LES TROIS VISAGES DU WEB SÉMANTIQUE

Guillaume SIRE
Maître de conférences en sciences de l'information et de la communication
Co-responsable de l'Unité régionale de formation à l'information scientifique et technique de la région Occitanie / Pyrénées-Méditerranée
Membre du Laboratoire d'Études et de Recherches Appliquées en Sciences Sociales, de l'Institut du Droit de l'Espace, des Territoires, de la Culture et de la Communication et du Centre d'Analyse et de Recherche Interdisciplinaire sur les Médias

■ Nous présentons les trois principales syntaxes du web sémantique, ainsi que leurs trois procédés de normalisation et le fonctionnement des institutions qui se chargent de normaliser chacune d'elles : le *World Wide Web Consortium* pour RDF/RDFa, *Commercenet* pour les microformats et le *Web Hypertext Application Technology Working Group* (*Whatwg*) pour les microdonnées. Tout en expliquant la différence entre ces syntaxes et leurs institutions, nous montrons ainsi pourquoi chaque syntaxe est le fruit de valeurs qui lui sont propres et entre lesquelles les développeurs, désireux de rendre "compréhensibles" leurs données, doivent arbitrer. L'enjeu n'est rien moins que la cartographie conceptuelle du plus grand espace informationnel jamais vu, qui pour l'instant est encore cartographié sur le mode hypertextuel qui a certes prouvé son efficacité, mais a également montré ses limites, notamment concernant le traitement algorithmique.

■ We stellen u de drie belangrijkste syntaxis van het semantisch web voor, evenals hun drie standaardisatieprocessen en de werking van de instellingen die zich bezighouden met het standaardiseren van elk van hen: het *World Wide Web Consortium* voor RDF/RDFa, *Commercenet* voor microformats en de *Web Hypertext Application Technology Working Group* (*Whatwg*) voor microdata. Terwijl we het verschil tussen deze syntaxis en hun instellingen uitleggen, tonen we aan waarom elke syntaxis het resultaat is van zijn eigen waarden en waartussen de ontwikkelaars die hun gegevens "verstaanbaar" willen maken, moeten bemiddelen. De uitdaging is niets minder dan het conceptueel in kaart brengen van de grootste informatieruimte ooit die op dit moment nog steeds in een hypertextuele modus wordt uitgebeeld, wat zeker zijn doeltreffendheid heeft bewezen, maar tevens zijn beperkingen heeft aangetoond, met name op het gebied van algoritmische verwerking.

On a parlé un peu vite, dans les années 90, de désintermédiation, et expliqué que grâce à Internet il n'y aurait désormais plus de "gatekeepers" aux portes de l'espace public, puisque n'importe qui, n'importe quand, pourrait s'exprimer à propos de n'importe quoi. Cette désintermédiation n'a en fin de compte jamais eu lieu. Les fameux "gatekeepers", journalistes, éditeurs, experts, existent encore. Les données de l'équation ne sont plus tout à fait les mêmes, certes, mais le problème général est inchangé : certaines informations accèdent à l'avant-scène de l'espace public, d'autres demeurent dans les coulisses. La chaîne de médiation, à bien y regarder, semble plus longue sur Internet que sur les autres médias. Des acteurs sont apparus qui ne produisent pas d'informations, mais hiérarchisent automatiquement les informations produites par d'autres. Ces acteurs, moteurs de recherche, réseaux sociaux, fonctionnent grâce à des algorithmes paramétrés de façon à "comprendre" autant que possible de quoi il est question dans un contenu, à "rapprocher" les contenus traitant de sujets similaires dans un index thématique et à" mesurer " la pertinence supposée de ces contenus les uns par rapport aux autres.

La tâche des ingénieurs qui conçoivent de tels algorithmes est d'autant plus difficile que sur le web, il est compliqué voire impossible d'imposer une même norme de publication à tous ceux qui publient des contenus. Or, sans norme commune, il est presque impossible de paramétrer a priori un outil de traitement automatique. C'est pourquoi sur le web les premiers moteurs de recherche fonctionnaient très mal. *Exalead*, *Altavista*, *Lycos* peinaient à répondre aux requêtes, tandis que *Yahoo*, annuaire composé à la main, était la voie d'accès privilégiée, tout du moins jusqu'à l'arrivée de *Google* en 1998. L'ingéniosité des concepteurs de *Google* consista à ajouter aux données fournies par les contenus (le texte) les données fournies par les liens hypertextes entre contenants (les pages). Le moteur pouvait ainsi dresser une topographie du web en positionnant les pages les unes par rapport aux autres, et les hiérarchiser en considérant que chaque lien pointant vers un document était un indice positif quant à la pertinence potentielle de ce document. Mais même si elle a permis de rendre plus pertinents les logiciels de traitement automatique, et notamment les moteurs de recherche, cette astuce n'était pas suffisante aux yeux de leurs concepteurs qui continuèrent à espérer qu'une harmonisation des formats méta-informatifs ait lieu, laquelle permettrait de perfectionner le paramétrage a priori des logiciels et de rendre plus pertinents les résultats du traitement automatique.

Le problème que nous évoquons ici tient moins au fait qu'il n'existe pas de norme de publication sur le web qu'au fait qu'il en existe plusieurs, émanant de différentes institutions qui, chacune, essayent

de s'imposer comme référence. Le langage HTML par exemple est normalisé au sein du *World Wide Web Consortium* (W3C). Il permet de décrire les documents de manière à optimiser leur affichage sur les navigateurs et leur traitement algorithmique par les moteurs de recherche et les réseaux sociaux. Le langage *Flash*, en revanche, édité par Adobe, est une norme de publication concurrente reconnue par les navigateurs à condition qu'ils soient équipés d'un module d'extension spécifique, mais qui n'est pas reconnue par les moteurs de recherche. Au final, une espèce de tectonique des normes se met en place, dont les modalités et l'issue dépendent des rapports de forces entre des acteurs aux intérêts différents. Et cette tectonique concerne, plus largement, chaque utilisateur d'Internet, dès lors qu'elle contribue à déterminer ce qu'il est possible de faire ou non sur Internet.

L'un des principaux (il y en a d'autres[1]) enjeux actuels de la normalisation a trait non pas vraiment à la hiérarchisation automatique, mais plutôt à ce qu'on pourrait appeler la *compréhension automatique* des contenus. Il s'agit de normaliser des systèmes méta-informatifs permettant d'apporter une couche de sens aux documents mis en ligne, de sorte que les intermédiaires puissent "comprendre" de quoi il est question dans tel ou tel document, et quels sont les liens entre les concepts mobilisés. Il s'agit autrement dit de prévenir les logiciels d'indexation de ce à quoi tel mot ou telle expression renvoie (s'agit-il d'un prénom ? du prénom d'un artiste ? s'agit-il d'une chose ? est-ce que cette chose est dangereuse ?) et quelles sont les relations entre les objets (quels sont les peintres qui ont inspiré ce peintre ? quels sont ceux qu'il a lui-même inspirés ? quels sont les objets dont l'usage est complémentaire à cet objet-là ? quels sont les produits susceptibles de s'y substituer ?). Si on parvient à rajouter cette couche d'informations sémantiques, il deviendra possible d'indexer non plus seulement les documents et leurs relations hypertextuelles, mais aussi les concepts et leurs relations conceptuelles (ou *prédicats*). Si je publie par exemple une recette de cuisine, et que j'emploie en la décrivant le mot "coriandre", je pourrai rajouter une balise dans le code HTML encadrant le mot "coriandre" pour prévenir les moteurs de recherche qu'il s'agit d'un ingrédient de la recette (contrairement aux mots qui précèdent : "*Ajoutez un peu de…*" et à certains mots qui pourraient avoir l'air de noms d'ingrédients mais qui n'en sont pas : "*Ne mettez pas de sucre*"). Je pourrai également dire au moteur que la coriandre est un végétal, et qu'on en trouve souvent dans la cuisine asiatique, et je pourrai enfin renseigner le temps de préparation et le temps de cuisson, de manière à ce que les logiciels de traitement puissent tisser un réseau de relations autour des concepts eux-mêmes, et non plus seulement des documents, et ainsi disposer d'un graphe plus complexe que celui qui lui était fourni par la seule topographie hypertextuelle des relations entre documents. Cet ajout de méta-informations normalisées directement à l'intérieur du code HTML renvoie à ce qu'on appelle "le web sémantique".

Il existe trois arènes principales de normalisation du web sémantique, chacune produisant, selon un processus qui lui est propre, sa propre grammatologie : le Word Wide Web Consortium produit le *Resource Description Framework* (RDF), *CommerceNet* produit les microformats et le *Web Hypertext Application Technology Working Group* (*Whatwg*) produit les microdonnées. Ci-dessous, nous présenterons ces trois institutions, leur histoire et leurs particularités ainsi que leurs visions du web sémantique, avant d'expliquer la manière dont ces trois formats tantôt se complètent, tantôt se concurrencent et de discuter des enjeux socioéconomiques et informationnels d'une telle situation.

W3C et RDF/RDFa

Le *World Wide Web Consortium* ou " W3C " a été créé par le fondateur du web : Tim Berners-Lee. C'est une organisation à but non lucratif ouverte à toutes les personnes morales qui souhaitent participer aux discussions menées en son sein. C'est au W3C que sont normalisés en particulier les trois protocoles centraux du web mis au point par Tim Berners-Lee : l'*Uniform Resource Locator* (URL), l'*HyperText Transfer Protocol* (HTTP) et le fameux *HyperText Markup Language* (HTML).Chaque proposition de modification d'un protocole existant ou d'ajout d'un nouveau protocole fait l'objet d'une discussion entre les membres réunis en un même groupe de travail. Elle donne ensuite lieu à un document censé franchir quatre étapes à l'issue desquelles, s'il les a toutes franchies, il obtiendra le statut de norme officielle du W3C. Le franchissement de chaque étape se fait par consensus, sachant que c'est Tim Berners-Lee, ou une personne qu'il a expressément mandatée, qui décrète qu'il y a consensus sur telle ou telle question et qu'il est temps de passer à l'étape suivante, ou qui décrète au contraire que le consensus n'a pas été trouvé et qu'il convient par conséquent de revenir à une étape précédente, ou tout simplement d'abandonner le projet. Le processus du W3C confère autrement dit un pouvoir considérable à Tim Berners-Lee, qui peut prendre éventuellement en dernière instance des décisions qui ne vont pas dans le sens de l'avis exprimé par la majorité des membres d'un groupe de travail ou bien trancher alors qu'il reste des objections formelles exprimées par tel ou tel membre et n'ayant pas été résolues. Pour cette raison, certains n'hésitent pas à le qualifier de "dictateur à perpétuité"[2].

Au sein du W3C, le groupe de travail intéressé par les langages sémantiques s'appelle le *Semantic Web Interest Group* (SWIG). C'est en son sein qu'a été développé le RDF, basé sur la syntaxe XML, et ayant atteint le statut de norme officielle en 2004 en version 1.1. Nous n'allons pas entrer dans le détail, mais il est important de comprendre pour la suite que le RDF est construit de telle sorte que soient renseignés des triplets sous la forme (sujet { prédicat } objet). Par exemple[3]pour la phrase : "Guillaume a écrit un article pour *Les Cahiers de la Documentation* intitulé "Le web sémantique", nous aurons les trois *prédicats* suivants : 1/ Guillaume { est rédacteur de } "Le triple visage du web sémantique" ; 2/ *Les Cahiers de la Documentation* { publie } "Le triple visage du web sémantique" ; 3/ "Le triple visage du web sémantique" { est } un Article. Nous pourrons ensuite créer d'autres relations : si Guillaume est rédacteur de "Le triple visage du web sémantique" et *Les Cahiers de la Documentation* publie "Le triple visage du web sémantique", alors Guillaume est un contributeur des Cahiers de la Documentation. Les sujets et les objets peuvent être des personnes, des choses, des concepts, des dates, des lieux, ou bien des adresses URL. Si on parvient à donner à chaque "élément" (personnes, choses, concepts, dates, lieux...) un seul Uniform Resource Identifier (URI), dans une base de données où sont renseignés les triplets, il devient possible ensuite de créer des relations conceptuelles à l'échelle du web.

Concrètement, pour employer le RDF, il faut utiliser un schéma de métadonnées permettant aux développeurs de décrire les éléments de leurs pages web de façon à être compris par les logiciels type moteurs de recherche, lesquels pourront ensuite faire les liens eux-mêmes entre sujets et objets, ou bien se référer à une base de données comme *DBpédia* (une collection de triplets constitués à partir des ressources RDF contenues dans *Wikipedia*), à condition que ces éléments y existent. Il existe plusieurs "schéma de métadonnées" ou "espaces de noms" utilisés en RDF, par exemple *Dublin Core* et *Friend Of A Friend* qui peuvent être utilisés en plus de l'espace de nom RDF qui lui aussi existe. Parce que plusieurs schémas peuvent être utilisés au sein d'une même page, il est indispensable de renseigner les préfixes à chaque fois dans les balises tout en ayant prévenu en tête de la page web quels étaient les différents schémas utilisés et à quelle adresse le logiciel pourra en trouver une description (on se sert pour cela du préfixe "xmlns" pour "xmlnamespace"). Cela se présente de la manière suivante[4] :

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/metadata/dublin_core#"
 xmlns:dcq="http://purl.org/metadata/dublin_core_qualifiers#">
<rdf:Description about="http://www.dlib.org/dlib/may98/05contents.
html">
<dc:Title>DLIB Magazine - The Magazine for Digital Library
Research
    - May 1998</dc:Title>
<dc:Description>D-LIB magazine is a monthly compilation of
contributed stories, commentary, and briefings.</dc:Description>
<dc:Contributorrdf:parseType="Resource">
<dcq:AgentType
        rdf:resource="http://purl.org/metadata/
dublin_core_qualifiers#Editor"/>
<rdf:value>Amy Friedlander</rdf:value>
</dc:Contributor>
<dc:Publisher>Corporation for National Research Initiatives</
dc:Publisher>
<dc:Date>1998-01-05</dc:Date>
<dc:Type>electronic journal</dc:Type>
<dc:Subject>
<rdf:Bag>
        <rdf:li>library use studies</rdf:li>
        <rdf:li>magazines and newspapers</rdf:li>
</rdf:Bag>
</dc:Subject>
<dc:Format>text/html</dc:Format>
<dc:Identifier>urn:issn:1082-9873</dc:Identifier>
<dc:Relationrdf:parseType="Resource">
<dcq:RelationType
        rdf:resource="http://purl.org/metadata/
dublin_core_qualifiers#IsPartOf"/>
<rdf:value resource="http://www.dlib.org"/>
</dc:Relation>
</rdf:Description>
</rdf:RDF>
```

Ici on voit dès les premières lignes qu'il y a trois schémas utilisés, correspondant chacun à un préfixe : RDF (rdf), *Dublin Core* (dc) et *Dublin Core* Qualifier (dcq). Et l'on voit dans le code du document que chaque fois qu'une balise est utilisée le préfixe est renseigné, de sorte qu'une machine puisse savoir à quel schéma se référer pour "comprendre" le contenu de telle ou telle balise.

Dans le HTML5, on utilise le *Resource Description Framework in Attributes* (RDFa) normalisé par le W3C en 2008 pour la version 1.0 et en 2013 pour la version 1.1. Il fonctionne comme le RDF, mais les préfixes cette fois s'inscrivent dans les attributs des balises HTML et non dans le nom des balises, ce qui simplifie considérablement l'écriture. Le RDFa utilise les attributs HTML existants (notamment class, id, rel, rev et href) et en institue de nouveaux (about, property, content, datatype, resource, typeof, prefix, vocab). Cela donne[5] :

```
<pxmlns:dc="http://purl.org/dc/elements/1.1/"
about="http://www.example.com/books/wikinomics">
   Dans son dernier livre
<emproperty="dc:title">Wikinomics</em>,
```

<spanproperty="dc:creator">Don Tapscott
explique les profonds changements technologiques, démographiques et économiques.
Ce livre a été publié en
<spanproperty="dc:date"content="2006-10-01">octobre 2006.
</p>

Cette fois-ci, les balises sont bien des balises HTML classiques, mais le "xmlns" renseigné à la première ligne donne lieu à des préfixes à l'intérieur des attributs, c'est-à-dire à l'intérieur des balises, de sorte qu'une même balise puisse contenir plusieurs attributs associés chacun à un schéma de métadonnées spécifique.

Le RDF comme le RDFa sont tous les deux normalisés, fixés par le W3C, qui en garantit d'une part l'efficacité technique et qui d'autre part garantit qu'aucun détenteur de brevet nécessaire à l'implémentation de ces formats n'essaiera d'en tirer parti financièrement[6].

CommerceNet et Microformats

Les microformats (auxquels renvoient les sigles µF ou uF) ont été développés sans politique de normalisation arrêtée, de façon participative et "sauvage". Cela les différencie des syntaxes RDF/RDFa normalisées selon la procédure du W3C. Il est d'ailleurs très facile de proposer un nouveau type de microformat en créant son propre espace de nom et sa propre syntaxe en fonction des besoins qu'on a soi-même et/ou qu'on a identifiés chez les autres. L'organisation à but non lucratif *CommerceNet*, dont la vocation est de promouvoir le commerce électronique, a aidé à mettre en place les microformats et à fédérer une communauté autour du wiki "Microformats.org" ouvert à tous.

Les microformats reposent sur trois attributs (class ; rel ; rev). Ces attributs peuvent être insérés dans n'importe quelle balise HTML et, dans le cas où il n'y en a pas déjà une à l'endroit où il faut insérer l'attribut, ils peuvent être ajoutés grâces aux balises ou . Tous les microformats existants sont répertoriés sur le wiki[7]. Chacun est associé à un type d'information (description d'une personne, localisation d'un lieu, CV, petites annonces...) et assorti d'une page web présentant aux développeurs comment l'utiliser. Il suffit d'annoncer à quel "microformat" on se réfère dans une balise HTML située en amont de la page et de suivre les instructions du wiki. Par exemple, pour le microformat "hCard", nous avons :

<pclass="h-card">
<imgclass="u-photo"src="http://example.org/photo.png"alt=""/>
<aclass="p-name u-url"href="http://example.org">Joe Bloggs
<aclass="u-email"href="mailto:joebloggs@example.com">joebloggs@example.com,
<spanclass="p-street-address">17 Austerstræti
<spanclass="p-locality">Reykjavík
<spanclass="p-country-name">Iceland
</p>

Nous voyons ici que le nom du microformat (hCard) joue le rôle joué par l'espace de nom dans RDF/RDFa. Mais contrairement aux espaces de noms, aucun lien n'est effectué vers le lieu où la syntaxe est décrite. Si le logiciel ne connaît pas déjà cette syntaxe (i.e. si son concepteur ne l'a pas paramétré en fonction de cette syntaxe), iln'aura aucun moyen de "comprendre" de quoi il est question.

Une autre des différences majeures entre le RDF/RDFa et les microformats tient au fait que les microformats ne permettent pas de renvoyer à des vocabulaires existants type *Dublin Core*. Il est impossible, autrement dit, de s'appuyer sur des bibliothèques de triplets déjà constituées dans d'autres langages sémantiques. Enfin, contrairement au RDF/RDFa, les microformats reprennent les attributs existants sans en ajouter de nouveaux.

Whatwg et Microdonnées

Le *Whatwg* a été créé à la suite d'un désaccord au sein du W3C. Une scission a eu lieu en 2005, lorsque Ian Hickson a proposé de travailler sur une version plus interactive du langage HTML, le HTML5, mais que Tim Berners-Lee a refusé d'ouvrir un groupe de travail, cela parce qu'il préférait se concentrer sur le XHTML et sur l'objectif de "sémantiser" le web. Ian Hickson ne renonça pas à son projet. Il collabora avec des ingénieurs d'Apple, Mozilla et Opera, puis *Google* à partir de 2005 (qui embaucha Ian Hickson) pour créer le *Web Hypertext Application Technology Working Group* (*Whatwg*). Ce groupe de travail fut doté d'une procédure extrêmement souple comparée à celle du W3C. L'objectif était d'y normaliser une version du code HTML en dehors du W3C, ce qui était possible dès lors que les membres du *Whatwg* étaient les propriétaires des navigateurs et des moteurs de recherche et pouvaient par conséquent paramétrer leurs logiciels de manière à exécuter les balises qu'ils auraient eux-mêmes mises au point, en plus de celles qui auraient été discutées et validées par le W3C.

Finalement, en 2007, le XHTML 2.0 était l'objet de plusieurs dysfonctionnements techniques, en conséquence de quoi Tim BernersLee proposa aux membres du *Whatwg* d'ouvrir un groupe de travail au sein du W3C concernant le HTML5, ce que Ian Hickson et ses collaborateurs acceptèrent, sans pour autant mettre un terme à l'activité du *Whatwg*,

disposant ainsi de deux structures où discuter du même protocole. Les navigateurs et les moteurs de toute manière reconnaîtraient les balises du code HTML5, et ce, qu'elles émanent du W3C ou bien du *Whatwg*. Les deux arènes éditeraient le langage conjointement, chacune selon sa propre procédure, et les concepteurs de navigateurs participeraient aussi bien à l'une qu'à l'autre, discutant entre eux au *Whatwg* et avec d'autres au W3C mais pouvant décider, dans le cas où une recommandation du W3C ne leur conviendrait pas, de ne pas l'implémenter et de lui préférer des modalités issues du seul *Whatwg* dans le cas où un arbitrage devrait avoir lieu.

C'est dans le cadre de ce *Whatwg* qu'ont été développées les microdonnées. Il est intéressant ici de bien considérer que ce sont les propriétaires de navigateurs et de moteurs de recherche qui, eux seuls, ont mis en place ce format. Il repose sur le principe du "Living Standard" : des modifications peuvent être faites rapidement, sans avoir à suivre les étapes et le protocole plus rigide des normes stabilisées par le W3C. Il n'y a pas non plus de politique de brevet afférente. Les microdonnées fonctionnent à partir d'attributs spécifiques : itemscope, itemtype, itemid, itemprop, itemref. L'attribut itemtype permet de renvoyer à un vocabulaire faisant office d'espace de nom, et présent sur le site "Schema.org". Par exemple ci-dessous, il s'agit d'une recette de cuisine dont la première ligne nous apprend qu'elle utilise le vocabulaire décrit ici : "http://schema.org/Recipe".

```
<divitemscopeitemtype="http://schema.org/Recipe">
  <spanitemprop="name">Mom's World Famous Banana Bread</span>
 By <spanitemprop="author">John Smith</span>,
 <metaitemprop="datePublished"content="2009-05-08">May 8, 2009
  <imgitemprop="image"src="bananabread.jpg"
     alt="Banana bread on a plate"/>
  <spanitemprop="description">This classic banana bread recipe comes
  from my mom -- the walnuts add a nice texture and flavor to the banana
  bread.</span>
 Prep Time: <metaitemprop="prepTime"content="PT15M">15 minutes
 Cook time: <metaitemprop="cookTime"content="PT1H">1 hour
  Yield: <spanitemprop="recipeYield">1 loaf</span>

[...]
</div>
```

En 2011, après l'officialisation des microdonnées, TantekCelik, l'un des créateurs des microformats, a accusé sur son compte Twitter le *Whatwg* de "*cracher dans les yeux de toutes les personnes et les organisations ayant œuvré à la conception des vocabulaires ouverts vCard, iCalendar, etc.*"[8]. De son côté, le développeur très influent Mark Pilgrim a prévenu que le Schema.org était l'illustration de l'échec du W3C avec son RDF/RDFa[9]. Manu Sporny, qui dirigeait le groupe de travail du W3C autour de la spécification RDF, a quant à lui pointé du doigt le fait que les microdonnées étaient le fait d'un petit groupe d'organisations, alors les RDFa et les microformats étaient le fruit du travail collectif de milliers de personnes, et que par conséquent les microdonnées ne pouvaient pas être considérées comme étant de véritables normes ouvertes[10]. La dispute a été virulente, et les arguments avancés pour l'une ou l'autre des trois possibilités visaient tous à acquérir une légitimité plus forte que les autres, soit en prétendant que la syntaxe défendue était plus efficace techniquement et plus facile à généraliser, soit en attaquant le fonctionnement des arènes de normalisation concurrente[11].

Entre complémentarité et concurrence

Nous avons vu que RDFa permettait de recourir à des vocabulaires RDF/RDFa existants (*Dublin Core*, *SKOS*, *OWL*...) alors que les microformats et les microdonnées utilisent des syntaxes et des vocabulaires spécifiquement faits pour eux et rassemblés sur le site Schema.org pour les microdonnées et sur le site Microformats.org pour les microformats. En outre, les RDFa peuvent être adossés à des bases de données déjà constituées comme *DBpédia* grâce aux URI, ce que ne permettent pas de faire les microformats et les microdonnées. C'est pourquoi certains développeurs en concluent que :"*RDFa offre donc une puissance supplémentaire, mais par ricochet, présente une complexité plus grande, d'où l'importance que les deux syntaxes coexistent*"[12].

Les microdonnées sont plus simples à utiliser que les RDFa et mieux uniformisées que les microformats. En revanche, elles émanent directement des concepteurs du moteur de recherche *Google*, ce qui laisse à penser qu'elles sont optimales pour les questions de référencement. A une époque, *Google* avait même prévenu les développeurs que son moteur privilégierait l'usage des microformats par rapport aux deux autres types de sémantisation (*Google* est revenu sur ces propos depuis). En revanche, le Schema.org est peu évolutif et les discussions à son sujet, contrairement aux deux autres formats, ne sont pas ouvertes. Les discussions concernant les microformats sont ouvertes en effet à tout le monde, tandis que les discussions concernant le RDF/RDFa ne le sont qu'aux personnes morales choisissant de devenir membres du W3C.

La politique de la signification

Dans cet article dont le niveau technique et les nombreux sigles n'auront, je l'espère, pas trop dégoûté le lecteur, j'ai voulu montrer comment sur le web trois normes différentes proposent trois visions à la fois de ce que peut être la sémantisation des données, et de ce que peut être la manière adéquate de prendre des décisions à son sujet. Ce n'est pas seulement une question d'ordre technique, mais aussi, et davantage, une question politique. Des rapports de force ont lieu entre des acteurs qui tous ont intérêt à ce qu'un protocole l'emporte sur les autres : intérêt technique, intérêt économique, intérêt juridique.

Dans son célèbre article "*Le code c'est la loi*", Lawrence Lessig prévenait que le code informatique déterminerait nos valeurs, avant de poser la question : "*Si c'est le code qui détermine nos valeurs, ne devons-nous pas intervenir dans le choix de ce code ? Devons-nous nous préoccuper de la manière dont les valeurs émergent ici ?*"13. La question ici se pose concernant les syntaxes qui sous-tendent la sémantisation des données d'Internet. Quelle est-elle, que fait-elle, qui la fait, et comment ? Quels sont les procédés de normalisation ? Dans le cas du W3C, la discussion est ouverte à toutes les personnes morales ; cependant les décisions sont prises par Tim Berners-Lee, le fondateur du W3C, qui seul peut décréter qu'il y a consensus. Récemment, son pouvoir a d'ailleurs fait l'objet de très vives remises en cause14. Chez *Commercenet*, la discussion est ouverte à tous, ce qui peut aboutir à un manque de cohérence globale. La sémantisation a lieu par îlots hypertextes, mais les pratiques des uns et des autres sont trop différentes pour qu'une base de données puisse être constituée et structurée à partir des seuls microformats à l'échelle du web tout entier. Enfin, les microdonnées émergent du *Whatwg*, c'est-à-dire d'une poignée d'acteurs beaucoup plus puissants que tous les autres. Elles sont extrêmement efficaces mais peu évolutives et de toute façon ne font pas l'objet d'un processus de normalisation ouvert.

Ces sujets sont techniques, c'est vrai, mais il n'empêche : nous sommes tous concernés. Choisir comment le web peut faire sens, c'est choisir comment nous-mêmes nous pourrons y trouver et y produire du sens. A titre personnel, je regrette que ces enjeux soient si peu considérés par les professionnels de la documentation, et, plus généralement, les citoyens, les politiques, etc. Car négliger ces questions, c'est laisser les ingénieurs, et les géants du web en particulier, s'en emparer et nous imposer leur vision, pourtant éminemment critiquable, de ce qu'est le sens et de ce que sont les conditions de sa production et de sa mise en circulation.

Guillaume Sire
Rue du Doyen-Gabriel-Marty, 2
31042 Toulouse
France
guillaume.sire@ut-capitole.fr
http://urfist.univ-toulouse.fr/
Décembre 2017

Notes

1. Sire, Guillaume. Lutte fratricide dans les coulisses du web. Ina Global, novembre 2017 (consulté le 20 décembre 2017). < http://www.inaglobal.fr/numerique/article/lutte-fratricide-dans-les-coulisses-du-web-10024>.
2. Voir par exemple : Malcolm, Jeremy. Multi-Stakeholder governance and the Internet governance forum. Terminus Press, 2008.
3. Cet exemple est une transposition de l'exemple donné par David Larlet dans son article très clair au sujet du web sémantique : Larlet, David. Le point sur RDF et RDFa avril 2008 (consulté le 20 décembre 2017). <https://larlet.fr/david/biologeek/archives/20080425-le-point-sur-rdf-et-rdfa/>
4. Exemple tiré du Site W3C.<https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/> (consulté le 20 décembre 2017)
5. Exemple provenant de Wikipédia.<https://fr.*wikipedia*.org/wiki/RDFa> (consulté le 20 décembre 2017)
6. Site du W3C.<https://www.w3.org/Consortium/Patent-Policy-20170801/> (consulté le 20 décembre 2017)
7. Le wiki des microformats.<http://microformats.org/wiki/Main_Page> (consulté le 20 décembre 2017)
8. Twitter.<https://twitter.com/t/status/770834814941429 76> (consulté le 20 décembre 2017)
9. Stevens, Luke. The Truth about HTML. 2012-2013 edition, Indie Digital, 2012, p. 87.
10. Sporny, Manu. The False Choice of Schema.org. The Beautiful, Tormented Machine, juin 2011 (consulté le 20 décembre 2017). <http://manu.sporny.org/2011/false-choice/>
11. Pour un aperçu de la dispute entourant les trois formats, voir le Site du W3C : <https://www.w3.org/2011/06/

semtech-bof-notes.html> (consulté le 20 décembre 2017)

12. Pour aller plus loin avec vos RDFa, Développez.com, mai 2011-janvier 2016 (consulté le 20 décembre 2017), <https://web-semantique.developpez.com/tutoriels/lpc/complements-rdfa/>
13. Lessig, Lawrence. *Le code c'est la loi*. Framablog (consulté le 20 décembre 2017), <https://framablog.org/2010/05/22/code-is-law-lessig/>
14. Sire, Guillaume. Lutte fratricide dans les coulisses du web. Ina Global, novembre 2017 (consulté le 20 décembre 2017). <http://www.inaglobal.fr/numerique/article/lutte-fratricide-dans-les-coulisses-du-web-10024>.