

Web sémantique : les politiques du sens et la rhétorique des données

Semantic Web: the politics of meaning and the rhetoric of data

Web Semántica: la política del significado y la retórica de los datos

Article inédit, mis en ligne le 15 novembre 2018.

Guillaume Sire

Maître de conférences en sciences de l'information et de la communication à l'Université Toulouse 1 Capitole et membre de l'Institut du Droit de l'Espace, des Territoires, de la Culture et de la Communication, et co-responsable de l'Unité Régionale de Formation à l'Information Scientifique et Technique d'Occitanie. Ses travaux portent sur la gouvernance du Web, les moteurs de recherche, le code informatique et la socio-économie des industries culturelles. / guillaume.sire@ut-capitole.fr

Plan de l'article

Introduction

W3C et RDF/RDFa

CommerceNet et microformats

Whatwg et Microdonnées

Conclusion : l'industrialisation du sens

Références bibliographiques

Résumé

Nous expliquons comment fonctionne le Web sémantique, et en particulier comment les producteurs de contenus peuvent « discrétiser » les informations de manière à ce que des logiciels comme les moteurs de recherche puissent ensuite les corréliser à l'échelle du Web en générant des ontologies. Nous comparons les trois syntaxes auxquels ils peuvent avoir recours : le *Resource Description Framework*, les microdonnées et les microformats, ainsi que les arènes et les processus de normalisation propres à chacune de ces syntaxes. En nous référant pour notre analyse à la théorie opérationnelle de l'écriture numérique développée à l'Université Technologique de Compiègne, nous montrons ainsi comment sont confrontées différentes « politiques du sens » sur le Web portées par des acteurs dont les objectifs et les intérêts diffèrent. Nous décryptons la teneur non pas simplement du mécanisme de description et de traitement de l'information mais aussi des allers-retours entre l'individu qui décrit et édite des informations et la machine qui traite et éditorialise des données.

Mots clés

Gouvernance d'Internet, Web sémantique, ontologies, HTML, RDF, microdonnées.

Abstract

We explain how the semantic Web works, and in particular how content providers can "discretize" information in order to correlate them within ontologies. We compare three syntaxes they can use: *Resource Description Framework*, microdata and microformats, and we compare the normalization processes of each of these syntaxes. We show how different "politics of meaning" are confronted on the Web and supported by companies whose objectives and interests are far to be the same. We analyze not just the mechanism of description operated by hypertext markups but also the mechanism of interaction between the people who describe and publish content on the Web and the machines that process and editorialize data.

Keywords

Internet Governance, Semantic Web, Ontologies, HTML, RDF, microdatas.

Resumen

Explicamos cómo funciona la Web semántica y, en particular, cómo los productores de contenido pueden "discretizar" la información para que software como los motores de búsqueda puedan correlacionarlos en la Web generando ontologías. Comparamos las tres sintaxis que pueden usar: el Marco de descripción de recursos, los microdatos y los microformatos, y las arenas y procesos de normalización específicos de cada una de estas sintaxis. De esta manera, mostramos cómo diferentes "políticas de significado" en la Web son confrontadas por actores cuyos objetivos e intereses difieren. Y desciframos el contenido no solo del mecanismo de descripción y procesamiento de la información, sino también del recorrido de ida y vuelta entre el hombre que describe y publica la información y la máquina que procesa y editorializa los datos.

Palabras clave

Gobernanza de Internet, Web Semántica, Ontologías, HTML, RDF, microdatos.

Introduction

Le terme « *Big Data* » désigne un phénomène double du point de vue de l'information. L'épithète *Big* renvoie à la surcharge informationnelle, le terme « *Data* » aux données, c'est-à-dire à un flux d'informations non pas linéaire mais « discrétisé ». Cette discrétisation a lieu lorsque « *le contenu est inscrit en un langage constitué d'unités discrètes indépendantes les unes des autres* » (Bachimont 1999, p. 200). Discrétiser une information autrement dit revient à la « *transformer en une suite d'unités élémentaires (...) manipulables par une machine* » (Goyet, 2017, p. 68). Les « *Big Data* » supposent de traiter une quantité indénombrable d'unités qualifiables. Comment et à quelles conditions l'information ainsi multipliée et discrétisée par une succession de dispositifs sociotechniques peut-elle avoir du « sens » ? Et que devient dans ce cas la tension inhérente à toute technique, « *entre l'ouverture au sens et la fermeture sur une utilité fixée a priori* » (Bachimont, 2010, p. 24) ? Voilà des questions qui relèvent du champ des sciences de l'information et de la communication et dont les enjeux ne sont pas seulement techniques, mais aussi, et tout autant, culturels, politiques, économiques et juridiques.

Le Web a été le théâtre de la multiplication des informations, d'une part, et, d'autre part, de leur discrétisation. Tandis que la multiplication avait lieu parce que l'action de publier était désormais possible pour n'importe qui dès lors qu'on possédait de quoi se connecter, la discrétisation eut lieu étant donné le besoin de normaliser les pratiques de production, de sorte que les informations une fois publiées pourraient effectivement être traitées par les navigateurs chargés de les afficher, et par les moteurs chargés de les indexer. Une des étapes fondamentales de cette discrétisation consiste à diviser le contenant (la page) en fragments (titre, sous-titre, corps du texte, pied de page) identifiés par des balises de code HTML. L'autre étape fondamentale consiste à discrétiser le contenu (le texte) en trouvant le moyen de faire voyager non pas seulement des signaux mais aussi des signes. Par exemple, pour un nombre, on prévient les navigateurs et les moteurs de recherche qu'il s'agit d'une taille, d'un poids, d'une date, etc. En se référant à la théorie opérationnelle de l'écriture numérique de Crozat *et al.* (2011), on peut dire qu'au lieu de se contenter de prendre en charge le « niveau techno-applicatif », en transformant les signaux binaires (une suite de 0 et 1) en nombre (mettons « 48 »), la machine prend également en charge le « niveau sémio-rhétorique » en attachant à ce nombre un sens (taille, poids, etc.). A la suite de ce processus, ce qui est manipulable correspond à ce qui est signifiant, la syntaxe ayant permis d'opérer une « *coïncidence entre le format technique et la forme sémiotique* » (Bouchardon *et al.*, 2011, p. 14). C'est ce qu'on appelle le Web sémantique. En plus de donner une structure à l'information, en discrétisant son contenant, on lui donne un sens, en discrétisant son contenu, puis on donne une structure aux unités de sens : une ontologie (Gruber, 1993).

Il existe plusieurs procédés de discrétisation sémantique. Chacun est normalisé de façon plus ou moins souple et transparente. Il y a différents codes (HTML et Javascript) et au sein d'un même code il peut y avoir différentes syntaxes utilisées pour transformer le flux d'information en unités signifiantes (RDFa, microdonnées et microformats pour le HTML, JSON-LD pour le Javascript), et différents formats de description ontologique utilisés pour structurer les relations entre ces unités (RDFS, OWL). Les conditions de production du sens – c'est-à-dire les conditions devant être remplies pour qu'il puisse y avoir « déprise et reprise » des unités discrètes (Bachimont, *op. cit.*, p. 29) – dépendent des spécificités de chacun de ces codes et de ces syntaxes, de leurs auteurs et de leurs modes de gouvernance. Les sciences de l'information et de la communication ont donc intérêt à investiguer la question de la concurrence et de la complémentarité entre ces codes, ces syntaxes et ces ontologies, et à analyser de quelles valeurs chacun d'entre eux est le véhicule et le garant. Cela car comme l'avait souligné Lawrence Lessig dans son célèbre article – *Code is Law* (1999) – les choix concernant le code sont des choix de valeurs. En comprenant les tenants des choix possibles en matière de Web sémantique, on comprendra non pas seulement les conditions de production du sens, du point de vue technique, mais vraiment ce que nous pourrions appeler « les politiques du sens ». On comprendra quels sont les intérêts qui prévalent, qui sont les acteurs qui dominent, et quels sont les effets concrets de cette prévalence et de cette domination en termes d'information et de communication. En outre, on comprendra où est *située* la sémantisation, en sondant non pas simplement le mécanisme de description et de traitement de l'information mais aussi la teneur des allers-retours entre l'homme qui décrit le contenu et la machine qui le traite.

Cadrage théorique et cadre méthodologique

Les syntaxes que nous étudierons ici sont des normes, en cela qu'elles se présentent toutes sous la forme de documents « *déterminant des spécifications techniques de biens, de services ou de processus qui ont vocation à être accessibles au public, [qui] résultent d'un choix collectif entre les parties intéressées à sa création, et [qui] servent de base pour la solution de problèmes répétitifs* » (Lelong et Mallard, 2000, p. 11). C'est pourquoi nous nous sommes placés dans la même perspective que ceux qui parmi nos collègues intéressés par la gouvernance de l'Internet étudient en particulier la normalisation des protocoles (De Nardis, 2009 ; Ermoshina et Musiani, 2018) en partant du principe qu'étudier comment une norme est construite et circule revient à étudier des rapports de pouvoir (Bowker et Star, 1999 ; Bush, 2011).

D'autre part, dans notre étude, l'information et la communication ne sont pas des variables explicatives, l'*explanans*, mais des variables expliquées : l'*explanandum*. Nous voulons comprendre pourquoi les syntaxes sont utilisées comme ça et pas autrement, et pourquoi la discrétisation de

l'information qu'elles opèrent a lieu de cette façon et pas d'une autre. C'est pourquoi nous nous sommes appuyés sur la théorie du support comme elle a été développée en particulier à l'Université Technologique de Compiègne par les chercheurs Bruno Bachimont, Serge Bouchardon, Isabelle Cailleau et Stéphane Crozat. En développant une théorie opérationnelle de l'écriture numérique, ceux-là se sont intéressés à la manière dont le numérique peut sortir de « l'autarcie » pour être « négocié dans le monde de la matière et du sens » (Bouchardon *et al.*, 2011, p. 13). Notre étude vise justement à comprendre de quoi est faite cette négociation dans le cas du web sémantique : comment, concrètement, les possibilités d'écriture sont négociées, en gardant à l'esprit que la compréhension de cette négociation éclairera les conditions *effectives* de production (information) et de mise en circulation (communication) du sens.

Pour comparer les organismes de normalisation, leurs prises d'intérêts et leurs procédures, nous avons commencé par identifier ceux qui tenaient les rênes des trois principaux organismes concernés par la standardisation des protocoles du Web sémantique (*World Wide Web Consortium*, *CommerceNet* et *Web Hypertext Application Technology Working Group*). Puis nous avons analysé comment au sein de ces organismes les décisions étaient prises. Enfin, nous avons comparé les modalités d'utilisation des différentes syntaxes du Web sémantique. Pour cela, nous avons nous-même produit du code. Cela nous a obligé à apprendre à maîtriser chacune des syntaxes de façon à être capable de comprendre ce que l'utilisation de l'une d'entre elles impliquait par rapport à l'utilisation d'une autre. Plus précisément, nous avons exprimé une même recette de cuisine dans les cinq principales syntaxes du Web sémantique (RDF, RDFa, Schema.org, hRecipe, JSON-LD). Cela nous a permis de représenter un objet dans différents langages de manière à pouvoir analyser et comparer ce que chacune de ces représentations faisait à/de l'objet (la recette de cuisine). Nous avons ainsi pu identifier non pas seulement les tenants de la concurrence entre syntaxes mais aussi les aboutissants sémiotiques de cette concurrence, *en les expérimentant*.

Nous nous sommes concentrés sur les syntaxes qui opèrent la discrétisation en structurant le contenu du document, et permettent de constituer des bases de données qui ensuite sont structurées grâce à des ontologies. Nous n'avons donc pas analysé les formats de description ontologique (dont les deux principales sont RDFS et sa spécification OWL), lesquels agissent une fois seulement que l'information a été discrétisée, et qui sont donc plus éloignées des créateurs de pages Web qui n'y ont jamais recours eux-mêmes. À l'inverse, tous les individus qui publient des informations sur le Web et décident d'opérer un balisage sémantique doivent obligatoirement choisir entre les différentes options que nous décrirons ci-après.

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 le souhaitent. Il leur suffit pour devenir membre de payer une somme pouvant aller d'un montant très modique pour une association, jusqu'à 59 000 euros par an pour une grande entreprise. Ils étaient au total 463 membres en septembre 2017. C'est au W3C que sont édités les trois standards mis au point par Tim Berners-Lee au début des années 1990 : l'*Uniform Resource Locator* (URL) pour localiser les données, l'*HyperText Transfer Protocol* (HTTP) pour les transférer et l'*HyperText Markup Language* (HTML) pour les décrire. Chaque proposition de modification ou de création d'un standard fait l'objet d'une discussion entre les membres du W3C ayant émis le souhait de participer au groupe de travail concerné. Elle donne ensuite lieu à un document censé franchir quatre étapes à l'issue desquelles, s'il les a toutes franchies, le document en question obtient le statut de norme officielle. C'est durant ces étapes que se joue ce qu'en sociologie des techniques on nomme la « stabilisation » (Bijker, 1995) et la « structuration » des relations technologiques (Benezech, 1996). Le franchissement de chacune d'elles se fait par consensus, sachant que c'est Tim Berners-Lee, ou une personne qu'il aura expressément mandatée pour ce faire, 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 de revenir à une étape précédente, ou tout simplement d'abandonner le projet. Le processus du W3C confère un pouvoir considérable à Tim Berners-Lee, qui peut prendre en dernière instance des décisions contraires à 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 (Halpin, 2017 ; Sire, 2017). Pour cette raison, certains observateurs n'hésitent pas à qualifier le fondateur du Web de « dictateur à perpétuité » (Malcom, 2008). Il n'empêche, le W3C est une arène dont la vocation est d'être ouverte, transparente, de garantir l'efficacité technique des normes, d'assurer l'interopérabilité des infrastructures (Virili, 2003), et d'empêcher, grâce à une politique de brevets originale, que des entreprises participent à la discussion dans le seul but de se rendre indispensables par la suite en tirant parti des brevets qu'elles détiennent (Russel, 2003 ; Halpin, 2017).

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 développée la syntaxe RDF, basé sur le code XML, et ayant atteint le statut de standard officiel en 2004 en version 1.1. Si on observe cette syntaxe au prisme de la théorie opérationnelle de l'écriture, on y retrouve le tropisme de « manipulabilité » : il est possible d'appliquer aux données discrétisées des règles formelles de traitement algorithmique. On retrouve également le principe de « fragmentation-agrégation » : le contenu est fragmenté en unités potentiellement signifiantes qu'on pourra éventuellement agréger différemment (Crozat *et al.*, 2011, p. 21-22).

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 pour la phrase : « *Octave a cuisiné un gâteau au chocolat pour Romane* », nous aurons les trois prédicats suivants :

- [Octave {est l'auteur du} gâteau au chocolat]
- [Romane {est le propriétaire du} gâteau au chocolat]
- [Le gâteau {est au} chocolat]

Nous pouvons créer d'autres relations grâce à des ontologies : si Octave est l'auteur du gâteau au chocolat et Romane son propriétaire, alors : [Octave {est un fournisseur de} Romane]. Il s'agira d'activer cette fois ce que la théorie opérationnelle de l'écriture nomme « la fonction *génération* » : on pourra créer « *des contenus automatiquement à partir des contenus existants* » (Crozat *et al.*, 2011, p. 22). 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 « fragment » (personnes, choses, concepts, dates, lieux...) un seul *Uniform Resource Identifier* (URI), dans une base de données où sont renseignés les triplets (qu'on appelle « dépôt » ou « banque »), il devient possible de créer des ontologies à l'échelle du Web. Après avoir transformé les informations continues en données discrètes (*manipulabilité*), on met de l'ordre non plus entre les contenants (i.e. en hiérarchisant les sites et les pages) mais entre les contenus (*fragmentation-agrégation*), c'est-à-dire entre les données elles-mêmes, archivées dans différentes ontologies interconnectées, à partir desquelles on pourra générer de nouveaux contenus (*génération*).

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 générer eux-mêmes des ontologies, ou bien se référer à une ontologie existante comme DBpédia (collection de triplets constitués à partir des informations contenues dans Wikipedia). C'est un cas typique du niveau « techno-applicatif » : il s'agit de proposer « un formatage qui prescrit *a priori* des signes et leur manipulation formelle » (Crozat *et al.*, *op. cit.*, p. 17). Il existe plusieurs schémas ou « espaces de noms » utilisés en RDF et correspondant à chaque fois à des ontologies spécifiques, par exemple « Dublin Core » et « Friend Of A Friend » qui peuvent être utilisés en plus de l'espace de nom RDF (RDFS). Certains de ces schémas, comme le Dublin Core, existait avant le RDF. Ainsi, le RDF permet de faire le lien avec des techniques de discrétisation préexistantes. Parce que plusieurs schémas peuvent être utilisés au sein d'un même document, il est indispensable de renseigner les préfixes à chaque fois dans les balises XML tout en ayant annoncé en tête du document 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 « *xml name space* »). Ci-dessous, nous avons écrit une recette de gâteau au chocolat, puis nous avons ensuite discrétisée en code XHTML (XML + HTML), grâce à la syntaxe RDF :

Version brute :

Recette du gâteau au chocolat d'Octave

Dessert pour 4 personnes / 15 minutes de préparation / 1er février 2018

Ingrédients :

200 g de chocolat noir à 70% de cacao

125 g de beurre demi-sel

100 g de farine

1 sachet de levure chimique (10 g)

4 oeufs

200 g de sucre en poudre

1 pincée de sel

Préparation :

Coupez le chocolat et le beurre en morceaux. Faites fondre au bain-marie. Retirez du feu.

Ajoutez la farine et la levure tamisées dans le chocolat fondu.

...

Version RDF :

```

1. <html xmlns=http://www.w3.org/1999/xhtml
2.   xmlns:foaf="http://xmlns.com/foaf/0.1/"
3.   xmlns:re="http://octave.com/recettes/gateaux">
4. <head>
5. <title>Recette du gâteau au chocolat d'Octave</title>
6. </head>
7. <body>
8. <h1> Recette du gâteau au chocolat d'Octave </h1>
9. <re:recipe>
10. <re:recipe_head>
11. <foaf:name>Octave</foaf:name>
12. <re:meal_type>Dessert</re:meal_type>
13. </re:recipe_head>
14. <re:recipe_body>
15. <re:ingredients>
16. <re:ingredient> 200 g de chocolat noir à 70% de cacao </re:ingredient>
17. <re:ingredient> 125 g de beurre demi-sel</re:ingredient>
18. <re:ingredient> 100 g de farine</re:ingredient>
19. <re:ingredient> 1 sachet de levure chimique (10 g)</re:ingredient>
20. <re:ingredient> 4 oeufs </re:ingredient>
21. <re:ingredient> 200 g de sucre en poudre</re:ingredient>
22. <re:ingredient> 1 pincée de sel</re:ingredient>
23. <re:directions>
24. <re:direction>Coupez le chocolat et le beurre en morceaux. Faites-les fondre en bain-marie.
    Retirez du feu. </re:direction>
25. <re:direction>Ajoutez la farine et la levure tamisées dans le chocolat fondu.</re:direction>
26. ...
27. </re:directions>

```

```

28. </re:recipe_body>
29. <re:recipe_footer>
30. <re:serving>4</re:serving>
31. <re:preparation_time>15 minutes</re:preparation_time>
32. </re:recipe_footer>
33. <re:document_info>
34. <re:document_author>Octave</re:document_author>
35. <re:date_updated>01/08/2018</re:date_updated>
36. </re:document_info>
37. </re:recipe>
38. </body>
39. </html>

```

Nous avons fait référence à deux espaces de noms (lignes 2 et 3). Le premier : *Friend of a Friend* (<http://xmlns.com/foaf/0.1/>) auquel nous avons donné le préfixe « foaf ». Et un deuxième que nous avons créé à titre d'exemple (<http://octave.com/recettes/gateaux>) auquel nous avons donné le préfixe « re » comme « recette ». Les balises spécifiques à chaque espace de nom sont assorties du préfixe correspondant, de telle sorte que les navigateurs et les moteurs de recherche puissent savoir que c'est bien à cet espace de nom, et aux ontologies qui lui sont propres, que la balise renvoie. Le principe de *fragmentation*, attaché au tropisme de *manipulabilité*, donne ainsi lieu à des possibilités de *génération* d'autant plus nombreuses que plusieurs espaces de noms pourront être invoqués.

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 au lieu de s'inscrire dans le nom des balises. Cela simplifie considérablement l'écriture, d'une part, et, d'autre part, cela permet de mieux normaliser l'usage des balises et de faire varier seulement les attributs à l'intérieur des balises. On augmente ici la *manipulabilité*, de sorte que les développeurs pourront opérer une *fragmentation* plus fine, ce qui augmentera l'efficacité des dispositifs visant à *générer* de nouveaux types de contenus à partir d'ontologies. 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), mais les balises restent les mêmes quel que soit l'espace de nom mobilisé. La fragmentation, autrement dit, se joue à un niveau inférieur de la syntaxe : ce ne sont plus les éléments syntaxiques qui varient, mais les attributs d'éléments qui sont les mêmes dans tous les cas, ce qui permet d'optimiser le paramétrage des logiciels chargés de comprendre et de traduire les informations ainsi transformées en données (navigateurs, moteurs de recherche). Pour notre recette, cela donne :

```

1. <p xmlns:re=http://octave.com/recettes/gateaux
  xmlns:foaf="http://xmlns.com/foaf/0.1/">
2. <head>
3. <title>Recette du gâteau au chocolat d'Octave</title>
4. </head>
5. <body>
6. <h1> Recette du gâteau au chocolat d'Octave </h1>
7. <span property="re:recipe_author" about="foaf:name">Octave</span>
8. <span property="re:meal_type">Dessert</span>
9. <span property="re:ingredient"> 200 g de chocolat noir à 70% de cacao </span>
10. <span property="re:ingredient"> 125 g de beurre demi-sel</span>
11. ...

```

Les balises sont des balises HTML classiques mais les « xmlns » renseignés à la première ligne donnent 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, et à une ontologie, spécifiques (cf. ligne 7). La discrétisation est plurielle : une même unité de sens renvoie à plusieurs modes de représentation.

Le RDF et le RDFa ne sont pas les seules syntaxes permettant d'opérer la discrétisation. Il y en a deux autres, les microformats et les microdonnées, qui chacune ont à la fois leurs propres modes de gouvernance, et leurs propres principes de fragmentation-agrégation, ayant eux-mêmes des effets sur les possibilités de génération de contenus (i.e. les moteurs de recherche n'effectueront pas les recoupements ontologiques *dans les mêmes conditions*), et, donc, sur le niveau « sémio-rhétorique ».

CommerceNet et Microformats

Les microformats (auxquels renvoient les sigles μ F ou uF) ont été développés sans politique de normalisation arrêtée, de façon participative. Cela les différencie des syntaxes RDF/RDFa normalisées selon la procédure du W3C. Le fait que les microformats n'aient jamais été « stabilisés » les apparente à ce que Ksenia Ermoshina et Francesca Musiani nomment un « quasi-standard » (Ermoshina et Musiani, 2018). N'importe qui peut en théorie proposer un nouveau type de microformat en créant son propre espace de noms et sa propre syntaxe. Il faut pour cela suivre une procédure qui consiste essentiellement à ouvrir une discussion sur le wiki de CommerceNet. Sur ce même site, il est dit qu'avant de proposer un nouveau type de microformat il faut implémenter sur son site les microformats qui existent déjà, comme un gage de bonne foi, et publier un « témoignage ». Mais toutes les propositions ne seront pas implémentées, et seules celles qui créeront autour d'elles un consensus sans que personne n'ait le pouvoir de *décréter* que c'est le cas, atteindront le statut de « quasi-standard ».

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 ». La documentation officielle prévient que les microformats sont « conçus d'abord pour les humains, ensuite pour les machines » (Khare et Çelik, 2006). L'accent est mis tout au long de la procédure informelle pour que les microformats soient lisibles pour des êtres humains, c'est-à-dire suffisamment proches du langage naturel pour être compréhensibles y compris en affichant le code source d'une page pour la lire à l'œil nu. Contrairement au RDF/RDFa, et en théorie (nous verrons qu'en pratique c'est différent), les microformats se situent au niveau sémio-rhétorique plutôt qu'au niveau techno-applicatif : ils ont vocation selon leurs concepteurs à créer du sens en interagissant avec des humains, et, dans un second temps seulement, à donner lieu à des traitements logiciels, en interagissant avec des machines.

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 grâce aux balises `<div>` ou ``. Tous les microformats existants sont répertoriés sur le wiki (http://microformats.org/wiki/Main_Page). 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 du document et de suivre les instructions du wiki pour renseigner les balises situées en aval. Par exemple, pour le microformat « hrecipe », nous avons :

1. `<div class="hrecipe">`
2. `<h1 class="fn"> Recette du gâteau au chocolat d'Octave</h1>`
3. `Contributed by Octave`
4. `<h2>Ingredients</h2>`
5. ``
6. `<li class="ingredient">`
7. `> 200 g de chocolat noir à 70% de cacao `
8. `...`
9. `<h2>Instructions</h2>`
10. `<ul class="instructions">`
11. ` Coupez le chocolat et le beurre en morceaux. Faites-les fondre en bain-marie. Retirez du feu. `
12. `...`

Le nom du microformat (`hrecipe`, ligne 1) 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. La fragmentation n'induit pas la même manipulabilité, et ne suppose pas de générer du contenu aussi aisément qu'avec RDF/RDFa. Si le logiciel ne connaît pas déjà la syntaxe dont il est question (i.e. si son concepteur ne l'a pas paramétré en fonction d'une syntaxe donnée *a priori*), il n'aura aucun moyen de « comprendre » de quoi il est question (niveau sémio-rhétorique) ou de « générer » de nouveaux contenus (niveau techno-applicatif). Ainsi les microformats ne permettent-ils pas de faire le lien avec d'autres vocabulaires (*dublin core*, *friend of a friend*, *rdf*), pas plus qu'ils ne permettent de faire le lien avec les ontologies constituées sur la base de ces vocabulaires. Il y a discrétisation, certes, mais il faut que le logiciel effectue certaines opérations sur ces données pour pouvoir faire le lien avec d'autres données situées ailleurs sur le Web.

Les microformats permettent aussi d'éditorialiser l'affichage dans les listes de moteurs de recherche. C'est le cas notamment depuis que Google a proposé en 2009 d'utiliser les microformats pour structurer l'apparence des recettes de cuisine dans les listes de résultats. A la suite de cette annonce et de sa mise en œuvre, Google a créé « *Recipe View* » en 2011, un module spécifique permettant notamment aux usagers du moteur de spécifier les ingrédients dont ils disposent pour savoir quel plat il leur est possible de cuisiner. De ce point de vue, les microformats auraient moins vocation à opérer une sémantisation qu'à éditorialiser l'affichage sur les moteurs de recherche. L'intention de ceux qui s'en saisissent serait alors moins située au niveau sémio-rhétorique qu'au niveau techno-applicatif, et ce alors même que, nous l'avons vu, l'intention de ceux qui ont conçu les microformats, elle, est au contraire davantage située au niveau sémio-rhétorique (« il faut pour voir être compris par des êtres humains...), qu'au niveau techno-applicatif (... et éventuellement pouvoir être opéré par des machines calculatoires »). On peut émettre l'hypothèse forte selon laquelle cette ambivalence est peut-être due au fait que les microformats sont implémentés par des entreprises de e-commerce essentiellement, qui ont besoin d'outils pour mettre en valeur leurs produits dans les listes de résultats des moteurs de recherche, mais qui n'ont pas forcément envie que des ontologies soient

constituées qui permettraient de comparer les prix et les caractéristiques des produits à vendre à grande échelle.

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 préférerait se concentrer sur le XHTML et sur l'objectif de sémantiser le Web (Sire, 2017). 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 de normaliser une version du code HTML en dehors du W3C, ce qui était possible dès lors que les premiers soutiens 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 (Sire, 2017).

Finalement, en 2007, le XHTML 2.0 était l'objet de plusieurs dysfonctionnements techniques, en conséquence de quoi Tim Berners-Lee 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.

Concernant la procédure de normalisation, n'importe qui peut théoriquement participer aux discussions du Whatwg. Cependant le dernier mot appartient au « Steering Group » composé en 2018 d'un représentant pour chacune de ces quatre entreprises : Mozilla, Microsoft, Google et Apple. D'autre part, alors qu'elle n'en avait pas jusque-là, le Whatwg a mis en place en janvier 2018 une politique de brevet similaire à celle du W3C, basée sur la gratuité et sur l'engagement pour chaque contributeur à ne pas tirer parti financièrement d'un brevet nécessaire à l'implémentation du standard (Russel, 2003).

C'est dans le cadre du Whatwg qu'ont été développées les microdonnées. Il est important ici de bien considérer que ce sont les propriétaires de navigateurs et de moteurs de recherche, qui 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. 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> ». En revanche, on ne peut pas renvoyer à d'autres vocabulaires qu'à ceux du Schema.org. Il y a fixation des balises et du vocabulaire, les variations ne concernant cette fois que les attributs, alors qu'elles concernaient la syntaxe et le vocabulaire dans le cas du RDF, et les attributs et le vocabulaire dans le cas du RDFa et des microformats. Autrement dit, le Schema.org, basé sur le tropisme de manipulabilité, limite le principe de fragmentation-agrégation pour mieux contrôler la fonction de traitement automatique et de génération des contenus.

```

1. <div itemscope itemtype="http://schema.org/Recipe">
2.   <span itemprop="name">Recette du gâteau au chocolat d'Octave</span>
3.   By <span itemprop="author">Octave</span>,
4.   Prep Time: <meta itemprop="prepTime" content="PT15M">15 minutes
5.   Ingredients:
6.   - <span itemprop="recipeIngredient">125 g de beurre demi-sel</span>
7.   - <span itemprop="recipeIngredient">100 g de farine</span>
8.   - <span itemprop="recipeIngredient">1 sachet de levure chimique (10 g)</span>
9.   ...
10.  Instructions:
11.  <span itemprop="recipeInstructions">
12.  Couper le chocolat et le beurre en morceaux
13.  </span>
14.  ...
15. </div>

```

On voit ici que la discrétisation avec des balises qui dans presque tous les cas (sauf ligne 4) ont recours à la même balise () et au même attribut (itemprop), pour faire recours à un schéma exclusif, sur lequel aucune ontologie publique ne repose, et qu'il faudra donc « traduire » en RDF si l'on veut faire le lien avec les banques de triplets du type DBpédia (<http://schema.org/docs/datamodel.html>). Les microdonnées constituent donc un dispositif rendant plus simple la discrétisation au niveau techno-applicatif (peu de balises, peu d'attributs, un schéma très facile à comprendre et à utiliser), mais rendant plus compliquée la sémantisation au niveau sémio-rhétorique (on ne peut pas créer d'ontologie sur la seule base des unités de sens exprimées sous forme de microdonnées).

En 2011, après l'officialisation des microdonnées, Tantek Celik, 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. ». Le développeur (très influent) Mark Pilgrim a quant à lui prévenu que le Schema.org était l'illustration de l'échec du W3C avec son RDF/RDFa. Manu Sporny, qui dirigeait le groupe de travail du W3C autour de la spécification RDF, a pointé du doigt le fait que les microdonnées étaient le fruit d'un petit groupe d'organisations alors que 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. 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 concurrentes.

En plus d'être le seul vocabulaire à être utilisable avec la syntaxe des microdonnées, le « shema.org » peut être utilisé comme vocabulaire avec les syntaxes RDF et RDFa. Il permet ainsi de faire le lien avec les ontologies appuyées sur ces syntaxes. Finalement, du côté syntaxe (au niveau techno-applicatif) les microdonnées concurrencent les microformats, et du côté vocabulaire (au niveau sémio-rhétorique) le shema.org concurrence les vocabulaires utilisés en RDF/RDFa.

Le Whatwg a également proposé d'intégrer son vocabulaire à une syntaxe pour Javascript standardisée par le W3C : le « *JavaScript Object Notation for Linked Data* » (JSON-LD). Le principe cette fois n'est plus de discrétiser directement l'information présente dans le document mais *de la recopier en la discrétisant* dans un espace réservé aux logiciels, en la traduisant dans un code de programmation (le Javascript) moins proche de l'écriture que ne l'est le HTML (qui est un code de description). Tout en restant dans le niveau techno-applicatif, on remonte ainsi vers le niveau théorico-idéal, « celui du code binaire défini dans son double isolement interprétatif et matériel » (Croizat *et al.*, 2011, p. 16), et, donc, en s'éloignant du niveau sémio-rhétorique. Le JSON-LD se

différencie du RDF et des microformats, où l'être humain passait avant la machine, et où la syntaxe de la discrétisation était fondue dans la syntaxe de la langue naturelle, comme une espèce de ponctuation sémantique. Il aboutit à la création d'un espace « plus abstrait » que seule la machine a vocation à consulter, tandis que les internautes continueront à consulter sur leurs navigateurs l'information contenue dans un fichier débarrassé des indications sémantiques, et qui, si on le recopie et qu'on le transporte ailleurs, « voyagera » sans sémantisation.

```

1. <script type="application/ld+json">
2. {
3.   "@context": "http://schema.org",
4.   "@type": "Recipe",
5.   "author": "Octave",
6.   "cookTime": "PT15M",
7.   "recipeIngredient": [
8.     "1 sachet de levure chimique (10 g)",
9.     ...
10.  "recipeInstructions": "Couper le chocolat et le beurre en morceaux",
11.  ...
12. </script>

```

L'usage du JSON-LD est recommandé par Google, car il est idéal pour le fonctionnement de son moteur de recherche. Il permet notamment à l'entreprise de créer sa propre ontologie, le *Knowledge Graph*, à partir des dépôts de triplets existants et des données que le moteur récolte lui-même, et de l'exploiter dans les résultats de son moteur de recherche. Mais cette ontologie, même si n'importe quel internaute peut l'interroger via le moteur, n'est pas accessible en tant que telle, ni réutilisable par une autre technologie. En devenant plus manipulables pour les machines, les données le sont moins, ou plus du tout, pour les êtres humains. Surtout, elles sont stockées dans un autre espace que celui où se trouvent les informations destinées aux humains. Le niveau techno-applicatif se dédouble : pour un même fond, on trouve d'un côté une forme adressée aux logiciels dont la vocation est d'agrèger les contenus et d'en générer de nouveaux (les moteurs de recherche), et de l'autre une forme adressée aux logiciels dont la vocation est de structurer l'affichage du contenu sur l'écran de l'utilisateur (les navigateurs). Ainsi, Google, après avoir proposé, et imposé, un nouveau vocabulaire, le Schema.org, et une nouvelle syntaxe pour le code HTML, les microdonnées, a proposé de traduire dans un autre code cette syntaxe et ce vocabulaire, pour mieux fluidifier la relation entre les niveaux théorético-idéal et techno-applicatif, quitte à opérer une rupture avec le niveau sémiotique.

Conclusion : l'industrialisation du sens

Nous avons vu comment la discrétisation des informations présentes sur le Web, c'est-à-dire leur transformation en *Big Data*, pouvait donner lieu à trois manières très différentes de tisser des liens entre « les trois niveaux du numérique » : théorético-idéal, techno-applicatif, sémio-rhétorique (Crozat *et al.*, 2011). Il y a là trois « politiques du sens ». En comparant les procédures de normalisation, nous avons montré que dans le cas des microdonnées et du RDF et du RDFa, même si la discussion est ouverte à tous, certains acteurs ont le pouvoir de validation en dernière instance (Microsoft, Mozilla, Google et Apple pour les microdonnées, et Tim Berners-Lee pour le RDF et le RDFa). En comparant les syntaxes, nous avons également montré que tandis que certaines d'entre elles (RDF/RDFa) servent à sémantiser les informations après les avoir discrétisées, d'autres (les microformats) sont au service d'intérêts commerciaux visant à éditorialiser l'affichage des données sans forcément permettre (voire en empêchant) de relier les données entre elle. La dernière syntaxe

analysée (les microdonnées du Schema.org) vise à imposer un vocabulaire en proposant une syntaxe, et d'influencer les autres syntaxes en proposant un vocabulaire. Elle résulte autrement dit d'une stratégie prédatrice puisque du côté syntaxe (au niveau techno-applicatif) elle concurrence les microformats, et du côté vocabulaire (au niveau sémio-rhétorique) elle concurrence tous les autres vocabulaires existants. Parmi les créateurs de cette syntaxe se trouve Google, qui grâce à cette stratégie de prédation a pu relier les données à l'échelle du Web mais qui, étant donné ses intérêts financiers, au lieu de publier l'ontologie résultant de cette réunification, s'est contenté de proposer un outil d'exploration : le Knowledge Graph. Enfin, nous avons vu que tandis que toutes les syntaxes du code HTML suivent la langue naturelle, en balisant les mots et les expressions présents sur les pages Web, dans le corps (<body>) du document, l'usage du code Javascript consiste quant à lui à dissocier les données des informations, c'est-à-dire à dissocier l'espace consacré à la machine (techno-applicatif) de l'espace consacré au lecteur (sémio-rhétorique). Ces deux espaces deviennent détachables : le document, s'il voyage (par exemple si le contenu d'une page Web est recopié sur une autre page Web) ne voyagera plus avec ses balises sémantiques, ce qui obligera les acteurs intermédiaires à opérer une nouvelle opération de discrétisation/sémantisation.

Nous avons vu que dans le cas du RDF/RDFa le principe de *fragmentation*, propre au tropisme de *manipulabilité* donnait lieu à des possibilités de *génération* de nouveaux contenus d'autant plus nombreuses que plusieurs espaces de noms pouvaient être invoqués, alors que dans le cas du Schema.org un seul espace de nom est mis au service d'une fonction de fragmentation qui dès lors est plus simple, mais moins souple, et qui offre une plus grande manipulabilité au niveau techno-applicatif mais entraîne une diminution des possibles au niveau sémio-rhétorique. Quant aux microformats, nous avons vu que l'intention de ceux qui s'en saisissent est moins située au niveau sémio-rhétorique qu'au niveau techno-applicatif, alors même que l'intention de ceux qui les ont conçus est au contraire davantage située au niveau sémio-rhétorique qu'au niveau techno-applicatif. Nous pouvons émettre l'hypothèse qu'une telle ambivalence est liée au statut de « quasi-standard » des microformats.

En étudiant la syntaxe de ce qu'il nomme les « petites formes » générées par les API sous forme de widgets de code HTML préécrits, Samuel Goyet a récemment montré comment la discrétisation était mise « *au service d'une plus grande circulation des formes du texte* » (Goyet, 2017, p. 70) et comment elle pouvait faire l'objet d'une « optimisation » visant « standardiser » et à « industrialiser » la circulation des textes de réseau (*ibid.*, p. 90). En étudiant nous-mêmes la discrétisation, dans le cas du Web sémantique, nous avons montré que cette industrialisation pouvait concerner non pas seulement la forme du texte mais aussi son sens, ou plutôt son « faire sens ». Nous avons également montré que des stratégies de prédation pouvaient être mises en œuvre, étant donné notamment les enjeux économiques de ce « faire sens ». Il n'y a pas d'industrialisation sans standardisation, et il n'y a pas standardisation sans rapports de forces (Bush, 2011). Des visions s'affrontent, des projets, des intérêts. Une tectonique a lieu, dont dépendent les modalités concrètes de production et de mise en circulation à grande échelle des signes, c'est-à-dire des informations transformées en données qualifiables, puis qualifiées. Cette tectonique produira l'ordre ou le chaos, c'est selon, les sciences de l'information et de la communication étant à coup sûr les mieux placées pour comprendre à quoi et à qui tient ce « selon ».

Références bibliographiques

Bachimont, Bruno (1999), « De l'hypertexte à l'hypotexte : les parcours de la mémoire documentaire », *Technologies, Idéologies, Pratiques*, n° 4, 195-225.

Bachimont, Bruno (2010), *Le sens de la technique : le numérique et le calcul*, Paris : Les belles lettres, (collection « encre marine »).

- Benezech, Danièle (1996), « La norme : une convention structurant les interactions technologiques et industrielles », *Revue d'Economie Industrielle*, n° 75, p. 27-44.
- Bijker, Wiebe E. (1995), *Of Bicycles, Bakelites and Bulbs: Steps towards a Theory of Socio-technical Change*, Cambridge, MA : MIT Press.
- Bouchardon Serge, Cailleau Isabelle, Crozat Stéphane, Bachimont Bruno, Hulin Thibaud (2011), « Explorer les possibles de l'écriture multimédia », in Paquienséguy Françoise (coord.), Dossier « Information publique : stratégies de production, dispositifs de diffusion et usages sociaux », *Les Enjeux de l'Information et de la Communication*, n° 12/2, 2011, p. 11-23, <https://lesenjeux.univ-grenoble-alpes.fr/pageshtml/art2011.html#dossier>
- Bowker, Geoffrey C., Star, Suzan Leigh (1999), *Sorting Things Out: Classification and Its Consequences*, Cambridge, MA: MIT Press.
- Busch, Lawrence (2011), *Standards: Recipe for Reality*, Cambridge, MA: MIT Press.
- DeNardis, Laura (2009), *Protocol Politics: The Globalization of Internet Governance*, Cambridge, MA: MIT Press.
- Crozat Stéphane, Bachimont Bruno, Cailleau Isabelle, Bouchardon Serge, Gaillard Ludovic (2011), « Éléments pour une théorie opérationnelle de l'écriture numérique », *Document numérique*, vol. 14, n°3, p. 9-33.
- Ermoshina, Ksenia, Musiani, Francesca (2018), « Standardizing by Running Code: The Signal Protocol and De Facto Standardization in End-to-End Encrypted Messaging », *GigaNet: Global Internet Governance Academic Network, Symposium annuel 2017*.
- Glen C. M. (2017), *Controlling Cyberspace. The politics of Internet Governance and Regulation*, Praeger.
- Goyet, Samuel (2017), « Outils d'écriture du Web et industrie du texte: Du code informatique comme pratique lettrée », *Réseaux*, vol. 206, n° 6, p. 61-94.
- Gruber, Thomas (1993), « A Translation Approach to Portable Ontology Specification », *Knowledge Acquisition*, vol. 5, p. 199-220.
- Halpin, Harry (2017), « The Crisis of Standardizing DRM: The Case of W3C Encrypted Media Extensions », *SPACE 2017 - Seventh International Conference on Security, Privacy, and Applied Cryptography Engineering*, Décembre 2017, Goa, India, Springer, 10662, p. 10-29.
- Khare, Rohit, Çelik, Tantek (2006), « Microformats: a pragmatic path to the semantic Web », *Proceedings of the 15th international conference on World Wide Web (WWW '06)*, ACM, New York, NY, p. 865-866.
- Lelong Benoît, Mallard Alexandre (2000), « Présentation : la fabrication des normes », *Réseaux*, vol. 18, n°102, p. 9-34.
- Lessig Lawrence (1999), *Code and Other Laws of Cyberspace*, New York, NY : Basic Books.
- Malcolm Jeremy (2008), *Multi-Stakeholder governance and the Internet governance forum*, Australia: Terminus press.
- Russel, Andrew L. (2003), « The W3C and its Patent Policy Controversy: A Case Study of Authority and Legitimacy in Internet Governance », *TPRC 2003*.
- Sire Guillaume (2017), « Gouverner le HTML. Analyse du processus de normalisation du code HTML5 et de la controverse "Encrypted Media Extensions" », *Réseaux*, vol. 206, no. 6, p. 37-60.
- Virili Francesco (2003), « Design, Sense-Making And Negotiation Activities In The " Web Services" Standardization Process », *MIS Quarterly Special Issue on Standard Making: A Critical Research Frontier for Information Systems*