer des hyperliens qui correspondent à des actions (accéder au document en ligne, le commander par prêt entre bibliothèques, laisser un commentaire...).

Quelle est la marche à suivre ?

- 1- Nous avons enrichi les pages HTML de Calames en y ajoutant des informations structurées, basées essentiellement sur le Dublin Core. Ces informations sont cachées : elles ne changent rien à l'affichage des pages dans le navigateur. (voir l'encart sur RDFa).
- 2- Yahoo (comme Google ou d'autres moteurs) moissonne ces pages HTML, mais désormais il sait en extraire ces informations structurées pour les enregistrer dans sa base de données.
- 3- Yahoo permet à n'importe qui d'écrire des programmes qui exploitent ces informations structurées pour modifier l'apparence des résultats de recherche. Ces programmes sont hébergés et exécutés sur les serveurs de Yahoo.
- 4- L'utilisateur final doit s'authentifier avec son compte Yahoo et, la première fois, activer une fonction qui, au moment d'une recherche qui renvoie des résultats Calames, lance le programme d'affichage spécifique.

Activer le nouvel affichage
des résultats Calames sur cette page :

<http://gallery.search.yahoo.com/application?smid=H50.s>



On le devine, l'étape la plus hasardeuse est la dernière : ne rêvons pas, peu de personnes prendront la peine de privilégier Yahoo aux dépens de Google, de s'authentifier puis d'activer une obscure fonction.

SearchMonkey prend tout son sens quand il s'applique par défaut sur certains sites, comme c'est le cas depuis peu pour Facebook :

Alex Moskalyuk on Facebook

[Add friend](#) | [Poke](#) | [Send message](#) | [View friends](#)

Alex Moskalyuk is on Facebook. Not the Alex Moskalyuk you were looking for? Visit Facebook to search for friends, family, coworkers.

www.facebook.com/people/Alex-Moskalyuk/500013218 - [Cached](#)



L'utilisateur profite alors des avantages de SearchMonkey sans y penser. Nous espérons qu'un jour Yahoo permettra à n'importe quel éditeur de site de faire de son affichage SearchMonkey l'affichage par défaut pour tous les utilisateurs. Heureusement, SearchMonkey n'est pas le seul débouché pour les informations structurées que Yahoo extrait des pages web. D'ailleurs, on pourrait imaginer qu'un moteur comme Yahoo utilise ces métadonnées pour renvoyer des résultats plus pertinents, et pas seulement plus présentables. Mais beaucoup d'aigrefins du web ne sauraient résister à la tentation de manipuler les algorithmes de pertinence.

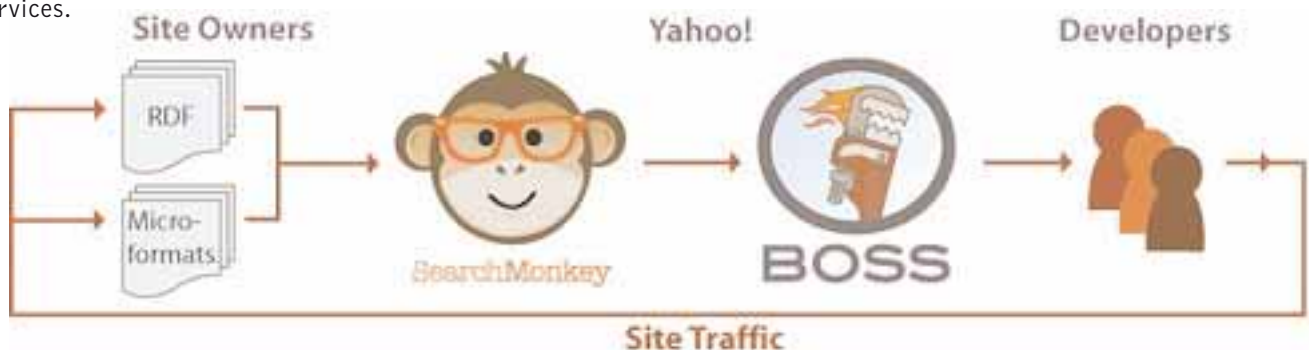
Yahoo BOSS, des web services pour nos métadonnées

Depuis quelques semaines, Yahoo a annoncé que les métadonnées RDF moissonnées par Yahoo seraient ouvertes, réutilisables par n'importe qui pour n'importe quel usage.³ C'est une révolution.

En effet, en lançant son nouveau service BOSS (*Build your Own Search System*), Yahoo propose à n'importe qui d'exploiter sa base de données pour construire de nouvelles applications de recherche (moteurs de recherche spécialisés dans un domaine, fusion des réponses fournies par différents moteurs...).⁴



Or, désormais, dans les données de la base Yahoo, on trouve aussi les métadonnées RDF extraites par SearchMonkey. Yahoo est donc aujourd'hui une base de métadonnées Dublin Core, FOAF⁵, vCard⁶, DBPedia⁷, Freebase⁸... accessibles par web services.



Et parmi ces métadonnées hébergées et exposées par Yahoo, il y a les métadonnées de Calames. Yahoo devient une API de Calames, son interface de programmation⁹. Cela signifie que n'importe qui peut construire une nouvelle interface de recherche sur les notices de Calames. Plus intéressant, Calames pourrait enrichir sa propre interface de consultation en allant puiser chez Yahoo des informations biographiques ou bibliographiques sur un scriptorium ou sur un écrivain. Plus prometteur encore, n'importe qui pourrait construire de nouvelles interfaces qui agrégeraient via Yahoo des métadonnées de manuscrits dispersés entre différentes institutions : reconstruire virtuellement la bibliothèque de l'abbaye de Clairvaux ou, pourquoi pas, un portail des manuscrits médiévaux... En d'autres termes, composer des mashups patrimoniaux.¹⁰

Le plus difficile n'est pas tant le développement des interfaces que l'accès aux métadonnées structurées. RDFa est le moyen le plus efficace pour exposer ses métadonnées à la face du web. L'effort demandé est modeste car il s'agit seulement d'ajouter quelques attributs au code HTML de son interface utilisateur.

L'effet est démultiplié à partir du moment où les grossistes du web que sont les moteurs de recherche comprennent vos données et les rendent exploitables par d'autres – et pas seulement par des outils propres aux bibliothèques comme le protocole z39.50.

Google s'y met

Pour ceux qui douteraient encore, le coup de grâce viendra de Google qui vient d'annoncer le 12 mai 2009 qu'il allait également tirer parti des métadonnées RDFa encodées dans les pages HTML, notamment pour enrichir la liste des résultats, comme le fait Yahoo.

Exemple Google

Drooling Dog Bar B.Q. - Colfax, CA

★★★★★ 15 reviews - Price range: \$\$

Drooling Dog has some really good BBQ. I had the pulled pork sandwich.

.... Drooling Dog BBQ is a great place to stop at on your way up the hill to Tahoe ...

www.yelp.com/biz/drooling-dog-bar-b-q-colfax - 75k - [Cached](#) - [Similar pages](#)

Mais, dans un premier temps, Google ne comprendra que ce qu'il veut bien comprendre (et associer à des liens sponsorisés). En effet, au lieu de supporter des vocabulaires RDF standard comme Dublin Core ou FOAF, Google oblige les propriétaires de sites à utiliser des vocabulaires *ad hoc*, forgés par Google pour décrire des produits, leurs prix et leur évaluation par les consommateurs.¹¹ Le web sémantique patrimonial ou scientifique attendra encore un peu.

Morale édifiante, irénique

et corporatiste :

le web sémantique est un catalogue

Ces annonces de Yahoo puis de Google signalent l'investissement des grands moteurs de recherche dans ce qu'on appelle maladroitement « web sémantique ». On préfère aujourd'hui utiliser l'expression « web des données », car elle exprime de manière plus exacte et moins pompeuse ce qui est en train de se passer. Historiquement et fonctionnellement, aujourd'hui encore, le web est une base de documents, et les moteurs nous aident à trouver ces documents (pages HTML, fichiers PDF, vidéos...). Progressivement, le web va devenir également une base de... données¹²,

RDFa : des métadonnées cachées dans les pages web

Le code HTML d'une page web a pour fonction première d'être correctement interprété par un navigateur qui va rendre l'information présentable et lisible par un humain.

Ainsi, la page

<http://www.calames.abes.fr/pub/ms/UNI160047> de Calames contient ce code

```
<td> Auteur : </td>
```

```
<td><span>Coat (Job)</span></td>
```

qui permet d'afficher cette zone :

Auteur : Coat (Job)

Mais seul un humain peut comprendre immédiatement que Coat (Job) est l'auteur du manuscrit décrit.

Pour le rendre compréhensible par des programmes, il faut expliciter cette information de la manière suivante :

```
<td> Auteur : </td>
```

```
<td><span property="dc:creator"
```

```
about="http://www.calames.abes.fr/pub/ms/res/UNI160047">
```

```
Coat (Job)</span></td>
```

On a simplement ajouté deux attributs (*property* et *about*) pour exprimer le **fait** qu'une personne nommée « Coat (Job) » est l'auteur (au sens du Dublin Core) du manuscrit identifié par le code

« <http://www.calames.abes.fr/pub/ms/res/UNI160047> »

(subtilité : il ne faut pas confondre ce code identifiant avec l'URL de la notice).

Ce fait peut aussi s'exprimer ainsi, sous une forme brute :

```
<http://www.calames.abes.fr/pub/ms/res/UNI160047>
```

```
<http://purl.org/dc/elements/1.1/creator> "Coat (Job)"
```

Cet ensemble de trois termes constitue une assertion élémentaire, une petite vérité, un « triplet RDF ».

Le web sémantique est la collection des triplets RDF accessibles sur le web.

RDFa (*Resource Description Framework* in attributes) est une méthode pour exprimer ces triplets dans les attributs au sein des éléments HTML.

et les moteurs de recherche vont nous aider à trouver de l'information structurée, des faits.

Nos catalogues ne sont pas une collection de documents qu'on appellerait « fiches » ou « notices ». Ils contiennent des données finement structurées, qui décrivent des documents (monographies, périodiques, articles, thèses, manuscrits), mais également des personnes, des collectivités, des concepts. Il ne suffit pas d'exposer aux moteurs de recherche nos notices bibliographiques et nos notices d'autorité ; il faut leur faire **comprendre** les données qui se trouvent dans ces documents-notices. RDFa est la voie la plus rapide dans cette direction.

Yann Nicolas

 nicolas@abes.fr

1 Sitemaps

est un moyen efficace d'indiquer aux moteurs les pages à indexer. Il faut avouer qu'à ce jeu, Google est bien plus fiable que Yahoo.

<http://www.sitemaps.org/>

2 www.calames.abes.fr

3 À ceci près que Yahoo

vient d'annoncer un système

de facturation, applicable à partir d'un certain nombre de requêtes par jour :

<http://developer.yahoo.com/search/boss/fees.html>

4 Une liste d'applications

qui s'appuient sur le service de recherche BOSS :

<http://www.programmableweb.com/api/yahoo-boss/mashups>

5 Friend-Of-A-Friend

est un vocabulaire qui peut modéliser les personnes et les relations interpersonnelles.

<http://www.foaf-project.org/>

6 vCard

est un vocabulaire qui permet de représenter une carte de visite électronique.

7 Dbpedia

est une version RDF de Wikipedia.

<http://dbpedia.org/About>

8 Freebase

est une encyclopédie collaborative entièrement constituée de données structurées compatibles avec les technologies du web sémantique.

<http://www.freebase.com/>

9 API

= *Application Programming Interface*

= interface de programmation.

10 NICOLAS, Yann, « Calames et après », in Bulletin des bibliothèques de France,

T. 53, n° 6, 2008, p. 29-33.

<http://bbf.enssib.fr/consulter/bbf-2008-06-0029-005>

11 <http://rdf.data-vocabulary.org>

12 Dans cette logique, Yahoo

propose un langage de requête qui permet d'interroger Yahoo et d'autres sites à la manière dont SQL permet d'interroger une base de données relationnelle.

Il s'agit de YQL (*Yahoo! query language*). Certes, YQL est encore limité et ne permet pas de tirer parti de toutes les possibilités des données RDF extraites par SearchMonkey.

<http://developer.yahoo.com/yql/>