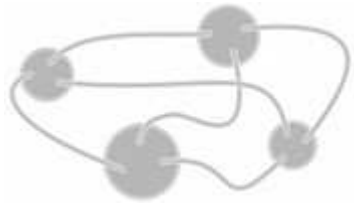
# Participer au web de données avec les données de la recherche en SHS : comment utiliser RDFa ?

Stéphane POUYLLAU, ingénieur de recherche au Centre national de la recherche scientifique.

Ce document a une vocation pédagogique : la formation par l'exemple à la mise en place d'une structuration RDF dans une page web contenant des informations pour la recherche en SHS.

Il a fait l'objet d'un billet de blog dans premier temps (http://blog.stephanepouyllau.org/401), puis d'une présentation orale.

Contexte : mise en place d'un connecteur/outil de collecte de l'information (métadonnées, données) dans le plate-forme ISIDORE utilisant les principes du RDFa 1.0 et la technologie XML « Sitemap ».

CORPUS IR


# Participer au web de données avec les données de la recherche en SHS : comment utiliser RDFa ?

Le web est l'un des vecteurs principaux de la diffusion des données de recherche en sciences humaines et sociales. Il permet de diffuser et d'éditer presque tous les matériaux utilisés par le chercheur et l'enseignant : de l'archive ou la bibliothèque à la publication électronique en passant par le séminaire, le colloque, la revues et le livre. L'utilisation du web comme outil d'édition, de publication et de diffusion a permis de démultiplier les accès aux documents et à l'information. Mais depuis 20 ans, l'effort a plus porté sur la mise à disposition de documents numériques (ouvrages, articles, corpus) que sur la structuration de l'information contenue dans ces documents : il est vrai que l'essor des moteurs de recherche traditionnels depuis les années 90 (d'Altavista à Google) ont permis d'atteindre et de s'y « retrouver » dans ces milliards de documents qui sont sur le web aujourd'hui.

En revanche, la publication électronique des contenus des bases de données – les données elles-même qui ont toujours leurs propres structurations, pose encore des questions et des difficultés qui font que le web, s'il est plein de documents et relativement vide de données et d'informations structurées. Ainsi, les outils d'exploitation des documents que nous utilisons aujourd'hui, tel les moteurs de recherche, fonctionnent sur des réservoirs de documents encore trop cloisonnés. Ainsi, construire une page web d'information sur l'historien Georges Duby nécessite toujours d'adresser plusieurs questions (requêtes) à plusieurs moteurs de recherche (généralistes et spécialisés) ou à plusieurs formulaires de bases de données et cela même si, depuis dix ans, les techniques de l'interopérabilité ont fait de très grand progrès. Ce web « cloisonné » ne permet pas aux machines de travailler et certaines parties du web deviennent invisibles aux moteurs de recherche et même parfois aux humains (qui s'est déjà retrouver devant un formulaire de bdd en ligne un peu froid ?). Bien sur, un homme peut le faire, à la main, mais s'il veut se faire aider de machine, pour gagner du temps ou mieux, traiter plus de données, cela devient assez complexe. Surtout pour un chercheur qui ne maîtrise pas forcement le SQL et dont ce n'est pas le métier. Ainsi, les données numériques sont bien rangées dans de multiples bases de données ou silos, mais nous n'avons construit que de simples petits « judas » afin de les regarder et l'éditorialisation des données ne fait pas tout, pis, elle cache parfois, sous une couche « cosmétique » (cela dit souvent nécessaire), une faible structuration des données. La faible structuration des données freine très souvent les modes de pérennisation de ces dernières donc la possibilité de leur ré-exploitation future. Il nous faut faire mieux.

**Comment dépasser cela ?**

Comment rendre plus accessible encore, non pas simplement les documents (au sens des fichiers) mais les informations contenues dans ces derniers sans appauvrir les formats de structuration de l'information. Comment se donner l'opportunité de construire des outils d'aide à la recherche permettant de construire – par exemple – la notice encyclopédique de George Duby, en présentant, non pas simplement la compilation du signalement de ses articles, ouvrages, conférences, mais aussi les thèmes qu'il a abordé au cours de sa carrière et en les reliant à des notions, des définitions, des illustrations, des ouvrages d'autres auteurs ? C'est tout l'enjeu de la construction du web de données, cette extension du web dont je parlais dans mon dernier billet. Il nous faut tout d'abord libérer les données après l'avoir fait avec les bases de données elles-même.

**Comment faire ?**

Tout d'abord un peu d'histoire. Dans les années 1995-2000, tous les acteurs de la recherche et de la culture ont massivement édité leurs bases de données sur le web, c'était l'enjeu du moment : tout le monde voulait mettre sa base en ligne, c'était un nouveau cycle dans la diffusion des documents (après le minitel, les connexions client/serveurs). Nous sommes entrés, depuis quelques années, dans un nouveau cycle dont la première phase (la première « marche » je préfère dire) a été l'interopérabilité des bases de données. En parallèle de cette phase, qui se poursuit, nous devons

« ouvrir les données ». Quel curieuse expression ! Simplement, il s'agit d'exposer les données, dans toutes leurs complexités, en utilisant le cadre de la modélisation en RDF. Pour cela, il nous faut apprendre et développer des modèles de données, faire des choix de vocabulaires documentaires afin de décrire l'information contenue dans une page web, un billet de blog, un article, un inventaire de fonds d'archive, un corpus, un thésaurus ou encore une notice de bibliothèque. Pour ouvrir ces données il faut être capable de dire : « tiens ça, c'est le titre et ça là, c'est l'auteur et je te prouve que c'est bien l'auteur car je suis capable de le relier, par un principe ouvert, normalisé et connu de tous, à un référentiel (les auteurs du SUDOC par exemple) et à une forme de vocabulaire (du mods, du dublin core simple, etc.) » : les documentalistes savent très bien faire cela. Ainsi, ouvrir ses données – participer à la construction du web de données – cela revient donc à structurer de l'information avec des règles communes, valables pour tout le monde du web et où donc l'implicite n'est pas le bienvenu. Ouvrir ses données au monde c'est donc vouloir diffuser les données et par uniquement les documents et surtout dire quel choix j'ai fais pour structurer l'information. Les documentalistes font (devraient) s'y régaler.

Avec l'aide de Gautier Poupeau (Antidot), je vais présenter un exemple simple. Il est possible d'exprimer selon RDF des données structurées dans une page web écrite en HTML : il s'agit de la syntaxe RDFa (pour *Resource Description Framework – in – attributes*). RDFa permet donc d'utiliser la mécanique du RDF tout en utilisant comme support les balises HTML.

Je prends comme exemple, très simple, une photographie et sa notice venant de MédiHAL, l'archive ouverte de photographies scientifiques que j'ai co-créé et qui est développée par le centre pour la communication scientifique directe[1] (CCSD) et le CN2SV[2]. Au travers de cet exemple, je souhaite montrer qu'il ne s'agit pas que de techniques documentaires, ou que de questions informatiques, ou encore que de questions d'édition : non, il s'agit de tous cela en même temps. Ainsi, construire le web de données c'est avant tout réunir plusieurs compétences et métiers pour envisager toutes les aspects.

La consultation avec un simple navigateur web de la notice exemple ne révèle pas la présence d'une structuration de l'information selon les principes RDF et pourtant, si l'on regarde le code source, il y a une structuration, des vocabulaires RDF et des étiquettes structurant l'information. Ainsi, dans un premier temps, il faut dire que cette page contiendra du RDFa : j'ai modifié le doctype XHTML. Il est remplacé par un doctype XHTML+RDFa :

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN"
"http://www.w3.org/MarkUp/DTD/xhtml-rdfa-1.dtd">
```

Notez ensuite la présence de plusieurs vocabulaires documentaires qui vont nous permettre de structurer l'information :

```
<html xml:lang="fr" version="XHTML+RDFa 1.0"
        xmlns="http://www.w3.org/1999/xhtml"
        xmlns:foaf="http://xmlns.com/foaf/0.1/"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:cc="http://creativecommons.org/ns#"
        xmlns:dcterms="http://purl.org/dc/terms/"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
        xmlns:xsd="http://www.w3.org/2001/XMLSchema#">
```

Pourquoi ? Puisque nous allons structurer les données contenues dans cette page web, il nous faut dire « ça, c'est le titre » : il s'agit de mettre une « étiquette » à une chaine de caractère du titre. Il nous faut construire des triplets RDF qui, par l'utilisation de prédicats (verbe), relient l'étiquette (l'objet) à la chaîne de caractère du titre (sujet). Puisque nous devons dire à quel vocabulaire nous faisons référence pour dire « c'est le titre », nous les déclarons en entête. Vous

1 Voir : http://www.ccsd.cnrs.fr
2 Voir : http://www.cn2sv.cnrs.fr

reconnaitrez sans doute « dc » pour le dublin core simple (*dublin core elements set* ou dces), « dcterms » pour le dublin core terms, « cc » pour signaler la présence de données sous licence creative commons, « geo » pour la géolocalisation GPS, « foaf » pour décrire le document qui est ici une notice MédiHAL, etc. Ainsi je déclare là l'ensemble des vocabulaires documentaires que je vais utiliser ensuite et j'en donne la référence en ligne : http://purl.org/dc/elements/1.1/ pour le dublin core simple. Ces référentiels sont eux-même décrits et structurés en RDF : ils sont utilisés par tous et sont donc le point de référence, la norme.

Je trouve ensuite le début de ma notice, qui est matérialisée par une balise  :

```
<div typeof="foaf:Image" about="http://medihal.archives-ouvertes.fr/medihal-
00501617">
```

Dans cette balise (fermante à la fin de ma notice), j'y mentionne que ce qui sera dans la balise  est une notice d'une image et que l'URL présente dans l'attribut *« about »* sera l'objet auquel se rapporte les informations que je vais structurer (donc ici, un conteneur, une notice, d'une image). Les informations décrites par la suite se rapportent à cette notice (rôle du *« about »*), ce conteneur, accessible à cette URL. Ma données est complexe, elle est composés d'une image (qui a plusieurs représentations : plusieurs vignettes, l'image déposée, etc.) et des métadonnées, voir des commentaires (publics, privés). Pour décrire ce conteneur, j'utilise le vocabulaire foaf qui permet de décrire des ressources, des personnes ou des institutions et je vais utiliser l'élément foaf:Image. Pour la syntaxe, je vous invite à lire ce billet de Gautier Poupeau qui présente très en détail et très clairement la syntaxe des CURIEs (ou *Compact URIs*) dans le monde RDF.

Dans ce , je vais pouvoir structurer l'information contenue dans la données en utilisant, dans cet exemple, la balise  ainsi que quelques attributs : « property » pour caractériser l'information avec un vocabulaire, « rel » pour relier de l'information directement au conteneur. Ainsi pour le titre de l'image, je vais utiliser le dublin core simple (dces), nous aurons :

```
<span property="dc:title">Madagascar : Vallée de l'Onive aux environs de
Tsinjoarivo</span>
```

Pour l'image en jpg présentée dans la notice (qui est l'une des représentations possibles de l'image) :

```
<span rel="foaf:thumbnail" about="http://medihal.archives-ouvertes.fr/medihal-
00501617">
<img title="Madagascar..." id="thumb320" src="http://medihal.archives-
ouvertes.fr/docs/00/50/16/17/archives/thumb320.jpg" border="0" /></span>
```

Là, nous caractérisons que le contenu de <img/>, c'est à dire une image en 320 pixels, est l'une des versions de l'image de la notice représentée par « http://medihal.archives-ouvertes.fr/medihal-00501617 » : il s'agit d'une vignette de l'image d'ou « foaf:thumbnail ». Dans ce cas, il possible d'implémenter les attributs rel et about dans la balise <img>. Je l'ai mis dans un  pour plus de clarté. Notez que j'ai repéré dans ce  l'attribut *« about »*, je n'y suis pas obligé, il est déjà signalé dans la balise  « mère ». Ce  structurant une version de l'image (une vignette de 320px de coté), j'ai préféré ré-indiquer ce *« about »* afin que vous compreniez bien que foaf:thumbnail (vignette) désigne une vignette de l'image déposée et dont l'URI est http://medihal.archives-ouvertes.fr/medihal-00501617.

Pour la légende, je vais utiliser le vocabulaire dublin core *terms*, le plus riche des dublin core avec l'étiquette dc:abstract (pour résumé) :

```
<span property="dcterms:abstract">Paysage rural de collines à proximité de
Tsinjoarivo ; Au premier plan le bord de la terrasse de la vallée de l'Onive ; A
l'arrière-plan, cultures en terrasse avec des rizières en escaliers, irriguées
par un affluent du fleuve</span>
```

Je pourrais aussi, plus simplement mais en introduisant un peu d'implicite, utiliser dces avec l'étiquette dc:description :

```
<span property="dc:description">Paysage rural de collines à proximité de
Tsinjoarivo ; Au premier plan le bord de la terrasse de la vallée de l'Onive ; A
l'arrière-plan, cultures en terrasse avec des rizières en escaliers, irriguées
par un affluent du fleuve</span>
```

Pour exprimer les mots clés, je vais utiliser une nouvelle fois le dces :

```
<span property="dc:subject"><a href="[lien vers mes mots-
clés]">Madagascar</a></span>
```

Il est possibilité là aussi d'être plus riche, en reliant mon mot-clés à un référentiel (thésaurus par exemple) en utilisant les vocabulaires sioc et skos pour exprimer des concepts et les liaisons.

Pour la géolocalisation de mon image, je vais utiliser le dublin core terms avec l'étiquette « spacial », qui va me permettre de relier mon conteneur (foaf:Image) à des valeurs de latitude et de longitude. Ainsi, j'exprime dans dcterms:spatial une latitude et une longitude issues d'un GPS ou d'une géolocalisation en spécifiant que je fais référence au vocabulaire WGS validé par le W3C (geo:lat et geo:long).

```
<span rel="dcterms:spatial">
<span property="geo:lat" content="-19.644527589975"></span>
<span property="geo:long" content="47.709846500067"></span>
</span>
```

Je me limite ici à quelques éléments de cette image (en prenant du DC simple pour être pédagogique), il est possible d'aller plus loin dans la structuration (en utilisant du DC terms ou d'autres vocabulaires).

**Conclusion**

Le web de données est une méthode qui consiste à utiliser le web comme un espace ou les données sont structurées : c'est à dire que l'information d'un document (pdf, jpg, txt, etc.) est cartographiée, repérée, signalée et reliée à des vocabulaires, accessibles eux-même sur le web et dont la structuration est connue et explicitée. C'est un formidable enjeux pour les documentalistes, les bibliothèques et les ingénieurs et techniciens en *digital humanities* qui construisent des corpus scientifiques et les diffusent en ligne. Le RDFa est l'une des techniques, l'une des mécaniques possible et elle est relativement simple à comprendre car elle s'inscrit dans une évolution naturelle des choses : une « sémantisation » de la page web via le code HTML. Il s'agit d'une révolution mais qui s'appuie sur des éléments que tout les professionnels de l'IST peuvent maîtriser. J'ai toujours pensé et dit que l'OAI-PMH était (est) la première marche vers le web de données, je pense qu'RDFa est la deuxième, du moins c'est un pont très simple pour mieux comprendre RDF et les techniques du web de données.

Liens utiles pour aller plus loin :

- Exemple utilisé : http://www.stephanepouyllau.org/webdedonnees/medihal/rdfa/
- Code source : http://www.stephanepouyllau.org/webdedonnees/medihal/rdfa/medihal-rdfa.txt
- Une vue sur le contenu RDFa de cet exemple.
- Comprendre RDF : http://www.lespetitescases.net/comprendre-rdf-en-moins-de-5-minutes
- Mettre du RDFa dans son blog : http://www.lespetitescases.net/rdfaiser-votre-blog-1-la-theorie ; http://www.lespetitescases.net/rdfaiser-votre-blog-2-la-pratique ; http://www.lespetitescases.net/rdfaiser-votre-blog-3-exploitation
- Vidéo de l'ADBS : Le Web de données : perspectives pour les métiers de l'information documentation