Métadonnée

Cours « Document et Web Sémantique »

Nicolas Delestre

INSA ROUEN NORMANDIE

Plan. . .

INSA ROUEN NORMANDIE

Quelques constats 1 / 2

Google's Algorithm Is Lying to You About Onions

`http://daringfireball.net/linked/2017/03/07/google-onions`

> « *Not only does Google, the world's preeminent index of information, tell its users that caramelizing onions takes "about 5 minutes" - it pulls that information from an article whose entire point was to tell people exactly the opposite. A block of text from the Times that I had published as a quote, to illustrate how it was a lie, had been extracted by the algorithm as the authoritative truth on the subject.* » *(Tom Scocca)*

INSA ROUEN NORMANDIE

Quelques constats 2 / 2

Pourquoi ?

- Les moteurs de recherche indexent des documents à destination des humains et non des machines
- Non prise en compte du contexte (à l'indexation et au moment de la recherche)
- Les moteurs de recherche utilisent les données (mots) pas l'information (relation entre les mots, les concepts)

- Les moteurs de recherche devraient demander des précisions concernant la recherche (interaction entre l'utilisateur et le moteur de recherche)

INSA ROUEN NORMANDIE

Rappels 1 / 2

Chaîne d'interprétation d'un document

1. Segmentation
   - Obtention des tokens
2. Analyse lexicale
   - Regroupement/séparation de token en lexème avec proposition d'hypothèses quant à leur rôle
3. Analyse syntaxique
   - Construction d'un arbre syntaxique qui va pouvoir vérifier que la phrase est bien construite et lever quelques hypothèses
4. Analyse sémantique
   - Interprétation sémantique afin de lever les dernières hypothèses
5. Analyse pragmatique
   - Contextualisation

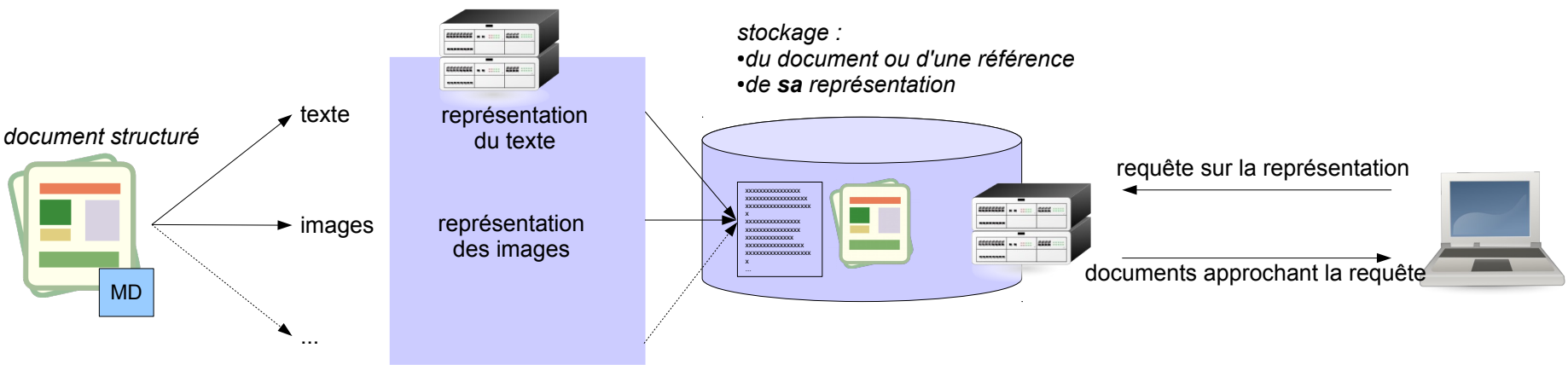
INSA ROUEN NORMANDIE

Rappels 2 / 2

Constats

- Aucun système (générique) aujourd'hui ne sait faire convenablement toutes ces étapes
- Les moteurs de recherche ne fond que la première ou deuxième étape
- Si on veut donner du sens à un document il faut lui adjoindre de l'information

Rôle des métadonnées



Métadonnées, Schéma, Profil d'applications 1 / 3

Définitions

≪ Une métadonnée (du grec meta "après" et du latin data "informations") est une donnée servant à définir ou décrire une autre donnée quel que soit son support (papier ou électronique). ≫ (Wikipédia)

≪ Les métadonnées sont, dans le cadre du Web sémantique, des données signifiantes qui permettent de faciliter l'accès au contenu informationnel d'une ressource informatique, une notice de contenu intégrée en quelque sorte (dans l'en-tête des documents HTML côté code source ou en tant que fichier XML autonome par exemple). ≫ (Wikipédia)

INSA ROUEN NORMANDIE

Métadonnées, Schéma, Profil d'applications 2 / 3

Métadonnée (en pratique)

- C'est une donnée sur une donnée :
  - clairement définie
  - avec une arité (attention au LangString)
  - a un type :
    - type énuméré (vocabulaire fermé)
    - type prédéfini (entier, date, etc.) avec formalisme de représentation

Schéma de métadonnées

- Ensemble de métadonnées défini par un organisme, une entreprise, etc.
- Exemples de schéma de métadonnées
  - Dublin Core
  - Learning Object Metadata

INSA ROUEN NORMANDIE

Métadonnées, Schéma, Profil d'applications 3 / 3

Profil d'application

- Définition

  « *...métadonnées issues d'un ou plusieurs schémas de métadonnées combinés afin d'améliorer et d'optimiser leur utilisation dans un cadre particulier* » *[HP00]*

  « *... est d'adapter et de combiner des schémas existants afin d'obtenir un nouveau schéma pour une application particulière tout en gardant l'intéropérabilité avec le ou les schémas de base* » *[DHSW02]*

- Un profil d'application est un schéma de métadonnées
- Exemples de profil d'application :
  - CanCore
  - LOM-FR

Le Dublin Core 1 / 2

Qu'est ce ?

- Géré par le (*Dublin Core Metadata Initiative*) : `http://dublincore.org/index.shtml`
- Schéma de métadonnées permettant de décrire des documents (numériques ou physiques)
  - 15 métadonnées forment le DCMES (*Dublin Core Metadata Element Set*). C'est aujourd'hui une norme (ISO 15836)
  - 55 (avec les 15 précédents) forment le *DCMI Metadata Terms*

Les métadonnées du DCMES (Wikipédia)

Elément	Commentaire
Title	Titre principal du document
Creator	Nom de la personne, de l'organisation ou du service à l'origine de la rédaction du document
Subject	Mots-clefs, phrases de résumé, ou codes de classement
Description	Résumé, table des matières, ou texte libre
Publisher	Nom de la personne, de l'organisation ou du service à l'origine de la publication du document

Le Dublin Core 2 / 2

Les métadonnées (suite)

Elément	Commentaire
Contributor	Nom d'une personne, d'une organisation ou d'un service qui contribue ou a contribué à l'élaboration du document. Chaque contributeur fait l'objet d'un élément Contributor séparé
Date	Date d'un évènement dans le cycle de vie du document
Type	Genre du contenu
Format	Type MIME, ou format physique du document
Identifier	Identificateur non ambigu : il est recommandé d'utiliser un système de référencement précis, afin que l'identifiant soit unique au sein du site, par exemple les URI ou les numéros ISBN
Source	Ressource dont dérive le document : le document peut découler en totalité ou en partie de la ressource en question. Il est recommandé d'utiliser une dénomination formelle des ressources, par exemple leur URI
Language	La langue du document
Relation	Lien avec d'autres ressources. De nombreux raffinements permettent d'établir des liens précis, par exemple de version, de chapitres, de standard, etc.
Coverage	Couverture spatiale (point géographique, pays, régions, noms de lieux) ou temporelle
Rights	Droits de propriété intellectuelle, Copyright, droits de propriété divers

Schema.org 1 / 3

Qu'est-ce ?

- `http://schema.org`
- Proposé par Google, Yahoo, Bing et Yandex
- Schéma de métadonnées généraliste (version 3.3 du 14/08/2017)
  - 597 classes (organisées en hiérarchie - héritage -)
  - 867 propriétés
  - 114 valeurs énumérées

Premier niveau de la hiérarchie

```
Thing
├── Action
├── CreativeWork
├── Event
├── Intangible
├── MedicalEntity
├── Organization
├── Person
├── Place
└── Product
```

Schema.org 2 / 3

Un extrait de la hiérarchie de classes

```
CreativeWork
├── Article
│   ├── NewsArticle
│   ├── Report
│   ├── ScholarlyArticle
│   │   └── MedicalScholarlyArticle
│   ├── SocialMediaPosting
│   │   └── ...
│   └── TechArticle
│       └── APIReference
├── Blog
├── Book
└── ...
```

Schema.org 3 / 3

Propriétés

- Les propriétés sont attachées aux classes :

  Thing additionalType, alternateName, description, image, name, potentialAction, sameAs, url

  Person additionalName, address, affiliation, alumniOf, award, ...

- Les valeurs des propriétés peuvent de type simples (Datatype) ou de classes :
  - Datatype : Boolean, Date, DateTime, Number (deux sous types : Float et Integer), Text (un sous type URL), Time
  - ex : addtionalName : Text
  - ex : address : PostalAddress
- Une sous classes peut utiliser les propriétés des super classes

INSA ROUEN NORMANDIE

Métadonnées pour les pages (HTML5)

- Métadonnée par défaut : balise *title*
- Métadonnées additionnelles, utilisation de la balise *meta* avec les attributs :

  name pour le nom de la métadonnée : *author*, *description*, *generator*, *keywords*

  content pour la valeur

Exemple : http://www.w3.org/(2015)

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/
    DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
<!-- Generated from data/head-home.php, ../../smarty/{head.tpl} -->
<head>
<title>World Wide Web Consortium (W3C)</title>
...
<meta name="description" content="The World Wide Web Consortium (W3C) is an
    international community where Member organizations, a full-time staff, and the
    public work together to develop Web standards." />
...
</head>
```

Microformat

Microformat (http://microformats.org)

- Spécifie une syntaxe et des schémas à utiliser
- Utilisation de l'attribut `class` pour spécifier la classe et le rôle de d'une valeur
- Plusieurs classes sont proposées (`hCard`, `hCalendar`, `rel-licence`, `hRecipe`, etc.). Par exemple `marmiton.org` utilise hRecipe.
- La version 2 propose une nomenclature de nommage (`h-*` pour les noms de classes, `p-*` pour les propriétés simples, etc.)

Exemple (http://microformats.org/wiki/microformats2-fr)

```
<h1 class="h-card">
  <span class="p-given-name">Chris</span>
  <abbr class="p-additional-name">R.</abbr>
  <span class="p-family-name">Messina</span>
</h1>
```

RDFa

- Mécanisme proposé par le w3c
- Définit la syntaxe, ne spécifie pas les schémas à utiliser
- Deux versions : RDFa v1.0 pour XHTML 1.0 et HTML 4.01 et **RDFa v1.1 pour HTML5** (deux versions HTML+RDFaLite et HTML+RDFa)
- Propriétés : vocab, typeof, property, etc.

Exemple HTML+RDFaLite (http://www.w3.org/TR/html-rdfa/)

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <title>Example Document</title>
  </head>
  <body vocab="http://schema.org/">
    <p typeof="Blog">
      Welcome to my <a property="url" href="http://example.org/">blog</a>.
    </p>
  </body>
</html>
```

Micro données (*microdata*)

- Syntaxe principalement associée à `Schema.org`
- Compatible avec HTML5, trois attributs principaux :
  - `itemscope` : Crée un élément et indique que les descendants de cette balise HTML contiennent des informations à son sujet.
  - `itemtype` : Un URL pointant vers un vocabulaire qui décrit l'élément et ses propriétés (`http://schema.org/XX`)
  - `itemprop` : Indique que la balise contient la valeur de la propriété indiquée.

Exemple (`http://schema.org/docs/gs.html`)

```
<div itemscope itemtype="http://schema.org/Movie">
  <h1 itemprop="name"&g;Avatar</h1>
  <div itemprop="director" itemscope itemtype="http://schema.org/Person">
    Director:
      <span itemprop="name">James Cameron</span>
      (born <span itemprop="birthDate">August 16, 1954)</span>
  </div>
  <span itemprop="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html" itemprop="trailer">Trailer</a>
</div>
```

JSON-LD

- Notation JSON avec des clés réservées (commençant par un @ pour le Linked Data)
- Format recommandé par Google

Exemple (http://schema.org/Movie)

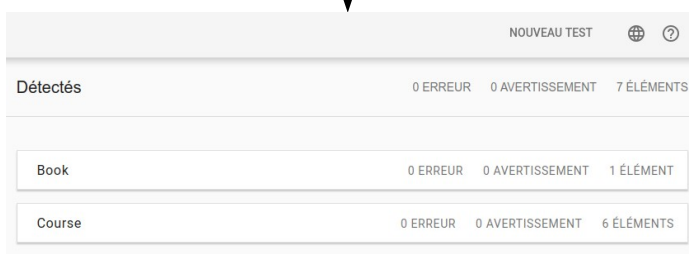
```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Movie",
  "name": "Pirates of the Carribean: On Stranger Tides (2011)",
  "actor": [{"@type": "Person", "name": "Johnny Depp"}, {"@type": "Person", "name": "
       Penelope Cruz"}],
  "aggregateRating": {
    "@type": "AggregateRating", "bestRating": "10", "ratingCount":
"200", "ratingValue": "8", "reviewCount": "50"
  },
  "author": [{"@type": "Person", "name": "Ted Elliott"}, {"@type": "Person", "name": "
       Terry Rossio"}],
  "description": "Jack Sparrow and Barbossa embark on a quest to find the elusive
       fountain of youth, only to discover that Blackbeard and his daughter are after
       it too.",
  "director": {"@type": "Person", "name": "Rob Marshall"}
  }
  </script>
```

Un exemple : http://asi.insa-rouen.fr/~delestre

```
<!doctype html>
<html lang="fr">
  <head>
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
….
   <script type="application/ld+json">
{
    "@context": {
          "sch": "http://schema.org/"
    },
    "@graph": [
          {
              "@id": "#me",
              "@type": "sch:Person",
              "sch:givenName" : "Nicolas",
              "sch:familyName" : "Delestre",
…
   </script>
  </head>
  <body >
    <div id="w">
      <header class="clearfix" >
          <div id="info">
            <h1>Nicolas Delestre</h1>
```

pour les machines

pour les humains



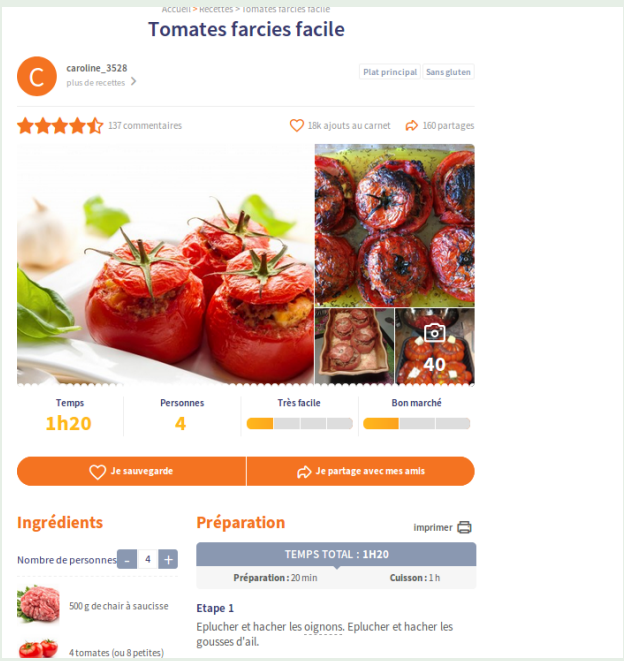
https://validator.schema.org/

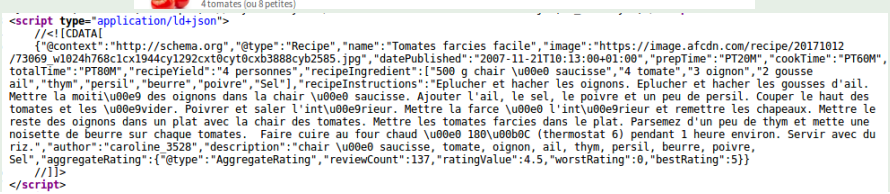
Rich Snippet 1 / 2

Qu'est ce ?

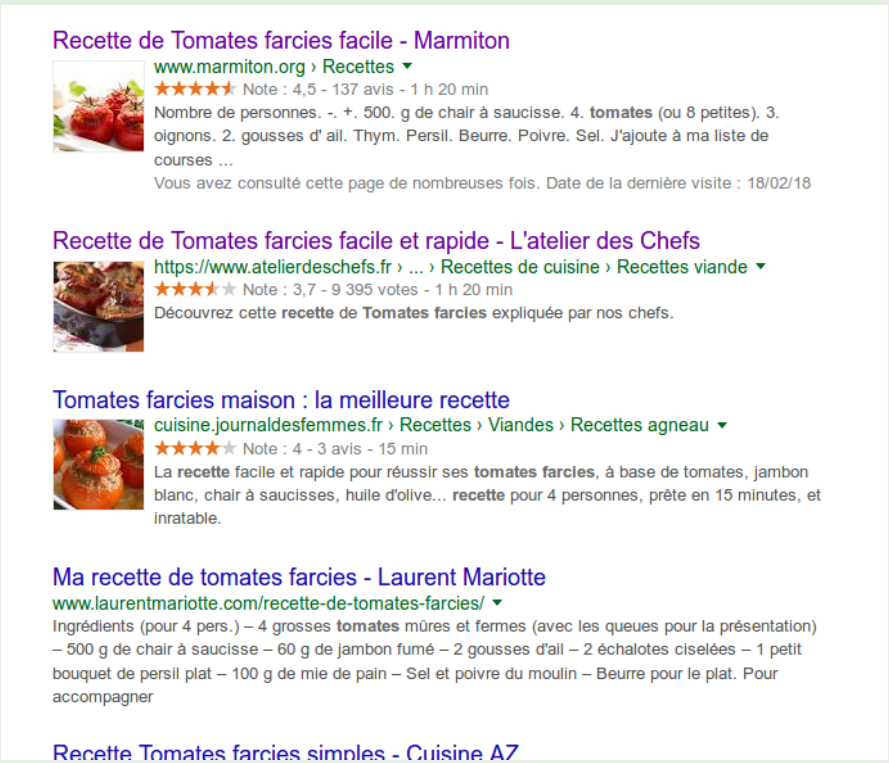
- Informations ajoutées dans la page des résultats
- Peut être des images ou du texte

Utilisation des métadonnées de www.marmiton.org par Google





```
<script type="application/ld+json">
    //<![CDATA[
    {"@context":"http://schema.org","@type":"Recipe","name":"Tomates farcies facile","image":"https://image.afcdn.com/recipe/20171012
/73069_w1024h768c1cx1944cy1292cxt0cyt0cxb3888cyb2585.jpg","datePublished":"2007-11-21T10:13:00+01:00","prepTime":"PT20M","cookTime":"PT60M",
totalTime":"PT80M","recipeYield":"4 personnes","recipeIngredient":["500 g chair \u00e0 saucisse","4 tomate","3 oignon","2 gousse
ail","thym","persil","beurre","poivre","Sel"],"recipeInstructions":"Eplucher et hacher les oignons. Eplucher et hacher les gousses d'ail.
Mettre la moiti\u00e9 des oignons dans la chair \u00e0 saucisse. Ajouter l'ail, le sel, le poivre et un peu de persil. Couper le haut des
tomates et les \u00e9vider. Poivrer et saler l'int\u00e9rieur. Mettre la farce \u00e0 l'int\u00e9rieur et remettre les chapeaux. Mettre le
reste des oignons dans un plat avec la chair des tomates. Mettre les tomates farcies dans le plat. Parsemez d'un peu de thym et mette une
noisette de beurre sur chaque tomates.  Faire cuire au four chaud \u00e0 180\u00b0C (thermostat 6) pendant 1 heure environ. Servir avec du
riz.","author":"caroline_3528","description":"chair \u00e0 saucisse, tomate, oignon, ail, thym, persil, beurre, poivre,
Sel","aggregateRating":{"@type":"AggregateRating","reviewCount":137,"ratingValue":4.5,"worstRating":0,"bestRating":5}}
    //]]>
</script>
```



Rich Snippet 2 / 2

Classes prises en compte par Google

- Product
- Recipe
- Review
- Event
- SoftwareApplication
- Video
- NewsArticle

OAI-PMH 1 / 2

Constat

- De nombreux formats ne permettent pas d'intégrer des métadonnées (PDF, images, vidéos, etc.) autres que celles prévues par défaut

Solutions

- Utilisation d'entrepôts de métadonnées contenant des fiches de description de documents
- L'*Open Archive Initiative* propose le *Protocol for Metadata Harvesting* qui permet aux entrepôts de s'échanger des fiches de métadonnées (v1.0 en 2001 et v2.0 en 2002) :
  - requête sous forme d'URL
  - échange d'information au format XML
  - indépendant des schémas mais incluant au moins les métadonnées du Dublin Core (par exemple BiblioML, LOM, etc.)
- Exemples d'entrepôts (français) :
  - UNIT : `http://www.unit.eu/ori-oai-repository/OAIHandler?`
  - HAL : `http://api.archives-ouvertes.fr/oai/hal/?`

OAI-PMH 2 / 2

Requêtes (Wikipédia)

Requête	Argument	Rôle
GetRecord	identifier (R), metadataPrefix (R)	Récupération d'un enregistrement donné
Identify		Informations sur l'entrepôt de données
ListIdentifiers	from (O), until (O), metadataPrefix (R), set (O), resumptionToken (E)	Récupère la liste des identifiants disponibles
ListMetadataFormats	identifier (O)	Demande la liste des formats de métadonnées disponibles. Sans paramètres tous les formats disponibles pour au moins un item sont retournés. Avec le paramètre identifier, ne sont retournés que les formats disponibles pour l'item concerné
ListRecords	from (O), until (O), metadataPrefix (R), set (O), resumptionToken (E)	Retourne une liste d'enregistrements correspondant aux différents paramètres (dates, ensemble) demandés
ListSets	resumptionToken (E)	Demande la liste des ensembles disponibles sur un entrepôt. La réponse peut être sur plusieurs pages

(R) Required, (O) Optional, (E) Exclusive

OAI-PMH, exemples de requête (UNIT) 1 / 3

Liste des schémas

- Requête : `http://www.unit.eu/ori-oai-repository/OAIHandler?verb=ListMetadataFormats`
- Réponse :

```
<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2015-03-03T14:14:55Z</responseDate>
  <request
  verb="ListMetadataFormats">http://www.unit.eu/ori-oai-repository/OAIHandler</
       request>
  <ListMetadataFormats>
    <metadataFormat>
      <metadataPrefix>oai_dc</metadataPrefix>
      <schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd</schema>
      <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</
           metadataNamespace>
    </metadataFormat>
    <metadataFormat>
      <metadataPrefix>lom</metadataPrefix>
      <schema>http://ltsc.ieee.org/xsd/lomv1.0/lom.xsd</schema>
      <metadataNamespace>http://ltsc.ieee.org/xsd/LOM</metadataNamespace>
    </metadataFormat>
  </ListMetadataFormats>
</OAI-PMH>
```

OAI-PMH, exemples de requête (UNIT) 2 / 3

Liste des fiches d'un certain schéma

- Requête : `http://www.unit.eu/ori-oai-repository/OAIHandler?verb=ListRecords&metadataPrefix=lom`
- Réponse :

```
<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.
     openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2015-03-03T14:24:00Z</responseDate>
  <request metadataPrefix="lom" verb="ListRecords">http://www.unit.eu/ori-oai-
       repository/OAIHandler</request>
  <ListRecords>
    <record>
      <header>
        <identifier>oai:oriwww.unit.eu:unit-ori-wf-1-5039</identifier>
        <datestamp>2011-10-26T00:00:00Z</datestamp>
        <setSpec>search_unit_taxonomie_regexp:3</setSpec>
        ...
      </header>
      <metadata>
        <lom:lom>
          <lom:general>
            <lom:identifier>
              <lom:catalog>URI</lom:catalog>
              <lom:entry>http://ori.unit-c.fr/uid/unit-ori-wf-1-5039</lom:entry>
            </lom:identifier>
            ...
```

OAI-PMH, exemples de requête (UNIT) 3 / 3

Métadonnées d'une fiche pour un schéma donné

- Requête : `http://www.unit.eu/ori-oai-repository/OAIHandler?verb=GetRecord&identifier=oai:oriwww.unit.eu:unit-ori-wf-1-5039&metadataPrefix=lom`
- Réponse :

```
<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.
     openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2015-03-03T14:34:46Z</responseDate>
  <request identifier="oai:oriwww.unit.eu:unit-ori-wf-1-5039" metadataPrefix="lom"
       verb="GetRecord">http://www.unit.eu/ori-oai-repository/OAIHandler</
      request>
  <GetRecord>
    <record>
      <header>
        ...
      </header>
      <metadata>
        <lom:lom>
          <lom:general>
            <lom:identifier>
              <lom:catalog>URI</lom:catalog>
              <lom:entry>http://ori.unit-c.fr/uid/unit-ori-wf-1-5039</lom:entry>
            </lom:identifier>
            <lom:title>
              <lom:string language="fre">Vibrations des Structures</lom:string>
            </lom:title>
```

ORI-OAI

Source : `http://www.ori-oai.org/`

Conclusion

- Les métadonnées permettent à des programmes de récupérer et d'interpréter des informations structurées

- C'est grâce à ces métadonnées que les moteurs de recherche peuvent ajouter des *Rich Snippets*
- Mais :
  - ces métadonnées ne sont pas vraiment liées
  - mélanger métadonnées et document peut ralentir les processus de chargement, d'extraction et d'interprétation des documents (le navigateur charge les métadonnées contenues dans les pages alors qu'il ne va pas en tirer parti...)
  - les représentations syntaxiques sont diverses
  - il n'y a pas de langage permettant d'interroger ces métadonnées (à l'image du SQL pour le monde relationnel)

INSA ROUEN NORMANDIE

Références

[DHSW02] E. Duval, W. Hodgins, S. Sutton, and S.L. Weibel.
Metadata principles and practicalities.
D-Lib Magazine, http ://www.dlib.org/dlib/april02/weibel/ 04weibel.html, 2002.

[HP00] R. Heery and M. Patel.
Application profiles : mixing and matching metadata schemas.
Ariadne issue 25, 2000.

INSA ROUEN NORMANDIE