Remerciements

Je tiens à exprimer ici ma reconnaissance envers Françoise Banat-Berger et Emmanuelle Bermès pour m'avoir soutenu et orienté dans cette étude.

Je remercie également Céline Guyon, sans qui ce projet n'aurait pas été possible.

Enfin, ma gratitude va à Michel Jacobson, Alexandre Monnin, Stéphane Pouyllau et Romain Wenz qui ont répondu à mes (nombreuses) questions.

Résumé

Cette étude interroge la rencontre entre les technologies du web sémantique et les collections numériques dans le web de données. Il s'agit d'abord de comprendre comment des contenus documentaires peuvent devenir signifiants pour les machines, en analysant le rôle des formalismes et de la représentation des connaissances par les vocabulaires contrôlés. Nous nous demandons ensuite dans quelle mesure la logique "rhizomatique" du web de données et les ontologies sont compatibles avec le fonctionnement des institutions patrimoniales et des communautés du web 2.0.

Abstract

The main goal of this document is to study the interplay of Semantic Web technologies and digital collections in the "web of data". First we shall try to understand how computers are able to make out information contents. To do so, we shall have to analyse the role of formalisms along with the use of controlled vocabularies for knowledge representation(s). Then we shall try to decide to what extent the "rhizomatic" logic of the web of data and ontologies can work together with cultural heritage institutions and web 2.0 communities.

Les mutations des collections numériques à l'heure du web de données

Vincent Ventresque

2 septembre 2013

Mémoire de deuxième année du master Archives Numériques à l'ENSSIB, dirigé par Françoise Banat-Berger et Emmanuelle Bermès.

Table des matières

1 Introduction : comprendre ou manipuler ?

> – Et penser qu'après six mille et quelques années d'une lacune aussi préjudiciable que celle de mon Phonographe, reprit-il, quantité de lazzis, émanés de l'indifférence humaine, ont salué l'apparition de mon premier essai !... « Jouet d'enfant ! » grommelait la foule. (...) Seulement pour satisfaire mes semblables, je sens bien qu'il faut que j'invente un instrument qui répète avant même qu'on ait parlé, – ou qui, si l'expérimentateur lui souffle : « Bonjour, monsieur ! » réponde : « Merci, comment vous portez-vous ? » Ou qui, s'il arrive qu'un oisif éternue dans l'auditoire, lui crie : « À vos souhaits ! » ou : « Dieu vous bénisse ! » etc.
>
> Villiers de L'Isle Adam, *L'Ève future*

> Il est tout à fait certain que les opérations de l'Automate sont réglées par l'esprit, et non par autre chose. On peut même dire que cette confirmation est susceptible d'une démonstration mathématique, a priori. La seule chose en question est donc la manière dont se produit l'intervention humaine.
>
> E. A. Poë, *Le joueur d'échecs de Maelzel*

Le web de données, nouvel avatar ou première étape du projet de web sémantique, peut laisser sceptique ou rêveur, tant son ambition semble démesurée. La vidéo de présentation proposée par le site de la bibliothèque numérique européenne [1] , qui nous explique en moins de quatre minutes comment les technologies du web de données vont révolutionner la recherche et le partage de connaissances, évoque en effet un service web qui pourrait aider l'internaute dans sa recherche d'information, « d'abord en désambiguïsant [sa] requête, puis en connectant toutes les informations pertinentes, mises à jour dynamiquement, dans un même espace ? ». Supposons que, pour trouver "le tableau de Vénus avec le coquillage", un internaute lance une requête sur le terme "Vénus" dans un moteur de recherche classique : des millions de pages lui seront proposées, par ordre de

1. Europeana : *What is linked data* [Europeana – LOD, a]. L'expression *linked data* désigne à la fois des ensembles de données liées et les technologies permettant de les interconnecter. Voir également le prototype de moteur de recherche [Europeana – LOD, b].

popularité, traitant aussi bien de la joueuse de tennis Venus Williams que de la déesse, de la planète, Vénus de Milo, d'un institut de beauté... Supposons maintenant que le moteur lui réponde : « voulez-vous parler de la planète, de la déesse, ou de la joueuse de tennis ? ». Eh bien, « *It's possible !* » nous dit la vidéo : non seulement le moteur d'Europeana met à disposition toute la culture européenne, mais, bien plus, il nous guide en ne présentant que des résultats pertinents, organisés et fiables – données produites par les institutions culturelles. Tout se passe comme si la machine était capable de dialoguer avec l'utilisateur, et d'évaluer la qualité des documents qu'elle lui renvoie, comme si elle *comprenait* le sens de la question et des informations contenues dans les pages web.

Même impression d'avoir affaire à un scénario de science-fiction à la lecture de l'article fondateur du projet de web sémantique [2] où l'on voit des machines jouer le rôle de secrétaires : les « agents sémantiques » auraient la capacité de d'organiser de manière autonome un rendez-vous médical en tenant compte des contraintes du traitement suivi, de la proximité de la clinique et de son classement dans les évaluations fiables, en coopérant pour harmoniser les plannings de la famille du patient en fonction des disponibilités du médecin... En particulier, la phrase suivante peut nous étonner :

> L'agent de Lucy, qui a une *confiance totale* [3] dans l'agent de Pete dans le contexte particulier de cette tâche lui a apporté automatiquement de l'aide en lui fournissant des codes d'accès et des raccourcis à partir des données qu'il avait déjà triées.

Comment une machine pourrait-elle en effet « faire confiance » à une autre machine ? La vision peut inquiéter ou faire sourire. Comme nous le verrons, l'ambition du web de données est bien, pourtant, de construire une infrastructure technologique telle que nous puissions accéder à des informations dignes de confi-

2. Tim Berners-Lee, James Hendler and Ora Lassila : voir extrait en annexe → 6.1.

3. Les italiques sont dans le texte d'origine et indiquent des mots-clés pris en compte par les machines.

ance, automatiquement sélectionnées par les ordinateurs. Le pari semble d'autant plus fou que le web est à l'heure actuelle un écosystème pour le moins chaotique, caractérisé par la surabondance de l'information, l'omniprésence de la falsification et de la publicité, et une extrême instabilité.

Si le web est capable de développer la connaissance et son partage, comme le laisse entendre la vidéo, cela implique que la recherche d'information atteigne de l'information pertinente. C'est devenu une banalité de parler de « société » ou d'« autoroutes » de l'information, mais nous n'avons toujours pas de mot pour dénommer précisément le décalage entre l'offre proliférante et la pertinence réelle de l'information. Certains proposent le terme d'« infobésité », d'autres préfèrent parler de saturation de l'attention (ou d'économie de l'attention)[4]. Le fait est qu'en peu de temps – le web n'a pas vingt ans –, le volume de données a littéralement explosé, jusqu'à devenir incommensurable. Allons plus loin : s'il est vrai que le web est un espace documentaire sans précédent, il ne suffit pas d'évoquer son immensité, il faut rappeler qu'il est d'abord un terrain d'échanges, une infrastructure de communication. De ce fait, il est difficile d'y évaluer le statut de l'information : avant d'être un espace documentaire, le web est un espace social. Tout le monde peut s'y exprimer. Chaque page est liée à d'autres pages qui lui donnent son sens, mais ces pages disparaissent et changent, dans un maelström de flux d'informations. C'est le règne du palimpseste, sans les ratures qui permettraient de savoir que le texte a été modifié, puisque la mise en forme et le texte sont des enregistrements séparés[5]. Outre ce caractère mouvant, il y règne le plus grand désordre ; on y trouve pêle-mêle toutes les opinions, y compris celles qui doivent contourner la censure. On le présente d'ailleurs souvent comme un territoire où tout est encore possible – de fait, la législation peine à suivre son

4. Cf. [Iskold, 2007].

5. Par exemple, les systèmes de gestion de contenu (CMS) et les blogs enregistrent le texte dans une base de données totalement indépendante de l'organisation de la page et du site dans lesquels il va être affiché.

développement. On y rencontre autant de « profils » que de personnes véritables : des pseudo-identités, des "avatars" du même individu. Enfin, il semble aujourd'hui inconcevable d'entreprendre le catalogage raisonné de la toile. Sans même parler du web « profond » ou « invisible » [6], les annuaires ont vécu. Tout devient accessible en ligne, et ce immédiatement, mais si tout est accessible en même temps, alors rien n'est vraiment ordonné. Tant et si bien que le problème se pose maintenant de reconnaître ce qu'on cherche au milieu d'un grand bazar. Et ce problème de la "botte de foin" se double d'une difficulté de distinguer le bon avatar du mauvais, qu'il s'agisse d'usurpation pure et simple ou seulement de la coexistence de différentes versions d'un même document [7]. En effet, parmi les millions de pages proposées par Google, il s'agit d'éviter d'ouvrir toutes celles qui se contentent de reprendre la même information, souvent en la déformant, et de trouver une source de première main.

Pour pouvoir dépasser ces difficultés, le web de données devrait donc proposer des outils incroyablement puissants, capables de collecter et articuler des informations pertinentes en vue de résoudre des problèmes aussi bien théoriques que pratiques. Serions-nous arrivés à l'ère de Hal, le fameux ordinateur de *2001, l'odyssée de l'espace* ? Après tout, la machine est devenue capable de répondre à des questions de culture générale, et, non contente de battre les meilleurs joueurs d'échecs, de surpasser les champions du jeu à questions *Jeopardy* [8]. En tout cas, l'objectif du web sémantique est bien de parvenir à rendre l'information signi-

6. Il s'agit du web qui n'est pas moissonné par les moteurs de recherche. Nous développerons ce point dans la première partie.

7. L'apparition du *phishing*, qui consiste à créer frauduleusement un double d'une page pour récupérer les informations que l'internaute hameçonné y saisit, est peut-être le cas le plus emblématique d'une nébuleuse de confusions. Concernant le problème des avatars et, plus généralement, de la dilution du document, voir notamment l'article de M.-A. Chabin qui présente le concept de « diplomatique numérique » : [Chabin, 2011].

8. Après Deep Blue, voici Watson, l'ordinateur d'IBM capable de répondre à des questions de culture générale : → `http://fr.wikipedia.org/wiki/Watson_%28intelligence_artificielle%29`.

fiante (« *meaningful content* » : cf. sous-titre de l'article précité de Berners-lee et *alii*) pour les ordinateurs.

Avant de nous demander comment cela est possible et afin d'écarter de faux problèmes, examinons ce que l'idée a de choquant. Nous sommes habitués à l'idée qu'une machine puisse effectuer des calculs, mais la pensée est affaire d'intentionnalité et ne se réduit pas au calcul. Il n'est pas nécessaire de supposer que la machine doit comprendre les opérations arithmétiques qu'elle effectue pour produire un résultat juste, mais comment peut-elle répondre à des questions, manipuler des informations et les évaluer sans les comprendre ? Si le calcul nous paraît aisément décomposable en processus mécaniques relativement simples, eux-mêmes réductibles à des échanges de signaux, la compréhension du langage humain nécessite en revanche une interprétation, donc une capacité de choisir entre des significations concurrentes. Le problème consiste donc à montrer comment on peut rendre un automate capable de décider devant des ambiguïtés, et à savoir si le sens est perçu par la machine qui le manipule et/ou si ce sens a une réalité objective [9] – indépendante d'un esprit qui le pense. Il semble que nous soyons devant l'alternative suivante : soit les machines n'ont pas d'esprit/âme et elles ne comprennent rien, mais alors comment expliquer qu'elles puissent interpréter, soit le sens ne se situe pas forcément dans l'esprit/âme, mais alors on pourrait dissoudre l'esprit dans la lettre, voire éliminer les notions d'esprit et de conscience, qui sont au fondement de notre civilisation.

La première solution conduit à l'abandon de l'explication rationnelle au profit de principes magiques, et rappelle les nombreux récits ayant pour thème l'homme artificiel depuis le mythe du Golem. Sans aller jusqu'au registre du fantastique,

9. Après tout, les récentes avancées des neuro-sciences, notamment grâce à l'imagerie cérébrale, et les progrès des « interfaces cerveau-machine », laissent penser que les représentations mentales sont "à portée de microscope", voire manipulables. Sur l'observation des représentations dans les circuits neuronaux, voir notamment le dialogue entre J.-P. Changeux et P. Ricœur *La nature et la règle – ce qui nous fait penser*.

le simple usage du terme "immatériel" pour qualifier l'information est révélateur d'une difficulté à penser la mémoire de l'ordinateur :

> Il faut se défaire de l'idée que les technologies cognitives et culturelles sont immatérielles : l'immatériel n'existe pas. La matière, devenue flux, est de moins en moins solide, elle n'en est pas pour cela immatérielle, et il faut au contraire en outre de plus en plus de matériels pour la transformer. Quand on parle d'immatérialité, on tente de désigner inadéquatement l'invisibilité de la matière, ou, plus profondément, on tente de réfléchir sur ce qui a considérablement bouleversé notre vision de la matière, à savoir la maîtrise relative de sa vitesse. Parler d'hypermatériels et d'hypermatérialité, c'est rappeler que ce qui est en jeu aujourd'hui est la maîtrise de la matière-énergie dans ses moindres états et à toutes échelles, non la supposée immatérialité de l'information. Le propre d'une technologie de l'esprit, qui est de produire des effets sur un esprit, n'est évidemment pas son « immatérialité [10]».

La deuxième solution, dite de "l'intelligence artificielle forte", est en apparence plus satisfaisante pour la raison scientifique, mais elle abolit les différences entre ordinateur et cerveau, programme et esprit : il ne s'agit plus alors de magie, mais de science-fiction. Il est vrai que les neurosciences nous donnent à voir dans les circuits du cerveau des combinaisons de neurones correspondant à nos représentations subjectives, et rendent concevable une manipulation directe de l'esprit ; mais il ne faut pas oublier que le fait de voir les zones activées lors d'une impression douloureuse ne nous dit rien de l'état mental ressenti par le sujet, de l'interprétation qu'il s'en fait. En d'autres termes, cette solution ignore la distinction fondamentale de ce qui existe *intrinsèquement* ou seulement *aux yeux d'un observateur* : à moins de considérer qu'une calculatrice comprend les opérations qu'elle effectue, on peut affirmer que la machine *simule* mais ne *reproduit* pas l'intelligence, le calcul étant une réalité mentale :

> Les manipulations de symboles formels n'ont par elles-mêmes aucune intentionnalité ; elles sont assez bien dépourvues de signification ; elles ne

10. B. Stiegler et Ars Industrialis. Voir également le concept de mnémotechnologie : «l'extériorisation se fait vers des appareils auxquels il est possible de déléguer de nouvelles fonctions cognitives» (Manifeste d'Ars Industrialis p. 152) et l'extrait en annexe : → 6.2.

> sont même pas des manipulations de symboles, puisque ces symboles ne symbolisent rien. En jargon linguistique, ils n'ont qu'une syntaxe et pas de sémantique. L'intentionnalité dont semblent dotés les ordinateurs est seulement dans 1'esprit de ceux qui les programment et les utilisent, de ceux qui font les entrées et interprètent les sorties [11].

Nous essaierons de montrer qu'il s'agit d'une alternative infondée – provenant en fait d'une conception simpliste des relations entre esprit et matière. En effet, l'analyse des technologies et des formalismes du web de données que nous nous proposons permettra d'entrevoir pourquoi les machines peuvent manipuler le sens sans pour autant le comprendre, ni même avoir besoin de l'interpréter. On a pu confondre le projet de web sémantique avec une volonté de développer l'intelligence artificielle dans la direction du traitement automatique de la langue naturelle [12], mais il consiste surtout en une nouvelle manière de représenter les connaissances. La linguistique montre que le sens est, en même temps qu'une réalité mentale, *un rapport réglé entre des signes* : dans cette mesure, les machines peuvent opérer des inférences alors qu'elles ne sont pas capables d'interpréter le discours humain. Les fondateurs du web sémantique l'ont d'ailleurs toujours dit, et cette partie de l'exposé a souvent été ignorée, c'est plus de structure et de logique que l'on parle, que d'intelligence artificielle [13]. Sans doute le malentendu provient-il de l'usage du terme sémantique, sans doute faut-il y voir l'une des raisons qui ont fait apparaître l'expression « web de données » en lieu et place de « web sémantique », et nous avançons l'hypothèse que le contresens est ali-

11. J. R. Searle, → `http://e-philo.univ-paris1.fr/Searle.pdf`. Pour une analyse des limites de la mécanisation de la pensée en rapport avec les ontologies, voir G. Declerck et J. Charlet : [Charlet et Declerck, 2011] .

12. Abrégé en TAL ou TALN ; l'expression anglaise est *Natural Language Processing*, NLP. Nous aborderons cette confusion dans la première partie.

13. Toujours dans le même article de Berners-Lee et *alii* : l'ordinateur « "saura" tout cela sans avoir recours à l'intelligence artificielle comme avec Hal (l'ordinateur de 2001 Odyssée de l'espace) ou avec C-3PO de la *Guerre des étoiles*. Au contraire, cette sémantique a été encodée dans la page Web au moment où le directeur de la clinique l'a mise en forme en utilisant un logiciel d'écriture de pages Web sémantique comprenant les ressources du site de l'Association de kinésithérapeutes. ».

menté par l'aspect fantasmagorique évoqué ci-dessus. Nous tenterons d'exposer les procédés du web de données de manière à départager ce qui relève de l'humain et de l'automatique, comme Poë l'a fait brillamment pour le canular de l'automate joueur d'échecs : si on admet que les machines se contentent de simuler le dialogue lorsqu'elles nous aident dans nos recherches d'information pertinente, reste à expliquer comment l'illusion fonctionne. Il est possible de renvoyer dos à dos les deux branches de l'alternative, en les affaiblissant et en posant à la fois que les machines ne comprennent rien et que le sens n'est pas la propriété exclusive de l'esprit mais s'incarne dans une forme manipulable.

Nous espérons ainsi réunir des éléments pour lever une grande partie des objections que l'on peut faire aux concepts de web de données et de web sémantique, et à leur application dans le domaine des données patrimoniales. En effet, l'interconnexion des bases de données de bibliothèques, d'archives et de musées, mentionnée par la vidéo d'Europeana, ne relève plus de la science-fiction, et ne signifie pas non plus que l'ordinateur ordonnera les collections sans l'aide de l'homme. Au contraire, on peut considérer le web de données comme une renaissance des vocabulaires contrôlés. La question de savoir si le contexte économique et juridique – notamment, les débats sur l'*Open Data* [14] – est favorable à cette interconnexion ne saurait être traitée dans le cadre de cet exposé, et il s'agira ici seulement d'étudier les contradictions et les conciliations possibles entre la logique des collections patrimoniales et la logique du web, en nous appuyant sur le concept deleuzien de rhizome.

À l'heure où le volume des collections numériques explose, où les catalogues des collections physiques sont très souvent en ligne, on peut se demander comment les institutions patrimoniales peuvent s'inscrire dans le web sans s'y perdre [15]. On peut se poser un certain nombre de questions : le web de données, en

14. Voir cette présentation : [OKFN, 2009].
15. Cette étude entendra le terme de collection dans l'acception très large de 'ce qui est collecté',

reliant des vocabulaires et des données qui sont produits dans des contextes différents, ne risque-t-il pas de produire du non-sens ou de l'ambiguïté, alors que sa vocation est précisément de désambiguïser et de distinguer le fiable du douteux ? La tentative n'est-elle pas nécessairement vouée à l'échec, du fait de l'irréductible diversité des langues et jargons ? Enfin, les organismes culturels sont attachés à l'idée d'une information de qualité, vérifiée, fiable, pérenne : comment cette conception peut-elle subsister dans l'écosystème du web ? En effet, contrairement à la conception classique du savoir où l'on doit, en théorie, pouvoir faire le tour d'une question en épuisant les ressources disponibles sur le sujet, l'architecture en réseau du web impose un changement de modèle :

> Les chercheurs du Web sémantique, au contraire, acceptent les paradoxes. Les questions sans réponse sont le prix à payer pour acquérir de la souplesse. Nous construisons un langage de règles aussi significatif que nécessaire pour permettre au Web de raisonner autant qu'on le veut. Cette philosophie ressemble à celle du Web classique : dès le développement du Web, ses détracteurs ont souligné qu'il ne pourrait jamais être une bibliothèque bien organisée, que sans base de données centrale et sans structure arborescente, on ne pourrait jamais être sûr de tout trouver. Ils avaient raison [16].

Dans un premier temps, l'étude présentera les principaux concepts du web de données, dans une démarche d'articulation progressive des couches (*semweb stack layers*, voir schéma ci-dessous [17]) qui le composent ; nous verrons alors comment le web de données s'appuie sur des descriptions faites par l'homme et permet de relier des collections hétérogènes. Cette méthode, qui relève de l'« ingénierie philosophique [18] » permettra, dans un second temps, de répondre aux objections que l'on a pu adresser au projet, en étudiant les propriétés architecturales

et englobe donc les fonds d'archives. Les collections du web 2.0, qu'il s'agisse d'objets multimedia comme Flickr ou de signets comme Delicio.us, seront également évoquées. Autre parti pris : nous considérerons que les instruments de recherche, les données bibliographiques et les descriptions des collections de musées, font partie des collections ; *idem* pour les « jeux de données » (datasets).

16. Berners-Lee et *alii*, *ibid*.
17. → `http://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(24)`.
18. Terme de T. Berners-Lee : voir le commentaire qu'en donne A. Monnin [Monnin, 2012a].

du *Giant Global Graph* (autre nom donné par Berners-Lee au web sémantique) et la manière dont les territoires peuvent y communiquer sans empiéter les uns sur les autres – grâce à un enchâssement des vocabulaires de description. L'interrogation portera également sur la possibilité de conserver une cohérence malgré l'hétérogénéité des données liées, notamment à travers les tentatives pour décrire leur contexte et leur provenance.

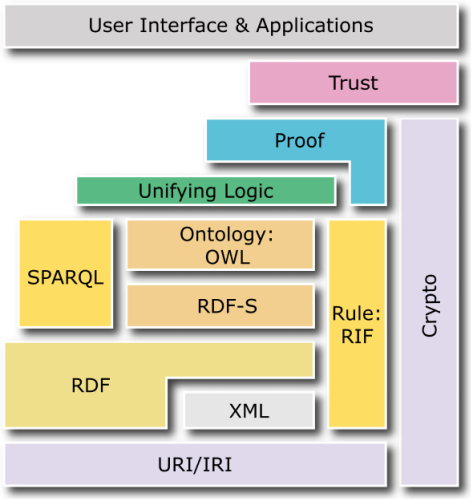


Figure 1 : le « Semweb Stack » (ou « Layer Cake ») de T. Berners-Lee

2 L'expression du sens : identifier, lier, inférer

> Adding semantics to the web involves two things : allowing documents which have information in machine-readable forms, and allowing links to be created with relationship values.
>
> T. Berners-Lee, *Plenary at WWW Geneva 94*.

> L'ordre et la connexion des idées sont les mêmes que l'ordre et la connexion des choses.
>
> B. Spinoza, *Éthique*, II, 7.

Comment encoder le sens de manière à ce que les machines puissent le manipuler ? Cette première partie a pour objectif d'étudier les briques du web de données, en suivant la logique d'empilement des couches de formalisme schématisée par le fameux Semweb Stack (fig. 1). Nous verrons que l'expression du sens [19] passe par des mécanismes permettant d'identifier les ressources de manière univoque, de les décrire selon une grammaire et des vocabulaires pour finalement réaliser des inférences ou « raisonnements » automatiques.

2.1 Structurer les données = typer les liens

Une limitation importante du web, qui a été perçue dès le départ, est l'absence de structuration des données qui sont mises à disposition. Par exemple, lorsque nous voyons l'image scannée d'un tableau dans une page web, il n'est pas possible d'effectuer des tris ou des calculs avec un tableur comme LibreOffice Calc. En d'autres termes, les informations contenues dans cette image que nous voyons ne sont pas reconnues comme des informations alphanumériques, à moins que nous utilisions un OCR, mais traitées comme un flux multimédia [20]. Ou encore, si nous cherchons le texte intégral de l'*Éthique* de Spinoza, il faudra éliminer

19. Nous empruntons ce titre à l'article déjà commenté de Berners-Lee et *alii*.
20. Exemple adapté de [Berners-Lee, 2006].

toutes les pages de catalogues de librairies, etc. Rien ne dit *a priori*, dans les liens que nous voyons parmi les résultats des moteurs de recherche ou sur d'autres pages, si on a affaire au texte ou à une simple citation. C'est que les pages ne sont pas catégorisées par un *type* identifié clairement : on doit s'en remettre au contexte de la page pour comprendre le sens du lien. Il existe bien en HTML une manière de caractériser le contenu des pages, à savoir le champ de métadonnées (i.e. données décrivant les données) inséré dans la partie en-tête du code de la page web (`<head> <meta>`) ; mais en admettant qu'une description soit renseignée dans ce champ, l'internaute ne la voit pas forcément. La seule exploitation qui en est faite est, la plupart du temps, l'extraction de mots-clés par les moteurs de recherche ; mais les sites commerciaux, entre autres, ayant abusé de ce procédé en utilisant un maximum de termes pour attirer des visiteurs, les mots-clés ne signifient plus grand'chose [21]. En résumé, une requête sur un sujet ne permet pas de préciser le type d'information attendu, et la navigation par les liens se fait à l'aveuglette, du fait des limites du langage HTML :

> Dans les liens hypertexte du web conventionnel, la nature de la relation entre deux documents liés est implicite, du fait que le format de données, i.e. HTML, n'est pas suffisamment expressif pour permettre que les entités décrites dans un document particulier soient connectées aux entités associées par des liens typés [22].

Or, c'est précisément à ce niveau qu'interviennent les technologies de représentation de l'information sur lesquelles repose le web de données : il s'agit d'offrir à l'internaute une présentation du web qui lui permette de voir immédiatement, en même temps que le lien, que "ceci n'est pas un texte mais une publicité pour une librairie", et même de lui permettre de spécifier dès le départ la catégorie d'information qu'il recherche.

21. À tel point que l'importance de ces descriptions dans le référencement par les moteurs de recherche semble tendre vers zéro depuis 1997. Cf. l'article de D. Sullivan : → `http://searchenginewatch.com/article/2066825/Death-Of-A-Meta-Tag`.

22. Cf. [Bizer *et al.*, 2009].

Il est vrai que le web offre également des informations structurées, comme les portails de consultation de catalogues : lorsqu'on remplit les champs d'un formulaire, on effectue bien une catégorisation des termes recherchés, on peut faire la différence entre une recherche des œuvres écrites *par* Spinoza, qui sera alors entré dans le champ 'auteur', et une recherche des œuvres *sur* Spinoza, s'il est entré dans le champ 'sujet'. Toutefois, ces bases de données ne sont pas toujours directement accessibles *via* les liens hypertexte et les moteurs de recherche. En effet, les pages que l'internaute voit s'afficher dans son navigateur sont créées au moment où il formule sa requête, par une extraction d'informations depuis la base de données qui sont ensuite 'incrustées' dans un cadre HTML ; ces pages dites « dynamiques » ne peuvent donc pas être désignées auparavant comme cibles d'un lien ni, du même coup, moissonnées par les robots d'indexation, puisque ceux-ci se contentent de suivre les liens. D'où l'existence d'un « web profond » ou « invisible », qui constituerait d'ailleurs selon certains la partie immergée de l'iceberg ; dans ce deuxième web, il faut déjà savoir précisément où chercher pour obtenir une information, il faut connaître le point d'accès, et souvent le titre, l'auteur du document – puisque celui-ci n'est pas indexé en plein texte par les moteurs de recherche disponibles à l'intérieur des sites.

Schématiquement, la situation actuelle peut se résumer par la coexistence du web visible et navigable, océan chaotique où seuls les moteurs de recherche permettent de se repérer, et de ce web structuré mais caché. Les institutions culturelles sont malheureusement souvent enfouies dans ce deuxième web. À moins que leurs données n'aient fait l'objet d'une démarche d'exposition au moissonnage ; mais même si c'est le cas, la structuration est perdue [23], et les informations ne sont pas regroupées avec les sources qui les compléteraient utilement :

23. Par exemple, un moteur de recherche web ne tient pas compte des différents types de descripteurs utilisés dans une notice : il ne reconnaît pas (encore) la *séquence alphanumérique* " 1632-1677 " comme des dates contenues dans la page web affichant la notice de Spinoza.

> Actuellement, les données de bibliothèque sont stockées dans des bases de données qui, même si elles peuvent avoir des interfaces web de recherche, ne sont pas réellement intégrées aux autres sources de données du web. Une quantité considérable de données bibliographiques et d'autres types de ressources sur le web partagent des données communes telles que des dates, des informations géographiques, des noms de personnes et d'organisations. Dans le futur environnement du web de données, tous ces points communs ont vocation à être connectés [24].

Le but du web de données est donc de sortir les informations de leurs « silos » et de faire du web une gigantesque base de données [25] que l'on pourrait parcourir au moyen de liens hypertexte.

Passer d'un « web de documents » à un « web de données », cela signifie donc que l'on change de niveau de granularité de l'information : on descend du document à la chose dont il parle. Il faut noter qu'on effectue le même mouvement lorsque l'on consulte des données bibliographiques, où la description du document est en même temps le nom du sujet qu'il traite, de sorte qu'elle en fournit un substitut condensé. Ainsi l'œuvre ayant pour titre *Soi-même comme un autre*, est caractérisée par le descripteur 'Altérité', qui est un *terme associé* à 'Autrui', lui-même étant un *terme spécifique* de 'Morale', qui est *un terme employé pour* 'Éthique' [26] : grâce à l'indexation par les vedettes-matière du vocabulaire contrôlé, on peut savoir que Spinoza et Ricœur se sont intéressés aux mêmes questions. Formulée ainsi, l'idée paraît simple et semble ne poser aucun problème. Pourtant, la mutation est radicale, puisque les liens n'ont plus seulement pour cible des ressources documentaires mais désignent des entités conceptuelles qui seront elles-mêmes traitées comme des ressources du web. Le projet suppose également que l'on puisse harmoniser, rendre « interopérables », les vocabulaires de descrip-

24. [W3C, 2011].
25. [Bizer et Heath, 2011].
26. Exemple tiré de RAMEAU, le vocabulaire contrôlé utilisé pour l'indexation à la Bibliothèque Nationale de France. Accès en ligne *via* le catalogue : → `http://catalogue.bnf.fr/`.

tion et les formats d'encodage des données. La difficulté est donc à la fois technique et organisationnelle.

2.2 La grammaire RDF et l'identification par les URIs

2.2.1 Une « gigantesque base de données » ?

Le web actuel est un espace interopérable dans le sens où l'on peut passer d'un site à un autre, sans s'en rendre compte, par un même mécanisme : la navigation par les liens. Toute page peut être connectée à n'importe quelle autre : comment étendre cette interopérabilité aux informations contenues dans les documents, et naviguer parmi des données ? Il 'suffit' de respecter les quatre règles suivantes, qui sont les quatre piliers du *linked data* :

> 1) Utiliser les URIs [27] comme des noms pour les choses.
> 2) Utiliser des URIs HTTP [28] de telle sorte que les gens puissent voir ce qu'il y a derrière ces noms.
> 3) Lorsque quelqu'un va voir derrière une URI, lui fournir des informations utiles au moyen des standards (RDF et SPARQL).
> 4) Inclure des liens vers d'autres URIs, de telle sorte qu'on puisse découvrir d'autres choses [29].

Les règles n° 2 et 4 sont relativement simples à comprendre pour une première approche : le protocole HTTP est ce qui permet de naviguer sur le web *via* les liens hypertexte, il faut s'assurer que les URIs soient accessibles par un navigateur et qu'elles permettent de consulter des ressources présentes sur la toile ; quant au tissage des liens pour enrichir l'information, c'est un principe évident pour tout internaute. En revanche, l'identification des choses/ressources par des adresses et

27. *Uniform Resource Identifier* : identifiants uniformes de ressources, ayant notamment pour propriété d'être uniques.
28. I.e. des URL (*Uniform Resource Locator*), des adresses accessibles via le protocole http, localisant des ressources sur le réseau et permettant de les afficher.
29. CF. [Berners-Lee, 2006].

l'utilisation de RDF, grammaire minimale pour décrire les choses/ressources [30], nécessitent quelques explications.

Nous avons dit que le projet se donne pour objectif de transformer le web en une gigantesque base de données, mais il faut préciser cette affirmation : les données tabulaires stockées dans des bases le sont sous forme binaire et dans de nombreux formats différents, c'est pourquoi il est impossible de les atteindre directement au moyen des liens et d'exploiter leur structure sans disposer du système de gestion de bases de données compatible avec le format [31]. En outre, la forme tabulaire présente l'inconvénient majeur d'exprimer l'élément enregistré et son type dans des "dimensions" différentes : le champ (colonne) et l'enregistrement (ligne) ne sont pas de même nature que les valeurs inscrites dans les 'cases' – comme des coordonnées permettant de repérer un point. Le type de la valeur ne lui est attribué que relativement, par sa position dans une table de telle base, à l'intersection du champ et de l'enregistrement.

Soit par exemple le cas de données bibliographiques sur *Les mots et les choses* contenues dans une table 'T_books' : on sait que Foucault est l'auteur de l'œuvre parce que la chaîne de caractères 'Foucault, M.' est une valeur enregistrée sur la ligne L dans la colonne 'author'. Sans ces coordonnées, impossible d'attribuer une catégorie à un élément : cela implique que l'ordinateur puisse parcourir la structure de la table pour que la chaîne de caractères soit identifiée et traitée comme 'author'. Il est donc très complexe de récupérer des données d'une base pour les agréger à une autre si les architectures diffèrent. De la même manière, la création de relations est alourdie par la nécessité de connaître la structure des tables, et d'ajouter des colonnes ou des tables intermédiaires : pour permettre le passage d'un

30. RDF signifie *Resource Description Framework*, ce qui peut être traduit par 'cadre de description de ressources'.

31. Il est nécessaire d'extraire et de convertir les données pour les transférer d'une base à l'autre, ou d'utiliser des protocoles complexes pour la recherche fédérée. Sur ce point, voir les pages de D. Lahary sur la norme Z39.50 : → [Lahary, 2003].

enregistrement à l'autre, il faut pouvoir réunir leurs identifiants. Par exemple, si l'on a une autre table 'T_authors', contenant des informations biographiques sur les auteurs, et que 'Foucault, Michel' est enregistré dans la colonne 'name' sur la ligne M, on peut souhaiter relier les données avec 'T_books' pour faire apparaître les dates de Foucault avec la présentation de l'œuvre. Il faudra créer une table intermédiaire 'T_link' de deux colonnes ('id_book' et 'id_author') pour y enregistrer sur la même ligne les identifiants L et M. Pour faire des liens entre les données, non seulement il faut modifier la structure de la base, mais encore ces liens sont perdus si l'on extrait les données d'une table sans celles des tables associées. En effet, les relations sont extérieures aux valeurs, donc implicites : du fait qu'elles sont induites, les relations entre les données rendent les tables interdépendantes et inexploitables dans un autre contexte [32].

Au contraire, le *linked data* a pour préalable que les données soient encodées sous forme de texte, ce qui les rend navigables – et accessibles aux grands moteurs de recherche. Cela permet que la structure des données, i.e. leur *type*, soit elle-même être exprimée sous forme textuelle, *avec* les données : ainsi, les limites des bases de données peuvent être dépassées.

2.2.2 Le triplet comme relation entre URIs

L'utilisation du formalisme de RDF permet de mettre les données et la structure au même niveau, dans une même "dimension", en les agrégeant sous la forme grammaticale ternaire et très simple nommée « triplet » (en anglais, *triple*), constituée de la séquence 'sujet, prédicat, objet' (ou 'sujet, propriété, valeur'). La structuration tabulaire laisse place à une description "à plat", et si l'on reformule les

32. Pour une analyse plus détaillée mais très abordable des limitations des bases de données, voir la présentation de Bertails et Poupeau : [Bertails et Poupeau, 2009].

informations précitées on obtient très simplement ces deux formulations :

"Les mots et les choses" "author" "Foucault, M."

"Michel Foucault" "dates" "1926-1984"

Il reste alors à insérer les éléments de cette grammaire dans l'enveloppe d'un fichier et à indiquer aux logiciels la localisation des ressources correspondantes.

Pour que cette grammaire soit prise en compte, il est en effet indispensable que ses syntagmes soient déclarés et reconnus par les logiciels, et l'on peut utiliser pour cela un langage à balises. Ainsi, le langage HTML est un code qui utilise des balises de mise en forme, de création de liens et autres instructions [33] ; le langage XML, lui aussi accessible aux navigateurs, offre la possibilité d'enregistrer des données structurées dans les subdivisions arborescentes d'un élément racine. Le formalisme RDF sera donc, par exemple [34], encapsulé dans un fichier au format XML.

Mais jusque là, nous n'avons qu'une coquille vide et des chaînes de caractères (appelées littéraux) ; or, si nous connaissons l'œuvre *Les mots et les choses*, rien n'indique au logiciel « ce qu'il y a derrière » cette expression qui fait sens pour nous. Il faut donc associer à cette entité/ressource conceptuelle un nom univoque, à destination de la machine : un identifiant qui servira à la fois d'*adresse* de référence – ou de métadonnée – pour l'ensemble des ressources documentaires la concernant et de *support pour des assertions* formulées en RDF. Ainsi, `http://dbpedia.org/resource/Les_Mots_et_les_choses` est une URI qui désigne l'œuvre et nous renverra automatiquement vers une page décrivant l'œuvre si on demande au navigateur de s'y rendre [35], mais c'est aussi un sujet pour

33. " `<html> <I> texte </I> </html>` " est transformé par le navigateur en " *texte* ". Les balises ouvrante et fermante " html " indiquent au logiciel le format de fichier, et les balises " I " la commande " afficher des italiques ".

34. Il existe d'autres manières d'écrire « sérialiser » le RDF, que nous utiliserons par la suite pour simplifier la lecture.

35. DBpedia, nœud majeur dans le réseau du web de données, est un site qui expose selon les principes du *linked data* une partie des données de Wikipedia. Lorsqu'on tape l'URI " http ://dbpedia.org/resource/etc... " on est automatiquement redirigé vers

l'attribution de la propriété "author" avec la valeur "Foucault" [36]. Si c'est un programme qui utilise cette URI, il sera redirigé vers un fichier contenant des données en RDF [37] , dont il pourra extraire immédiatement des éléments, et ce avec leur structure :

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF (...)>
(...)
<dbpedia-owl:author
rdf:resource="http://dbpedia.org/resource/Michel_Foucault"/>
(...)
</rdf:RDF>
```

Nous avons ici une description de l'œuvre (l'URI sujet étant sous-entendue, puisqu'elle est aussi l'identifiant de la ressource à laquelle le fichier se rapporte) qui la met en relation avec une autre ressource (concernant Foucault et également exprimée par une URI : `http://dbpedia.org/resource/Michel_Foucault`).

Première remarque : cette description nous fait connaître immédiatement le type de cette relation : 'avoir comme auteur' : le triplet forme un ensemble autonome, contenant la structure. Bien plus : cette relation est elle-même identifiée par une URI, ici abrégée par le préfixe `dbpedia-owl:` devant `author`. Autrement dit, cette relation est également une ressource, qui peut faire l'objet de nouvelles assertions en RDF ; tous les éléments de la description peuvent faire l'objet d'une description dans le même formalisme, dans la même 'dimension'. C'est pourquoi RDF est aussi un *métalangage*, à savoir un langage sur le langage – nous y reviendrons.

Deuxième remarque : l'utilisation des URIs permettra de retrouver partout sur le web des informations structurées sous la forme de triplets et se rapportant bien à

" http ://dbpedia.org/page/etc... ". Ce mécanisme de redirection porte le nom de « négociation de contenu ».

36. Bien sûr, une URI étant le nom d'une ressource conceptuelle, elle peut aussi bien être prédicat ou objet.

37. Pour simplifier l'exemple, nous extrayons le triplet du fichier `http://dbpedia.org/data/The_Order_of_Things.rdf`.

la même ressource – c'est d'ailleurs ainsi qu'Europeana peut nous aider à désambiguïser nos requêtes (cf. introduction). Le protocole SPARQL [38] , mentionné dans les quatre principes du linked data, donne la possibilité d'interroger de manière structurée (comme SQL pour les bases de données) n'importe quel ensemble de triplets (ou « graph »). Signalons également l'existence de moteurs de recherche sémantiques, comme Sindice, qui indexent une multitude de jeux de données (*datasets*) et permettent d'utiliser toute la richesse du langage de requête fourni par SPARQL [39]. Mais il y a plus : à la condition que l'on réutilise ces identifiants, ou que l'on établisse une équivalence entre identifiants, une machine peut réunir des triplets qui apparaissent dans des contextes différents, les en extraire automatiquement – et ce sans qu'ils perdent leur sens, puisqu'il suffit de suivre les URIs pour retrouver leur signification. L'URI `http://www.w3.org/2002/07/owl#sameAs`, entre autres, permet d'exprimer qu'une URI est identique à une autre, et par ce biais on peut relier les données concernant Foucault qui se trouvent dans deux jeux de données différents.

```
<http://data.bnf.fr/ark:/12148/cb11903202t#foaf:Person>
<http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Michel_Foucault>
```

est un triplet posant l'égalité entre la ressource identifiée par le numéro `11903202` dans les données de la BnF et `Michel_Foucault` dans DBpedia.

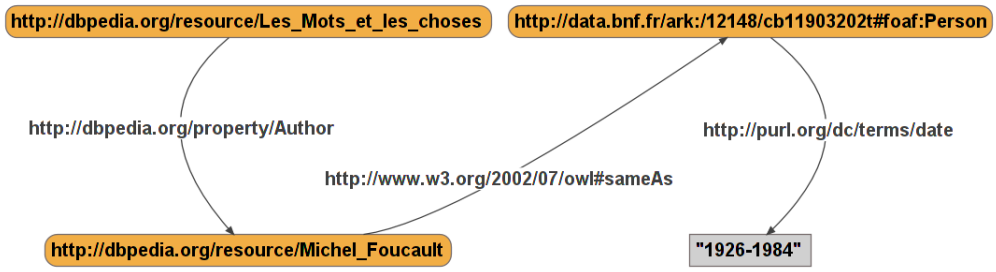
Si ce triplet est présent dans DBpedia, cela permet à une machine de récupérer très facilement l'information suivante à la BnF :

38. Pour *SPARQL Protocol and RDF Query Language*. Voir par exemple le point d'accès (*sparql endpoint*) au jeu de données de DBpedia : → `http://dbpedia.org/sparql`.

39. Voir notamment l'interface avancée de Sindice, qui présente l'avantage d'être très intuitive (puisqu'elle ne nécessite pas de connaître la syntaxe de SPARQL) : → [DERI,]. D'autres moteurs de recherche sont disponibles à partir de cette page du W3C : `http://www.w3.org/2001/sw/wiki/Category:Search_Engine`. Enfin, mentionnons le site du Datahub, qui recense plus de 700 *datasets*, dont 90 pour le groupe Bibliographic Data : → [OKFN,].

```
<http://data.bnf.fr/ark:/12148/cb11903202t#foaf:Person>
<http://purl.org/dc/terms/date>
"1926-1984"
```

On voit que les informations de notre exemple s'agencent dans une série de triplets permettant de circuler d'une ressource à l'autre et éventuellement de récupérer, en suivant les liens, des informations issues d'une autre source. Pour conclure ce point, nous pouvons résumer l'enchaînement par le « graphe » ci-dessous.



2.3 Ontologies et inférences

2.3.1 Pourquoi des ontologies ?

L'analyse de notre exemple mobilise plusieurs catégories de ressources, et présente une difficulté qu'il faut maintenant expliciter. À première vue l'entité 'Foucault', pour commencer, ne semble pas poser problème : c'est qu'il s'agit d'un concept désignant une réalité singulière, que nous connaissons, et dont la description est immédiatement compréhensible par l'homme, puisqu'une information fournie sous la forme d'un littéral, "1926-1984".

Or jusqu'ici nous avons montré comment le formalisme RDF permet d'identifier le littéral et de le catégoriser comme des dates, mais cela n'explique pas comment une machine est capable de traiter ces chiffres comme des dates. De même, pour la propriété `owl:sameAs` : comment apprend-t-on à l'ordinateur à faire équivaloir deux ressources ? Ou encore, comment éviter que les machines

infèrent des liens aberrants, comme celui qui consisterait à attribuer des dates à un concept ? En effet, *Foucault* est aussi le titre d'une œuvre de Deleuze, donc l'URI désignant la personne 'Foucault' peut servir d'objet dans le triplet

```
<URI_œuvre_de_Deleuze> dc:subject <URI_Foucault>
```

et l'on peut souhaiter la réutiliser en tant que concept ou « tag » dans un système de catégorisation de type folksonomie [40].

La résolution de ces problèmes est possible grâce à l'utilisation de vocabulaires nommés « ontologies », qui, nous le verrons, expliquent l'usage du terme « web sémantique », mais doivent être différenciés des technologies du traitement de la langue. Commençons en effet par bien distinguer les deux domaines.

Le traitement automatique du langage naturel est un ensemble de procédés qui permettent de reconnaître des formes linguistiques et donc d'extraire le sens d'un texte rédigé par l'homme. Le programme est capable de reconnaître une phrase dans une séquence de caractères à partir du moment où on lui indique une liste de séparateurs (points, majuscules...), puis de reconnaître des mots à l'intérieur de cette phrase si on lui indique de repérer des espaces. Le sens et la fonction des mots sont ensuite reconnus grâce à des dictionnaires qui donnent leurs différentes formes (conjugaisons, accords...) et des listes de modèles grammaticaux : il y a plus de chances que "souris" soit un verbe qu'un nom si le terme se trouve après "je" [41]. L'analyse effectuée se situe donc au niveau *terminologique*.

40. Le terme `dc:subject` est une propriété définie dans les métadonnées du Dublin Core, tout comme `dc:date` utilisé plus haut. L'œuvre de Deleuze est identifiée par l'URI `http://data.bnf.fr/ark:/12148/cb34875610x` ; voir également la notice correspondante qui donne Foucault comme sujet (→ `http://catalogue.bnf.fr`). Il existe une URI pour Foucault déclarée en tant que concept sur le site communautaire http ://www.xomba.com/. Cela peut sembler une erreur manifeste de typer la même URI comme personne et comme concept, ou de relier les deux identifiants par `owl:sameAs`, mais l'exemple a pour seule fonction de présenter les risques de la modélisation. Une explication détaillée du problème est donnée ici : → [Johnston, 2011a].

41. Pour une introduction, voir [Chaudiron, 2007]. Pour une analyse détaillée de la reconnaissance des modèles grammaticaux, voir [Lallich-Boidin et Maret, 2005].

Au contraire, les ontologies sont des systèmes de représentation de connaissances, i.e. de relations entre des *concepts*[42], comme le sont les classifications et les thésaurus. Le terme, emprunté à la philosophie, a une connotation métaphysique, puisqu'il renvoie aux tentatives de construire un système unique de représentation du monde, partant le plus souvent d'un principe transcendant appelé, par exemple, Dieu. Dans ce genre de systèmes, on a pu chercher à prouver l'existence de Dieu par la simple analyse de son concept : cette forme de "preuve" a été qualifiée d'ontologique par Kant. Mais il faut se défaire de cette connotation si l'on s'en tient la définition devenue canonique :

> Une ontologie est la spécification d'une conceptualisation[43].

Comme le dit Gruber, il faut avant tout se demander quelle est l'utilité des ontologies pour comprendre leur nature : on peut dire schématiquement que ce sont des modèles permettant aux ordinateurs de manipuler de grandes quantités de connaissances et d'effectuer des inférences sur les contenus en manipulant les relations formalisées par les modèles. Pour simplifier, on peut rappeler la forme bien connue du syllogisme : si Foucault est une instance (élément) de la classe (ensemble) des personnes, alors l'ordinateur peut déduire que Foucault a des dates de naissance et de décès, si on lui a auparavant spécifié que la classe des personnes est elle-même une instance de la classe des "instances ayant des dates de naissance et de décès"[44]. Reste alors à ranger correctement les instances dans les classes, c'est-à-dire à décrire les composantes des concepts et leurs relations aussi précisément que possible. Entre cette simple déduction et la capacité de proposer un diagnostic médical, il n'y a pas de différence de nature mais seulement de degré. Et cette différence de capacité est fonction de la complexité du modèle on-

42. Sur la différence et la complémentarité des deux approches, voir [Calderan *et al.*, 2012] chap. 5.

43. T. Gruber : → [Gruber, 1992].

44. Nous ne pouvons aborder ici que des exemples simples ; pour des précisions voir [Eiter *et al.*, 2008].

tologique : un vocabulaire suffisamment expressif permet, par exemple, d'analyser le dossier médical d'un patient et de détecter des signaux d'effets indésirables liés aux médicaments en fouillant dans de gros volumes d'informations. Autre illustration de la puissance des ontologies et du calcul logique : le projet collaboratif Gene Ontology, qui associe trois modèles (processus biologique, composant cellulaire, fonction moléculaire) de manière à permettre de raisonner sur de gigantesques bases de connaissances hétérogènes – notamment, découvrir un lien entre un gène et une fonction dans un processus biologique [45].

2.3.2 RDFS et OWL pour spécifier les concepts

Toutefois cette première approche ne nous dit pas encore comment les ontologies vont pouvoir être construites à partir de RDF – qui leur sert également d'enveloppe. Son formalisme permettant de relier des ressources et de décrire leurs relations au moyen de ressources – puisque les concepts sont à la fois des ressources, et en même temps des éléments entrant dans les descriptions – nous l'avons qualifié de métalangage : la question est donc de savoir comment les liens entre concepts peuvent être formulés de telle sorte que les machines puissent "s'y retrouver". RDF fournit une enveloppe (*wrapper*), qui peut s'insérer dans une enveloppe XML, et permet d'y « délimiter des frontières, à l'intérieur desquelles le contenu renvoie explicitement à un modèle de données [46] ».

Son vocabulaire est donc très restreint, puisqu'il se contente d'indiquer la fonction, le statut des éléments, dont le sens est spécifié par un vocabulaire ex-

45. Gene Ontology est souvent cité pour l'ampleur des jeux de données et la précision de son ontologie : → `http://www.geneontology.org/GO.contents.doc.shtml`. En qui concerne l'exemple de la détection des effets indésirables, voir l'article précité de G. Declerck et J. Charlet.

46. Plus précisément : « The RDF element is a simple wrapper that marks the boundaries in an XML document between which the content is explicitly intended to be mappable into an RDF data model instance. ». Voir la recommandation du W3C de 1999 : → `http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/`.

térieur : il se limite à une vingtaine de termes[47], et s'il existe une classe nommée `Property`, c'est seulement pour signifier que les éléments qui en font partie peuvent être employés comme prédicats dans un triplet. Autre cas significatif, `XMLLitteral` : ce terme permet de dire que dans tel attribut de telle balise, seul un littéral peut figurer et non une URI, sous peine de générer un fichier incorrect et illisible.

Mais ce qui nous intéresse est au-delà de l'incorrect : il s'agit, entre autres d'éviter les confusions possibles entre les propriétés d'une ressource qui peut appartenir à plusieurs types (toujours selon notre exemple, personne ou concept) alors même que ces confusions seraient valides d'un point de vue syntaxique, conformes aux exigences du formalisme RDF. Il faut donc disposer d'un langage qui permette de définir les concepts des vocabulaires qui s'appuient sur RDF : c'est le rôle de RDF Schema, ou RDFS.

RDFS permet de « déclarer » les concepts, i.e. essentiellement de caractériser les relations entre classes et propriétés. La propriété `subClassOf` permet d'exprimer que les instances d'une classe sont également des instances d'une autre classe qui la contient : par exemple, l'ontologie FOAF définit un vocabulaire pour décrire les personnes, et la classe `foaf:Person`[48] est une sous-classe de `foaf:Agent`, au même niveau que `foaf:Organization` et `foaf:Group`. Cette propriété permet donc de définir des hiérarchies entre classes, et donc des taxonomies. Mais la relation entre classes, propriétés et instances n'est pas nécessairement une subsomption : RDFS offre notamment la relation `seeAlso`, très utilisée, qui a le même sens que la notion « terme associé » dans un thésaurus, et peut

47. Dont, entre autres : `rdf:subject`, `rdf:predicate`, `rdf:object`, `rdf:type`, `rdf:Property`, `rdf:XMLLiteral`. Pour une présentation plus complète du rôle du vocabulaire dans la syntaxe et la sémantique de RDF voir les recommandations du W3C : [W3C, 2004e] (section "Namespace") et [W3C, 2004c] (section RDF interpretations).

48. Classe qui apparaît dans l'URI de Foucault utilisée à la Bnf : voir point précédent → 2.2.2. Le projet FOAF (*Friend Of A Friend*) est présenté ici [FOAF, 2009]. Nous reviendrons sur la réutilisation de FOAF.

être utilisée comme une relation d'équivalence moins forte que `owl:sameAs`. De même, une relation très intéressante est la propriété `isDefinedBy`, qui permet d'indiquer qu'une ressource contient la définition d'une autre. Enfin et surtout [49], les propriétés `Domain` (domaine) et `Range` (co-domaine ou "portée") déterminent pour une propriété donnée quels types de ressources peuvent jouer les rôles, respectivement, de sujet et d'objet dans un triplet. Par exemple, la propriété `dc:date` évoquée ci-dessus a pour portée un littéral, et son domaine n'est pas précisé : on peut donc exprimer qu'un livre appartient à une période mais pas qu'elle est 'la belle époque' désignée par `http://dbpedia.org/resource/Belle_Epoque`. En revanche, les limitations sont plus précises avec la propriété `owl-dbpedia:birthDate`, qui a comme domaine `Person` et comme portée `xsd:date` (i.e. un type de donnée défini par un schéma xml comme séquence de chiffres sous la forme AAAA-MM-JJ) : cette contrainte répond bien à notre désir d'éviter qu'un concept ait une date de naissance [50].

Néanmoins, les relations que RDFS permet de décrire restent relativement simples, d'où la création d'un langage plus riche, OWL [51], qui rend possible notamment l'expression du fait que des classes résultent de l'assemblage d'autres classes, ou sont disjointes. FOAF, en l'occurrence, spécifie qu'aucune instance de la classe `Person` n'est en même temps une instance de la classe `Organization` alors qu'une personne peut être membre d'une organisation, et, pour reprendre notre exemple de Foucault, on peut déclarer que les concepts sont disjoints des personnes. Ou encore on peut exprimer qu'une propriété est symétrique (comme 'frère de'), inverse (fils/père), transitive (si A est inférieur à B et B inférieur à C...).

49. Pour une présentation complète, voir la recommandation du W3C : [W3C, 2004d].

50. Voir les déclarations des classes du Dublin Core et de Dbpedia : → `http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms` et `http://mappings.dbpedia.org/server/ontology/classes/Person`.

51. *Ontology Web Language*, existant en trois versions : Lite, DL, Full, de la plus simple à la plus complète.

2.3.3 Exemples d'applications

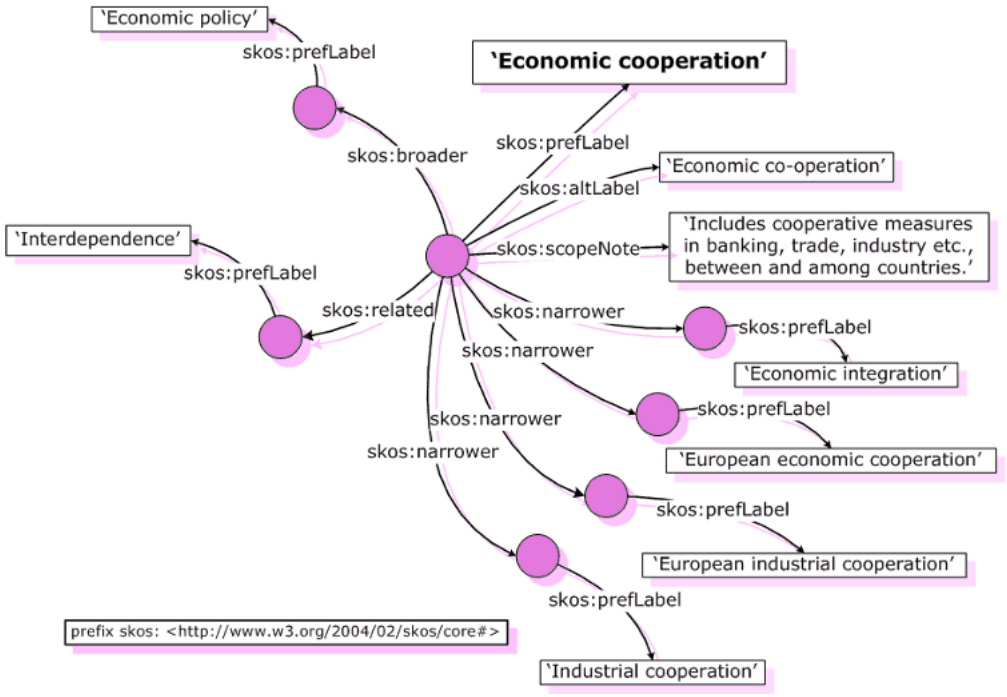
Sans aller plus loin dans la présentation des langages permettant de construire des ontologies, nous pouvons maintenant comprendre leur fonctionnement ; mais afin de mesurer leur utilité et leur intérêt, il est bon d'évoquer quelques applications dans des domaines distincts.

FOAF, que nous avons déjà mentionnée, est une brique essentielle du web de données, dans la mesure où elle fournit un vocabulaire pour décrire des agents et leurs relations. Si l'on en croit T. Berners-Lee, son utilisation rend possible le rêve de réseaux sociaux indépendants, ouverts, où les internautes ne seraient plus les otages de plate-formes commerciales qui détiennent leurs données et interdisent de relier les différents profils qu'ils ont créés en divers points de la toile [52] : grâce aux outils de recherche sémantique, il est possible de relier toutes les contributions d'une même personne, et il existe déjà des logiciels qui pourraient contribuer à la création de réseaux sociaux simplement à partir de pages web. Le site Advogato.org, notamment, constitue une communauté de travail collaboratif sur le développement du logiciel libre où les participants échangent dans le cadre d'une procédure de certification mutuelle : FOAF permet de vérifier l'authenticité des interventions – en fournissant d'identification, comme la signature cryptée de l'adresse mail – et permet de relier les compétences et les centres d'intérêts autour de projets de logiciels.

SKOS (Simple Knowledge Organization System) est une autre ontologie très utilisée, dédiée à la description de vocabulaires contrôlés. Pour décrire le réseau conceptuel d'un thésaurus, par exemple, SKOS fournit des propriétés comme `broader` (concept générique), `narrower` (spécifique), `related` (associé). De même, pour les termes traduisant les concepts, on identifie les formes préférées et rejetées

52. Cf. annexe : un *Giant Global Graph* → 6.3. Berners-Lee mentionne en particulier cette "déclaration des droits du web social ouvert" `http://opensocialweb.org/2007/09/05/bill-of-rights/`. En ce sens, le web de données est un prolongement du web 2.0.

au moyen des propriétés `prefLabel` et `altLabel`. D'où le graphe ci-dessous [53]



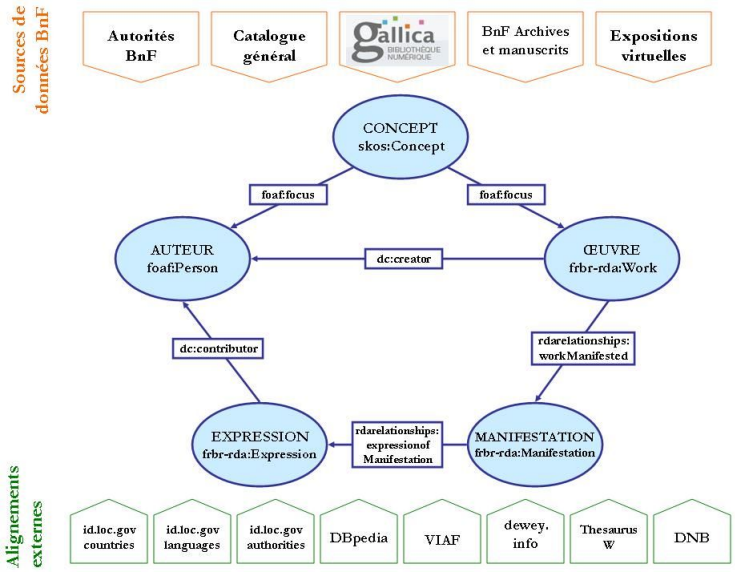
Mais cette description prend tout son sens dans la possibilité de donner des équivalences avec des concepts ou termes d'autres référentiels, avec plus de précision que `owl:sameAs` ou `rdfs:seeAlso` [54] : en effet, cela ouvre la porte au multilinguisme [55], et à toutes sortes de réutilisations de vocabulaires. Signalons par exemple la plate-forme Isidore : destinée à collecter ainsi qu'à communiquer les (méta)données de la recherche en sciences humaines parmi de nombreuses bases de données (plus de 2000 sources regroupées en 79 collections, plus de deux millions de ressources), elle exploite de nombreux référentiels (RAMEAU,

53. Tiré d'une présentation du Thésaurus pour l'indexation des archives locales : [SIAF,].

54. Il faut remarquer que les relations de similarité définies par `exactMatch` et `broadMatch` ne sont pas transitives, contrairement à `owl:sameAs`, comme le précise la présentation du W3C : [W3C, 2009](Mapping Concept Schemes). Les correspondances sont également plus précises que `rdfs:seeAlso` : il est possible de classer un concept appartenant à un vocabulaire dans un autre référentiel en précisant si les relations sont génériques ou associatives (`narrowMatch`, `broadMatch`, `relatedMatch`).

55. Voir par exemple la présentation des expérimentations sur RAMEAU de la BnF et LCSH de la Bibliothèque du Congrès : `http://rameau.bnf.fr/informations/projint.htm#macs`.

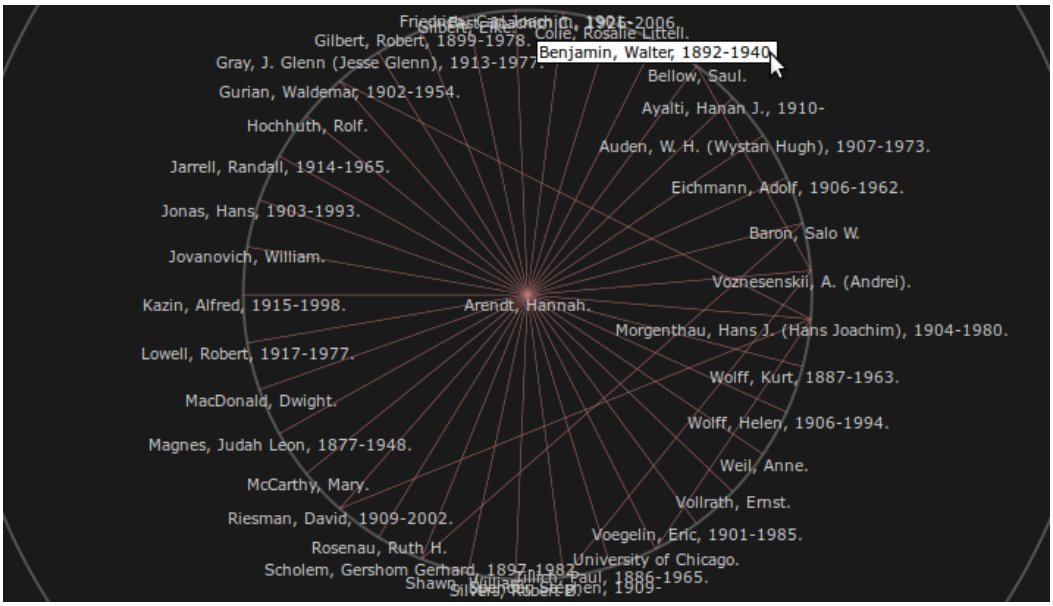
PACTOLS, Catégories OpenEdition, Disciplines HAL) qui fournissent des facettes dynamiques pour orienter la navigation via le moteur de recherche [56]. Autre exemple d'utilisation de SKOS : le projet data.Bnf.fr [57] fait correspondre, « aligne » les sujets de son vocabulaire RAMEAU avec des jeux de données issus de domaines hétérogènes (par exemple, le Thésaurus pour l'indexation des archives locales, la base de noms de lieux Geonames, le thésaurus Agrovoc) et d'institutions non francophones (Library of Congress, Deutsche Nationalbibliothek). Pour les notices d'auteurs, les données sont liées au Fichier d'Autorité International Virtuel (VIAF, *Virtual Identifier for Authority File*), au référentiel du Système Universitaire de Documentation (SUDOC – IDRef), et même à DBpedia. Il faut préciser cependant que SKOS n'est qu'un élément du modèle de données, dont nous reproduisons le schéma simplifié ci-dessous afin de donner à voir l'interconnexion des ontologies et des jeux de données.



56. Isidore fait partie de la Très Grande Infrastructure de Recherche Huma-num ; pour plus de détails sur l'utilisation des référentiels et le moissonnage, voir [Capelli *et al.*, 2012]. Le jeu de données peut également être interrogé avec SPARQL.

57. Cf. [data.bnf.fr, 2012].

Le projet SNAC enfin (*Social Networks and Archival Contexts* [58]), auquel participent d'ailleurs la BnF et les Archives nationales, nous paraît particulièrement représentatif de ce que les ontologies rendent possible en termes d'interconnexion de sources de données hétérogènes et d'exploitation de la description des relations entre ressources. Cette plate-forme destinée à donner accès à des informations biographiques et bibliographiques permet en effet de rebondir de lien en lien entre des données d'archives, de bibliothèques, de musées, de manière à reconstituer le milieu socio-culturel des personnes. L'objectif est de donner accès directement à la description des personnes, sans avoir à réaliser une fouille complexe de la description des ressources archivistiques. Ainsi, des propriétés comme `correspondedWith` (a correspondu avec) ou `appearsWith` (apparaît avec) permettent de représenter et d'analyser le réseau des relations sociales et intellectuelles d'un individu, soit par l'utilisation du point d'accès SPARQL, soit en naviguant de proche en proche l'interface de navigation très intuitive reproduite ci-dessous.



58. Voir [SNAC Project, 2010].

3 Le *Giant Global Graph*

> The problem with trees : Many systems are organised hierarchically. The CERNDOC documentation system is an example, as is the Unix file system, and the VMS/HELP system. A tree has the practical advantage of giving every node a unique name. However, it does not allow the system to model the real world.
>
> T. Berners-Lee, *HyperText and CERN* (1989).

> Les systèmes arborescents sont des systèmes hiérarchiques qui comportent des centres de signifiance et de subjectivation, des automates centraux comme des mémoires organisées. C'est que les modèles correspondants sont tels qu'un élément n'y reçoit ses informations que d'une unité supérieure, et une affectation subjective, de liaisons préétablies. On le voit bien dans les problèmes actuels d'informatique et de machines électroniques, qui conservent encore la plus vieille pensée dans la mesure où ils confèrent le pouvoir à une mémoire ou à un organe central.
>
> G. Deleuze et F. Guattari, *Rhizome* (1976).

La présentation des technologies du web de données a montré que la sémantique manipulée par les machines repose sur un empilement de couches de formalismes : s'il est possible de relier des données structurées selon le principe de l'hypertexte, c'est parce que dans les descriptions formées par les triplets on utilise comme éléments des identifiants de ressources conformes aux exigences du protocole HTTP. Grâce à ce principe très simple et aux langages dédiés à la description des relations entre instances, classes et propriétés, on construit ensuite des vocabulaires de métadonnées ou de valeurs [59] qui servent eux-mêmes à typer les ressources. Il s'agit donc plus de représentation des connaissances que d'intelligence artificielle, et cela suppose une adoption très large des standards, ainsi que

59. Un thésaurus par exemple donne un réseau de termes et de concepts qui vont permettre d'indexer des ressources par leur sujet : c'est un système d'organisation de connaissances, il stocke des informations et ne se contente pas de donner une structure de description. Un vocabulaire de métadonnées en revanche exprime seulement la différence entre divers champs de description (qui vont renseignés être par les éléments des vocabulaires de valeurs) : 'auteur' et 'sujet' sont des facettes distinctes de la même ressource. Cette distinction est proposée par le LLD Incubator Group : [W3C, 2011] (*value vocabularies* | *metadata element sets*, Annexe A).

le partage et la réutilisation des vocabulaires. Mais l'interconnexion de données issues de domaines hétérogènes et la normalisation des pratiques de description soulève un certain nombre de difficultés : comment concilier la logique du web et son architecture décentralisée avec la logique des collections et des institutions patrimoniales ?

3.1 Difficultés

3.1.1 Modéliser = réduire

Le projet de web sémantique a pu soulever des critiques, parfois virulentes, venant à la fois des communautés du web et des communautés du monde de la culture. Outre la confusion avec le traitement automatique de la langue (→ 2.3.1), l'utilisation des URI et des ontologies peut sembler supposer un accord sur une représentation unifiée du monde, sur une modélisation impliquant une certaine idéologie.

Le web s'est développé de manière "anarchique", sans suivre un plan préalable et sans être dirigé par une organisation centrale : tout le monde peut s'y exprimer, ce qui pose d'ailleurs parfois problème aux institutions politiques. Le World Wide Web Consortium n'est pas un centre de contrôle des contenus, il s'occupe seulement des questions techniques et des formats. Nous nous sommes habitués à cette liberté de communiquer par-delà les frontières et de tisser des liens entre acteurs d'une même communauté d'intérêts, liberté d'auto-organisation qui caractérise le web 2.0. La question n'est pas ici de savoir dans quelle mesure cette conception est illusoire, mais d'abord de comprendre en quoi le web de données peut être perçu comme une tentative d'uniformisation.

Si l'on compare l'espace documentaire du web avec celui d'une bibliothèque, l'opposition est radicale. On accède aux documents du web par un moteur de

recherche qui indexe le texte des documents, ou bien on s'oriente grâce aux liens et aux « tags » librement apposés par les internautes sur les pages de leurs blogs et même sur les documents qui sont mis en partage par d'autres : cette indexation collaborative donne lieu aux systèmes de classement nommés « folksonomies ». Lorsque l'on utilise un catalogue de bibliothèque, au contraire, le moteur de recherche renvoie à des notices, et non au contenu du texte ; ces notices sont codifiées, dans le sens où l'on utilise un vocabulaire contrôlé pour décrire le document, et elles renvoient à un système d'organisation des connaissances le plus souvent hiérarchique, arborescent.

Dans un tel système, les catégories obéissent à un principe de subsomption et de dichotomie, comme dans une taxinomie biologique : les mammifères appartiennent à la classe des vertébrés et se distinguent des reptiles et des oiseaux, et il est impossible de ranger un moineau à la fois dans la classe des oiseaux et dans celle des mammifères. Cela ne paraît pas poser de problème pour les espèces vivantes, mais l'application au domaine de l'information est critiquable dans la mesure où elle peut avoir plusieurs usages qu'il est impossible d'anticiper. C'est l'argument que développe C. Shirky dans un billet sur les limites des ontologies, qui seraient surévaluées selon lui [60]. L'objection repose sur le constat qu'une ontologie ou une taxinomie sert à catégoriser à l'avance l'information et l'enferme dans un système de classement figé. Shirky donne en exemple la disparition des annuaires comme Yahoo et DMOZ. Cette logique qui consiste à faire entrer une œuvre dans une case correspond bien à l'âge des collections physiques, où la taille et la localisation des étagères est un paramètre essentiel [61] mais elle atteint

60. Voir [Shirky, 2005].

61. Voir aussi E. Bermès, Petite histoire des classifications : « Plus le modèle est simple, plus il est efficace pour remplir son but premier : la "mise en espace", ranger des livres. Dans un but pratique (le libre accès) on va donc privilégier un modèle médiocre avec une notation simple à un modèle performant avec une notation complexe. » → `http://www.figoblog.org/node/803`.

vite ses limites. D'où l'existence de systèmes de renvois et d'alias, pour pouvoir faire figurer, par exemple chez Yahoo, la littérature dans la catégorie 'loisirs', alors qu'elle n'y figure pas "vraiment". Ce parti pris est le signe que les concepteurs de Yahoo n'avaient pas compris que sur le web, il n'y a plus d'étagères, et que l'absence (ou presque) de contraintes physiques permet de donner plusieurs accès à la même ressource. Mais le choix est aussi révélateur d'un jugement de valeur, qui donne plus de prix au savoir et aux arts qu'au divertissement. Shirky conclut que les ontologies devraient être réservées à des domaines très spécifiques, où la modélisation est efficace, où l'aspect technique du sujet et la complexité des modèles garantissent une absence de parti pris.

Une objection du même type a été avancée par K. Sparck Jones [62] qui compare le projet du web sémantique à la tentative métaphysique de Leibniz d'élaborer une « caractéristique universelle », un discours formel qui serait le parfait reflet du monde et permettrait de raisonner sur n'importe quel objet avec une certitude mathématique. Sparck Jones pose la question suivante :

> Il y a là une supposition implicite. C'est qu'un seul modèle du monde suffit à de multiples tâches, et qu'il existe un tel dispositif de raisonnement, exactement comme il y en a pour les humains. En fait, on prend l'accès à l'information pour une seule et même tâche générique. Mais il y a la recherche de texte (*text retrieval*), le filtrage de document, les questions-réponses, l'extraction d'information, la sélection de passages, les résumés, les requêtes sur des données... Tout cela ne serait-il que les variations d'un même processus, tout cela se résoudrait-il en une modélisation commune du monde et des hypothèses de raisonnement ?

Il y a en effet une naïveté évidente à croire que tout se laisse mettre sous forme de syllogismes, dans la mesure où l'on oublie les limites de la déduction et les incohérences de toute tentative de généralisation [63].

62. Dans l'article *What's new about the Semantic Web ? Some questions* (auquel répond T. Berners-Lee, nous le verrons) : [Sparck-Jones, 2004].

63. Voir également le billet très ironique de C. Shirky, qui reprend les jeux de logique de L. Carroll : [Shirky, 2003].

Les critiques que nous avons évoquées ont en commun d'accuser le web sémantique d'être « antisocial » [64], dans le sens où une structure réductrice viendrait s'appliquer d'en haut (*top-down*) à un milieu auto-organisé et horizontal, où les structures émergent de la base (*bottom-up*). Mais une inquiétude symétrique inverse est également exprimée, cette fois dans les milieux des institutions : dans l'interconnexion des données, il y a un risque d'appauvrissement des structures. K. Coyle, par exemple, soulève le problème de la réutilisation de FOAF par les bibliothèques : ce qu'une notice de bibliothèque entend par auteur et personne n'a rien à voir avec l'entité `foaf:Person` [65]. En effet, FOAF suppose qu'il suffit d'ajouter au nom de la personne un ou plusieurs identifiants de type login pour distinguer les individus, le nom n'est pas un identifiant ; au contraire, dans les données d'autorité de bibliothèques, les informations sur le nom sont plus riches et collectées dans le but de réunir les différentes formes qu'il peut prendre sous un seul enregistrement. Coyle fait également remarquer l'importance des standards dans l'organisation d'une bibliothèque : les structures sont lourdes et rigides parce qu'il s'agit de centraliser des notices produites dans le cadre d'une division du travail très complexe. Les contextes d'usage des vocabulaires sont donc très différents, et il n'est pas anodin de connecter les descriptions issues des bibliothèques avec celles provenant des archives ou des musées. Le modèle des bibliothèques est conçu pour que les données sur un auteur puissent resservir dans plusieurs notices bibliographiques ; la description archivistique, suivant le principe de respect des fonds, est centrée sur les concepts de contexte, d'organicité et de hiérarchie. Quant aux musées, en plus de la considération de l'unicité de la pièce, c'est le concept d'événement qui domine : il faut pouvoir rendre compte des différentes phases que l'objet traverse (exposition, restauration...) [66].

64. Expression empruntée à F. Gandon : [Gandon, 2006]
65. Voir [Coyle, 2010].
66. Nous reprenons l'analyse d'E. Bermès : [Bermès, 2011].

3.1.2 La crise identitaire

Une autre difficulté majeure est soulevée par le passage du web de documents au web de données : le fait que les URIs ne désignent plus seulement des pages web mais également n'importe quel type d'entité conceptuelle ou réelle ne va pas de soi.

La critique développée par le groupe Pédauque revient sur les problèmes de réductionnisme évoqués ci-dessus mais va plus loin [67], en s'attaquant aux présupposés linguistiques du formalisme de RDF et des ontologies. À travers l'omniprésence des métadonnées ainsi que la volonté de développer des ontologies générales et multilingues [68], le web sémantique aboutirait à substituer aux textes leur représentation formelle. Le cœur de l'argumentation consiste à rappeler qu'un texte est un ensemble complexe, qui ne se laisse pas réduire à une série de phrases et encore moins à un descripteur : le primat de la phrase et du mot en linguistique remontent à une tradition qui rejette la rhétorique pour se concentrer sur les questions d'analyse de la proposition, de citation et de référence ; mais ce primat est contestable et contesté [69]. En donnant la priorité aux idées sur leur formulation, en supposant que les concepts sont indépendants des termes et du contexte, on s'enferme dans une vision utopiste et inquiétante où la diversité des langues et des dialectes ne signifie plus rien. Dans cette vision, plus de place pour l'expression idiomatique, tout serait traduisible sans perte et sans ambiguïté, le signifié prenant

67. Roger T. Pédauque, *Le texte en jeu* : [Pédauque, 2005].

68. Pédauque mentionne en particulier les tentatives de construire des ontologies « de haut niveau » (i.e. non spécifiques) comme Wordnet et Cyc, qui sont comparées au système aristotélicien – tentative pour organiser l'ensemble des connaissances et des domaines de la réalité selon quelques principes métaphysiques ; → `http://www.cyc.com/kb` et `http://wordnetweb.princeton.edu/perl/webwn`. Voir également annexe pour des formes simplifiées d'une ontologie de haut niveau et d'une spécifique : → 6.4.

69. Par des « théories qui s'attachent à étudier les structures argumentatives des textes, les interactions et actes du langage. Elles relèvent de théories communicationnelles et, en cela, d'approches fonctionnalistes caractérisant les textes par des critères externes (but, interactions typiques, etc.). Elles décrivent des normes de production textuelle dans des espaces sociolinguistiques identifiés. »

totalement le pas sur le signifiant [70]. La tentative du web sémantique serait donc d'une ambition démesurée, puisque toute chose devrait être décrite (« grammatisation »), et ce dans une inter-langue minimale, comme le « Global English » et ses 850 mots – les autres langues étant naturellement vouées à disparaître ou à en devenir des dialectes. Le concept de « grammatisation » utilisé par Pédauque donne une première approche : le risque est de confondre les mots et les choses, de s'enfermer dans une construction mentale qui se substitue à la réalité.

Le problème de l'ambivalence des URIs reste toutefois implicite, il est maintenant nécessaire de présenter les débats auxquels il a donné lieu, et de préciser l'analyse en étudiant le double statut des identifiants. Sachant que l'URI peut à la fois pointer vers une ressource informationnelle – comme un document, une information avec son support –, et vers une chose du monde, un fait réel ou une représentation mentale, la question est de savoir comment éviter que l'ambivalence soit une ambiguïté.

Nous avons également déjà remarqué qu'une URI a deux fonctions [71] : elle sert à la fois de référence (c'est-à-dire de signifiant qui renvoie à une ressource, un signifié) et de support pour des assertions (puisqu'elle s'articule avec d'autres URIs dans les triplets). Essayons de préciser l'ambivalence : d'un côté l'URI renvoie à une ressource, et dans cette mesure elle n'a de signification que par cette ressource qu'on peut lui substituer ; d'un autre côté, lorsqu'elle joue le rôle de sujet dans un triplet, elle est elle-même ce qui est désigné, sans quoi les programmes ne pourraient "raisonner".

Dans le premier cas elle est une adresse, elle acquiert une signification en donnant accès à une représentation de la ressource informationnelle. Elle permet d'afficher une page dans un navigateur par exemple : elle fonctionne comme un

70. Selon la distinction de Saussure : schématiquement, le signifié est la représentation mentale d'un concept, et le signifiant la forme perceptible du signe (le son) qui renvoie à ce signifié.

71. Voir la présentation de RDF et des URIs : → [2.2.2].

mot, dont le son s'efface devant le sens du discours dans lequel il est pris. Dans le deuxième cas elle fonctionne comme un concept : sa signification lui vient de sa position dans un réseau sémantique plus ou moins complexe, à commencer par le contexte du triplet. Comment distinguer les deux fonctions ? H. Halpin [72], présente le problème de manière grotesque :

> Sur le web sémantique, puis-je faire une assertion au sujet de T. Berners-Lee en faisant une assertion à propos de sa page perso ? S'il utilise une URI distincte pour lui-même, se doit-il de fournir une représentation de lui-même ?

Pour reprendre notre exemple du livre sur Foucault, le risque de confusion est le même qu'entre "Foucault a écrit sur les prisons" et "*Foucault* a été écrit par Deleuze".

L'ensemble des données liées sous forme de triplets, censé constituer un réseau de connaissances et pas seulement de documents, n'est-il pas menacé d'écroulement si l'on ne peut jamais clairement dire qu'une URI désigne une ressource informationnelle ou pointe directement une chose ? C'est la question que soulève H. Halpin et qui a été largement débattue, au point qu'on a cherché à utiliser un autre type d'identifiants que les URIs-URLs [73].

72. À qui nous empruntons le titre de cette section. Cf. [Halpin, 2006].

73. Nous avons vu que le deuxième et le troisième principe du *linked data* recommandent l'utilisation des URIs HTTP, i.e. des URLs, qui permettent de localiser et représenter les ressources. Pour être précis, il faut ajouter que le document qui s'affiche sur notre écran n'est pas lui-même une ressource, c'est une représentation construite par les logiciels. Les URNs (*Uniform Resource Name*) sont l'autre type d'URIs qui a été proposé : ces identifiants étaient destinés à être indépendants des représentations de ressources, donc des opérations logicielles effectuées pour les manipuler. Cette indépendance devait leur donner une plus grande stabilité, une pérennité. Comme le dit Halpin, il a été proposé d'utiliser les URNs pour désigner les entités non informationnelles et ainsi échapper à l'ambiguïté. Pour une présentation plus détaillée voir [Calderan *et al.*, 2012] chap. 9.

3.2 Questions d'architecture

Ces critiques ont été entendues, et l'on peut considérer que la proposition du terme « web de données » pour rebaptiser le projet du web sémantique est le résultat d'un effort pour préciser le concept et pas seulement une volonté d'éviter la confusion avec l'acception de "sémantique" renvoyant au TALN [74]. L'analyse des réponses qui ont été données montre que les objections proviennent le plus souvent d'une absence de prise en compte de la nouveauté radicale du modèle de données fourni par le cadre de description RDF : sa structure en graphe [75] et sa souplesse ont été sous-évaluées, de la même manière que l'on oublie que le web sémantique n'est rien d'autre qu'une extension du web et possède les mêmes propriétés architecturales.

3.2.1 URIs et parallélisme architectural

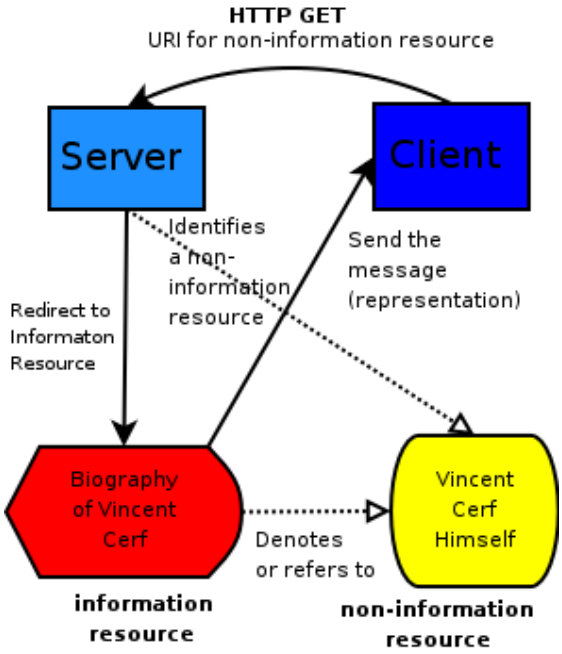
Le problème de la crise identitaire est le plus urgent à résoudre, puisqu'il pointe un risque d'inconsistance du projet lui-même : le web sémantique, qui est censé fournir des moyens techniques pour lever les ambiguïtés ne ferait que les déplacer du sens des termes vers le statut des identifiants. C'est également la question la plus rapide à traiter, puisqu'il suffit de l'écarter. En effet, les interrogations que nous avons présentées sur l'ambivalence des URIs touchent deux dimensions distinctes, l'une concernant l'architecture du web et l'autre la métaphysique. Halpin cite la position pragmatique de T. Berners-Lee :

> Nous ne sommes pas en train d'analyser un monde mais de le construire. Nous ne sommes pas des philosophes expérimentaux mais des ingénieurs philosophes.

74. Nous avons déjà évoqué ce malentendu : → [1] ; mais ce n'est qu'une des raisons de l'évolution. Voir l'article de Berners-Lee, Hall et Shadbolt qui recentre le projet sur la mise à disposition et la réutilisation des données ([Berners-Lee *et al.*, 2006]). Une présentation synthétique de cette évolution a été proposée par G. Poupeau : [Poupeau, 2008a].

75. L'arbre est une forme simple de graphe, mais nous utiliserons le terme de graphe en l'opposant aux structures arborescentes.

Une solution architecturale a été donnée, dont le principe est relativement simple : si un utilisateur envoie une requête pour obtenir une représentation de ressource alors que l'URI (1) désigne une entité non informationnelle mais conceptuelle, soit il est redirigé vers une autre URI (2) qui lui fournit une représentation, soit il reçoit un code d'erreur. La page que l'internaute voit s'afficher est la représentation d'une ressource qui existe parallèlement et indépendamment de la ressource conceptuelle : elle est censée l'expliciter mais peut très bien être dépassée. Par exemple, si l'URI (1) identifie Vinton Cerf, l'utilisateur peut trouver dans la page construite à partir de l'URI (2) l'affirmation "V. Cerf est étudiant à l'Université de Californie" alors que V. Cerf est à ce moment-là employé par Google [76]. Les ressources non informationnelles sont donc destinées aux machines et se situent sur un autre plan, comme le montre le schéma de Halpin :



L'autre dimension du problème est métaphysique, en ce qu'elle concerne la nature de l'information et de la ressource sur le web, ce qui présuppose de s'in-

76. Voir aussi l'exemple de DBpedia plus haut : → [2.2.2]. Ce paragraphe est une paraphrase de l'article de Halpin précité.

terroger sur les relations entre la pensée, le langage, le document, le numérique. Comme le dit Halpin, la question réactualise de vieux débats philosophiques que nous ne sommes pas près de clore : mais est-ce nécessaire pour construire le web de données ? Pour commencer, on peut remarquer que le web de données ne fait que révéler une difficulté qui est consubstantielle au web lui-même : ce qui est difficile à penser, c'est le statut de la ressource elle-même bien plus que celui de l'identifiant. La ressource est en effet un concept abstrait : nous ne manipulons que des représentations qui sont construites par les navigateurs (entre autres) et non les ressources elles-mêmes, on peut les comparer à des ombres [77]. Mais surtout, on peut se demander si cette ambivalence des identifiants et des ressources n'est pas la même que celle du signe et du concept en général. La relation qui associe en un seul signe une face signifiante et une face signifiée est aussi inexplicable et évidente que la relation entre l'esprit et le corps. Le fait qu'un concept puisse à la fois avoir un contenu informationnel (i.e. représenter la nature d'un objet) et avoir une forme (être lui-même l'objet d'autres concepts qui le définissent formellement) est une propriété de la pensée dont il n'est peut-être pas nécessaire de chercher à rendre compte mais qu'on peut considérer comme une donnée de l'expérience [78]. C'est bien ce que la psychologie puis la linguistique ont fait, et c'est aussi le sens de la démarche pragmatique d'« ingénierie philosophique » prônée par Berners-Lee.

Par ailleurs, ce parallélisme ne signifie pas forcément que l'on chercherait à construire un monde d'idées-URIs au-dessus de la réalité. Selon l'analyse qui

77. Cf. A. Monnin *Les ressources, des ombres récalcitrantes* : [Monnin, 2013].

78. Pour n'en citer qu'une seule, on peut rappeler l'opposition entre Descartes et Spinoza : Descartes considère que la matière et l'esprit sont deux substances distinctes, et que l'union de l'âme et du corps est évidente en même temps qu'incompréhensible. Spinoza fait l'hypothèse que l'esprit et la matière sont deux attributs (expressions) de la même substance ; sans chercher à expliquer la relation entre les deux attributs, il propose l'analogie avec la relation entre l'idée et l'objet qu'elle représente. Cette relation est également la même que la différence entre ce que nous avons appelé contenu informationnel (essence objective) et forme du concept (essence formelle). Voir Descartes, *Méditations métaphysiques*, VI et Spinoza, *Éthique* II, 5, 7 et sqq.

précède, on peut dire qu'une URI n'a jamais de sens dans l'absolu mais seulement en tant que *nœud d'un réseau* sémantique plus ou moins complexe ou en tant qu'adresse pour des représentations variables. La critique de Pédauque semble supposer que la fonction des ontologies est d'assigner un sens déterminé à un terme afin de pouvoir l'utiliser pour décrire les ressources auxquelles il serait substitué [79]. Effectivement, en prenant le modèle d'ontologies de haut niveau ou de vocabulaires très complets comme Wordnet, mentionné par Pédauque, on peut croire que les technologies sémantiques proposent un modèle de traduction mot à mot, une équivalence terme à terme avec la réalité. Mais on peut se demander si cette conception ne doit pas être précisée en rappelant la différence entre les ontologies proprement dites et l'usage que l'on en fait dans les jeux de données ou bases de connaissances.

En effet, les ontologies sont des modèles formels qui permettent de manipuler des ensembles de connaissances en les catégorisant par des classes, des propriétés et des relations. Les données de Wordnet sont au contraire des ensembles de connaissances qui ont été organisées selon un modèle ontologique – éventuellement contestable. Par exemple, Wordnet va établir une relation de généralisation (hyperonymie) entre 'marronnier' et 'arbre'. Mais cette information n'est pas au même niveau que la proposition : "la propriété 'hyperonyme' est l'inverse de 'hyponyme'". Autrement dit, la critique nous semble porter sur l'usage de vocabulaires de valeurs [80], de vocabulaires contrôlés, et donc sur le contenu informa-

79. « La réponse du Web sémantique est d'indexer les textes avec les concepts d'une ontologie partagée par une large communauté. La recherche se fait alors sur les concepts avec lesquels les textes sont indexés. Cet index est une représentation formelle du document. Cette représentation se substitue au texte pour la phase de recherche et même plus : le texte est enrégimenté à la représentation formelle (...). Mais le texte laisse ouvert toute la richesse de l'interprétation tandis que l'ontologie, en privilégiant un (ou deux au plus) contexte, bloque toute autre capacité d'interprétation de l'homme comme de la machine. C'est le présupposé du Web sémantique et d'autres applications des ontologies dans une recherche d'automatisation des traitements. » Pédauque, *op. cit*.

80. Sur ce concept, voir note plus haut : → 59. Il nous semble que l'expression *indexer les textes avec les concepts d'une ontologie* (cf. citation dans la note précédente) montre que la critique ne

tionnel des jeux de données plutôt que sur le modèle formel, à savoir l'ontologie vide [81], qui organise les connaissances. Cette différence nous paraît importante dans la mesure où l'on peut faire coexister plusieurs accès aux ressources puisque les données et l'organisation peuvent être dissociées.

Nous posons donc les questions suivantes : si l'on conteste la possibilité de définir le sens d'un terme en arguant du fait que le contexte change toujours le sens des mots, ne faut-il pas aussi renoncer à l'usage des dictionnaires et rejeter, en même temps que le projet du web de données, toute tentative d'ordre encyclopédique ? Et si c'est réellement l'utilisation d'ontologies que l'on conteste, veut-on dire que la connaissance devrait se passer de modéliser le monde, ou que les technologies sémantiques ne sont pas neutres et imposent une vision du monde sans le dire et sans laisser d'alternative ? C'est sans doute ce dernier point qui pose problème, et il faut maintenant reprendre l'étude de l'application des ontologies dans le web de données tel qu'il s'est développé depuis le texte de Pédauque, et analyser ses principes architecturaux afin de savoir si « l'ontologie, en privilégiant un (ou deux au plus) contexte, bloque toute autre capacité d'interprétation de l'homme comme de la machine [82]. »

3.2.2 Arbre et rhizome

Reprenons l'affirmation selon laquelle le web de données n'est qu'une extension [83] du web et non un autre web : cela signifie que les propriétés architecturales sont les mêmes, et qu'il n'y a pas plus de risque de voir s'y établir la domination

porte pas sur les ontologies en général.

81. On parle parfois de « peuplement des ontologies » pour signifier le fait que des entités sont des instances d'une classe : [Aussenac-Gilles, 2012]. On peut donc considérer les vocabulaires de valeurs comme des ontologies peuplées.

82. Un autre texte de Pédauque a été entrepris, où les critiques sont plus nuancées : [Pédauque, 2011].

83. « Le web sémantique n'est pas un web séparé mais une extension du web actuel, dans laquelle l'information reçoit un sens bien défini, rendant possible une meilleure coopération entre les hommes et les machines. » Berners-Lee et *alii*, *op. cit*.

d'un modèle unique que dans le web 2.0. C'est aussi la raison pour laquelle le web de données a pu être rebaptisé *Giant Global Graph* [84]. Le web se développe selon une logique imprévisible, celle des échanges sociaux et des flux d'informations :

> Le web est une architecture décentralisée car elle ne s'appuie pas sur un seul serveur principal dont dépendrait l'ensemble du réseau, mais sur plusieurs serveurs répartis. La disparition d'un serveur ou l'ajout d'une nouvelle machine ne déstabilise pas l'ensemble du réseau. C'est la première condition de la construction du web « par le bas » (*bottom-up*) : il se construit par l'agrégation des informations qui sont publiées, et non suivant un plan prédéfini [85].

Il s'agit maintenant de montrer que le modèle de données fourni par RDF [86], et sur lequel sont construites les ontologies, non seulement apporte de la structure sans imposer un cadre unique, mais bien plus, donne une nouvelle souplesse en ce qu'il permet de faire coexister plusieurs points de vues sur les mêmes données, plusieurs contextes d'usage.

Un premier point technique notable est l'extensibilité des vocabulaires ou modèles de données que l'on formalise en RDF. Comme les modèles peuvent être étendus à volonté et ne sont pas figés, on peut toujours enrichir un vocabulaire qui serait trop réducteur. Nous avons déjà vu que la structure d'une base de données est rigide, dans la mesure où la création de nouvelles relations peut avoir des incidences sur l'ensemble [87]. Dans le cas du formalisme RDF, on ajoute autant de triplets que l'on veut, par exemple en réutilisant une partie d'une ontologie définie ailleurs (cf. le cas de FOAF dans le modèle de la BnF ; nous y reviendrons).

Autre nouveauté architecturale, cette fois par rapport aux schémas XML : les triplets sont indépendants d'une organisation d'ensemble, les fichiers au format RDF ne nécessitant pas d'être validés. Qui plus est, on sort du simple cadre des structures arborescentes pour passer à un type de graphes réticulaires, que Deleuze

84. Cf. plus haut : → 6.3.
85. Cf. [Calderan *et al.*, 2012], chap 3.
86. RDF peut être considéré à la fois comme un modèle de données, comme un cadre d'échange (data interchange) et de description, et un vocabulaire. Cf. [Bergman, 2009a].
87. Cf. *supra* : → 2.2.1.

et Guattari appellent « rhizome » :

> Résumons les caractères principaux d'un rhizome : à la différence des arbres ou de leurs racines, le rhizome connecte un point quelconque avec un autre point quelconque, et chacun de ses traits ne renvoie pas nécessairement à des traits de même nature, il met en jeu des régimes de signes très différents et même des états de non-signes. Le rhizome ne se laisse ramener ni à l'Un ni au multiple. Il n'est pas l'Un qui devient deux, ni même qui deviendrait directement trois, quatre ou cinq, etc. Il n'est pas un multiple qui dérive de l'Un, ni auquel l'Un s'ajouterait (n+1). Il n'est pas fait d'unités, mais de dimensions, ou plutôt de directions mouvantes [88].

En effet, la structure d'un fichier xml est nécessairement celle d'un arbre : pour être « bien formé », le document doit comporter un élément ou nœud appelé racine, délimité par des balises ouvrante et fermante, et tous les éléments en sont des subdivisions qui elles-mêmes se divisent selon le même principe. Le type de document peut être défini par un schéma qui précise l'organisation de ces subdivisions et leur contenu. Soit par exemple la notice d'une œuvre : l'élément `<auteur>` contient un élément `<nom>` qui contient nécessairement – champ obligatoire – une valeur textuelle, un autre élément `<sujet>` contenant nécessairement une séquence de chiffres correspondant au numéro de la discipline dans la classification Dewey. L'avantage du schéma est de pouvoir fixer des contraintes, de telle sorte que l'on ne trouve pas de chiffres à la place du nom, et que l'on ait bien un nom d'auteur même si l'on n'a pas de sujet. Évidemment, le schéma peut renvoyer à un autre document XML contenant les valeurs de la classification, et l'on peut demander à un programme de contrôler l'existence des numéros entrés ainsi que d'importer le nom du domaine correspondant.

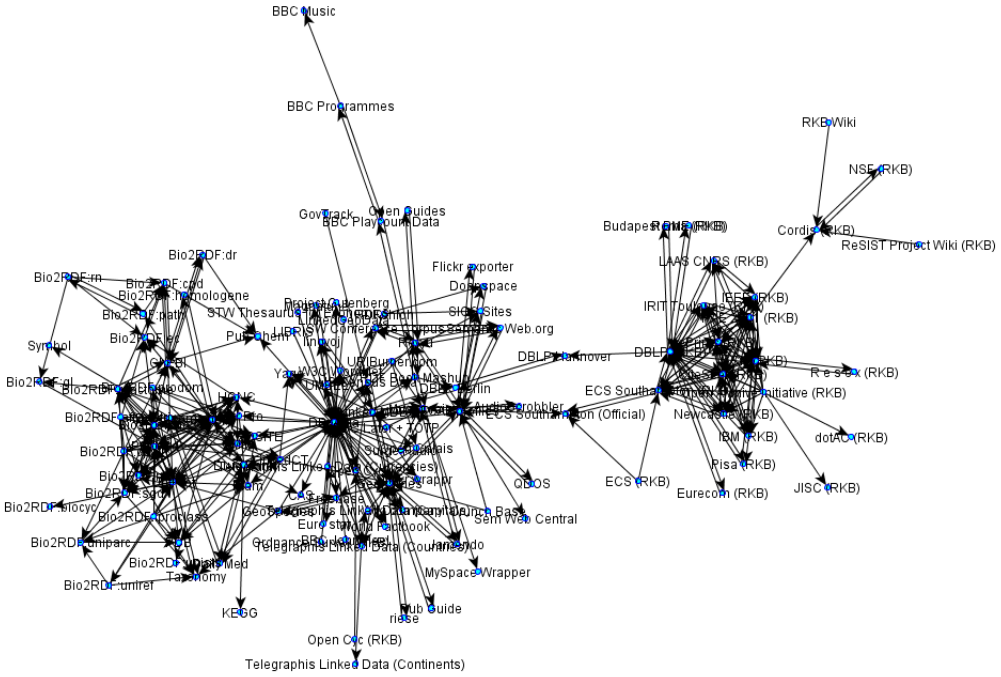
Par rapport aux bases de données, XML présente l'avantage d'être un format textuel et d'être interopérable (donc plus pérenne). En revanche, les schémas imposent un processus de validation qui fait que la moindre déviation par rap-

88. Deleuze et Guattari, *op. cit*. Le terme de rhizome a été repris par un site orienté web sémantique [GRIHO,] et par un logiciel de wiki sémantique : → `http://www.w3.org/2001/sw/Europe/events/foaf-galway/papers/pp/rhizome_position_paper/`.

port au plan génère des erreurs, et que l'on retrouve les mêmes lourdeurs pour faire évoluer un modèle de données. Enfin, XML n'est pas conçu pour exprimer des relations entre ressources et permettre de raisonner sur les liens entre URIs. Au contraire, RDF permet de multiplier les liens à volonté sans avoir à respecter un schéma ayant une racine et des subdivisions, ni même un plan de classement préalable en général : seules les contraintes logiques déterminent le tissage des liens. Or il semble que les objections présentées plus haut aient négligé cette nature rhizomatique du modèle de données RDF [89]. D'une part les relations entre les catégories ne se limitent pas à des relations de subsomption – nous avons vu l'exemple de relations d'équivalence, d'association, de définition. D'autre part, même dans le cas de la subsomption on peut ranger un même objet sous plusieurs concepts – à condition que les classes ne soient pas disjointes : c'est ainsi que la réutilisation des URIs est possible, notamment dans le cas des alignements de vocabulaires.

Toutefois, avant d'étudier plus précisément l'articulation des jeux de données et l'utilisation des vocabulaires, ces considérations architecturales peuvent être confirmées par un survol du web de données afin d'en appréhender la géographie : comme le montrait la carte ci-dessous en 2010, on pouvait déjà distinguer au moins trois webs de données.

89. Voir en particulier les réponses de G. Poupeau à Pédauque et à Shirky : [Poupeau, 2008b] et [Poupeau, 2006].



L'interconnexion des données ne doit pas être identifiée à une confusion des genres, puisque des communautés d'intérêt se forment, tout comme dans le web 2.0, et donc des « regroupements assez bien interconnectés, mais reliés par un très petit nombre de jeux de données pivots [90] ». La représentation classique du *Linked Open Data Cloud* permet également de voir apparaître des domaines formant des territoires relativement autonomes : données gouvernementales, données bibliographiques, médicales, géographiques [91]... Et si l'on considère souvent DBpedia comme une entité centrale, en se référant justement à cette représentation, il faut rappeler que ce sont des données issues de Wikipedia, projet collaboratif s'il en est, et que l'on peut toujours modifier les sources : en admettant que DBpedia soit une référence-pivot, il serait abusif de parler d'une hiérarchie.

90. A. Isaac commente ainsi la carte reproduite : [Calderan *et al.*, 2012] chap. 4 et → `http://blog.larkc.eu/?p=1941` pour plus de précisions.
91. Voir annexe : → 6.5. Les couleurs représentent les différents domaines.

3.3 Un monde ouvert

3.3.1 Interconnexion ≠ confusion

Bien que le web sémantique ait pu être considéré comme « antisocial », la préoccupation de ses concepteurs a toujours été de développer la logique du web et non un plan prédéfini : on peut remarquer d'ailleurs une récurrence de l'idée selon laquelle il est impossible de savoir quel usage sera fait des technologies et des données mises à disposition, et que c'est là justement tout l'intérêt de la réutilisation [92].

Le web de données est un monde ouvert, et il est notable que le tissage des liens qui l'a construit ait fait coexister et proliférer de nombreux vocabulaires, dont la complexité est très variable et qui se répartissent dans les deux catégories déjà évoquées, à savoir les jeux de métadonnées et les vocabulaires contrôlés. D'une part, il n'y a pas à craindre la prédominance d'un système de classification aux dépens des folksonomies, qui ont d'ailleurs leurs limites lorsqu'il s'agit d'annoter de très grands nombres de ressources [93]. D'autre part, on peut dire que la réutilisation des modèles est partielle et ne se traduit pas par un aplatissement des différences. C'est avec un petit nombre de relations d'équivalence que les passerelles se développent, et l'alignement d'un terme du *dataset* X avec un terme du *dataset* Y n'a pas forcément de conséquence sur l'ensemble des données ni ne signifie leur uniformisation. Hendler fait remarquer que le simple fait qu'un ensemble d'internautes décident d'employer un identifiant pour un concept donne la possibilité de le distinguer de concepts différents qui seraient représentés par le même terme et que cela n'empêche pas d'autres internautes d'utiliser un autre

92. Cf. par exemple la conférence de T. Berners-Lee en 2009 et l'image des fleurs : [Berners-Lee, 2009].

93. Cf. [Hendler, 2007]. Une application des technologies sémantiques pourrait justement être la gestion des systèmes d'annotation collaborative ; l'exemple pour les collections de documents vidéo est analysé sur le site du W3C : `http://www.w3.org/TR/2004/REC-webont-req-20040210/` 2.2.

identifiant pour le même concept et d'établir un lien ensuite. Prenant également l'exemple de FOAF, il explique qu'une simple équivalence entre deux *logins* sur des sites différents (équivalence établie grâce à l'adresse mail commune) permet de très nombreux rapprochements à l'échelle des millions de pages dispersées sur les sites de communautés scientifiques par exemple [94]. En d'autres termes, peu de sémantique permet beaucoup de liens utiles. Si l'on considère de plus près les passerelles établies entre les jeux de données et la réutilisation des URIs par les institutions patrimoniales, on se rend compte que les alignements sont réalisés avec beaucoup de prudence.

Même dans le cas de projets d'harmonisation de grande envergure, qui pourraient paraître les plus risqués, chaque modèle garde son autonomie. À ce titre, l'union des FRBR avec CIDOC-CRM est significative puisqu'il s'agit de rendre interopérables des cadres de description issus des bibliothèques et des musées. L'ontologie réalisée est le fruit d'une synthèse entre les recommandations FRBR pour les bibliothèques et du modèle conceptuel CRM destiné aux musées, le second fournissant un cadre plus large et qui permet, au prix de quelques modifications, l'articulation des deux domaines [95]. Les tenants du projet insistent sur le fait qu'il ne s'agit pas d'un abandon des spécificités des FRBR mais d'un changement de méthode, et que l'expression du modèle dans un autre formalisme permet de vérifier sa cohérence et de le clarifier.

Bien plus, la confrontation devient complémentarité : les entités bibliographiques, qui étaient considérées comme statiques et sorties de nulle part, reçoivent un contexte ; les descriptions muséographiques s'enrichissent de concepts leur per-

94. *Ibid*. Cf. aussi la vision de Berners-Lee du *Giant Global Graph* présentée plus haut : → 2.3.3.

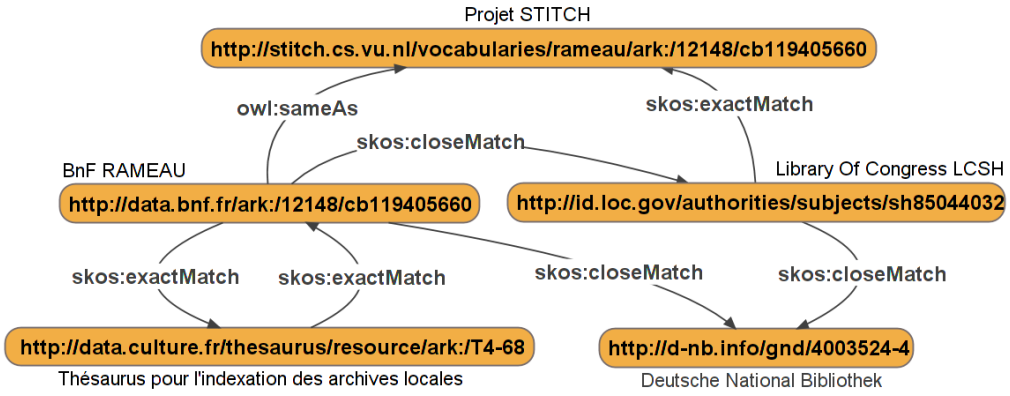
95. FRBR pour *Functional Requirements for Bibliographic Records*, Fonctionnalités requises des notices bibliographiques (FRBR) ; CIDOC pour Comité international de documentation et CRM pour Conceptual Reference Model : [ICOM-CIDOC,]. Le CIDOC-CRM pourrait également faire alliance avec le cadre de description archivistique EAD (*Encoded Archival Description*).

mettant de mieux décrire les objets produits en série par exemple, ou de catégories transverses pour appréhender la création artistique en général. Loin d'être un nivellement qui aplatit les différences, l'interconnexion des cadres de description apporte donc plus de précision. Pour ne prendre qu'un seul exemple, le fait de déclarer certaines classes des FRBR (conception de l'œuvre, publication, représentation théâtrale...) comme des instances de classes d'entités temporelles de CRM (création, activité...) donne la possibilité de retracer le parcours d'une œuvre en reliant les données de contexte qui sont présentes mais enfouies dans les notices. En d'autres termes, le CRM apporte une dimension supplémentaire aux données bibliographiques puisque le point de vue temporel ajouté au cadre de description fait apparaître des relations entre des éléments dispersés. Ce cas nous paraît particulièrement représentatif de la manière dont le formalisme des *linked data* rend possible une conjonction de points de vue hétérogènes : alors que le modèle des FRBR est construit autour d'une arborescence centrale (œuvre > expression > manifestation > exemplaire), il est possible de lui ajouter un axe temporel sans le dénaturer.

L'autre approche, sans doute plus répandue, consiste à construire, plutôt qu'un modèle ontologique complexe, un système d'équivalence entre identifiants. La démarche peut paraître moins ambitieuse, mais le propos de Hendler se vérifie : peu de sémantique permet beaucoup de valorisation et de réutilisation des données.

Des propriétés comme `owl:sameAs`, `skos:exactMatch` et `skos:closeMatch` suffisent à établir des correspondances très utiles entre des jeux de données très hétérogènes : ainsi, RAMEAU peut nous renvoyer au *Thésaurus pour l'indexation des archives locales* aussi bien qu'à la Bibliothèque du Congrès (LOC) ou à la Bibliothèque Nationale Allemande (DNB), comme le montre le graphe ci-dessous avec l'exemple de la vedette matière "Mouvement des Lumières" [96].

96. Exemple adapté d'une présentation donnée à la BPI par E. Bermès.



Le lien des archives vers la bibliothèque permet de préciser le sens du contexte historique en donnant un accès immédiat à une documentation très complète mais très facile à exploiter puisque les ressources sont catégorisées par type et accompagnées de présentations générales de la philosophie des Lumières. Dans l'autre sens, un programme pourrait récupérer des documents iconographiques de la base de données Archim, en passant par le terme générique "XVIII^{e} siècle", et incruster automatiquement dans la page, par exemple, une vignette de la plaque de bronze sur laquelle avait été gravée la Déclaration des droits de l'homme [97]. De même, il est possible de rebondir vers LOC et DNB par des liens directs, donc d'interroger immédiatement des sources issues de deux autres pays en ayant la certitude que les sujets sont identiques ou très proches ; mais cette proximité ne conduit pas forcément à une substitution terme à terme. Il est significatif en particulier que les relations d'équivalence ne soient pas les mêmes selon qu'elles viennent de la BnF ou de LOC : la BnF est connectée à LOC à la fois directement par une relation de proche équivalence et indirectement par un alignement qui passe par le projet STITCH [98] ; depuis LOC la relation est d'exacte équivalence mais cela n'engage pas la BnF.

97. La base de données est disponible ici : → `http://www.culture.gouv.fr/documentation/archim/dossiers.htm`. L'image est bien dans Archim mais l'exemple est fictif.

98. Ce projet consiste à réunir et rendre disponibles des vocabulaires selon les standards du web de données ainsi qu'à fournir des outils pour aider à la création d'alignements : [CATCH,].

Autre cas remarquable, le projet VIAF déjà mentionné [99] constitue une passerelle entre les données d'autorités de bibliothèques d'une quinzaine de pays : il s'agit d'abord d'apparier automatiquement, en repérant les concordances de dates et de titres d'œuvres, les différentes formes possibles que prennent les noms d'une personne – mais le projet pourrait être étendu aux titres – dans les différents pays. De cette manière, on peut produire des alignements entre URIs et mutualiser les données, qui sont alors réutilisées dans d'autres contextes, comme le projet SNAC.

Soit par exemple une recherche dans un jeu de données comme data.bnf.fr sur Jacques Derrida : étant donné que l'URI équivalente sur VIAF nous est fournie, il est possible de la retrouver dans les données du projet SNAC et de rebondir vers des documents de premier ordre, accessibles via les archives en ligne de Californie [100]. Ou encore, l'existence de VIAF a servi de point de départ pour la jonction entre la BnF et le SUDOC, et qui permet à data.bnf.fr de renvoyer aux données d'autorités d'IdRef :

> C'est paradoxalement via des initiatives internationales que l'on voit se mettre en place la relation entre les données d'autorités des réservoirs distincts. Les données d'autorités de la BnF et IdRef de l'ABES sont en effet intégrées à VIAF, pièce maîtresse dans le mouvement d'ouverture et d'interconnexion des données d'autorité à l'échelle mondiale [101].

Autre application notable de VIAF : une amélioration de la qualité du catalogage "par dérivation". En effet, dans la mutualisation des informations constituant les notices, le lien entre les données bibliographiques et la source première des données d'autorités peut se perdre, par le jeu des duplications. Par exemple, une bibliothèque peut s'approvisionner dans Worldcat mais n'y trouver qu'un double incomplet, où le nom de l'auteur figure sans l'identifiant : les catalogueurs sont

99. → 2.3.3.

100. Voir le fonds *Jacques Derrida Papers*, très important puisque Derrida a enseigné et vécu aux États-Unis : → `http://www.oac.cdlib.org`. Toutefois ce rebond nécessite une recherche, les données de SNAC n'étant pas (encore ?) directement connectées à celles de la BnF.

101. Service Interministériel des Archives de France, *Bulletin sur les ressources archivistiques* numériques, n°45 (→ `www.archivesdefrance.culture.gouv.fr/static/6589`).

donc obligés de rechercher les notices d'autorités originales, tâche qui peut être épargnée grâce à VIAF [102]. On peut remarquer à ce propos que l'ontologie repose entièrement sur un triangle noms–identifiant–institution, ce qui donne toujours la possibilité de tracer la provenance de l'information [103].

3.3.2 Contexte et provenance

Les exemples présentés ci-dessus montrent deux manières de lier les données, qui correspondent toutes deux à un tissage de proche en proche : c'est d'abord autour d'intérêts partagés que des communautés en viennent à exposer leurs données et à les aligner. Malgré l'hétérogénéité des modèles, le domaine reste celui du patrimoine. Mais il n'en reste pas moins que d'autres types de connexions sont possibles et commencent à se multiplier, du fait de l'appartenance des jeux de données à la structure du web. Il est significatif en particulier que de nombreuses institutions patrimoniales s'allient avec DBpedia [104] puisque le monde du patrimoine y rencontre le web 2.0.

Toutefois, l'exposition des données et l'interaction des vocabulaires n'est pas sans soulever un certain nombre de questions. Que signifie le fait d'être dans le web pour les collections d'archives scientifiques numérisées, de données bibliographiques, muséographiques, archivistiques ? Que se passe-t-il si les sites des institutions culturelles sont automatiquement enrichis par des données issues du web ? Et dans l'autre sens, qu'advient-il des données institutionnelles une fois qu'elles sont exposées ? N'y a-t-il pas un risque de noyade dans l'océan chaotique de l'infobésité, encore accru par le fait que les données deviennent indépendantes de leur contexte documentaire ? Ces questions ne peuvent évidemment être

102. Voir le blog de l'ABES : [Nicolas, 2012].
103. Voir [Boulet, 2011], p. 13.
104. C'est d'ailleurs le cas pour le *Thésaurus pour l'indexation des archives locales*, pour les jeux de données des bibliothèques mentionnées ci-dessus, ainsi que pour le projet VIAF.

traitées exhaustivement ici, et relèvent de la prospective dans le sens où le web de données est encore très jeune. En revanche, il est possible d'achever cette réflexion en s'interrogeant sur ce que l'on appelle un monde ouvert et d'évoquer quelques pistes pour la contextualisation des informations. Cette interrogation est d'autant plus importante que le web de données se donne pour objectif d'augmenter la fiabilité des informations : ainsi la dernière couche du "gâteau" porte le nom de *Trust*. C'est aussi ce que le scénario de science-fiction de l'article de 2001 mettait en scène, et c'est enfin l'idée qu'il manque à nos navigateurs un bouton "ah, vraiment ?" qui nous permettrait de vérifier les sources d'une assertion [105].

Tout d'abord, il faut préciser que la notion de « monde ouvert » n'est pas en elle-même incompatible avec l'idée de cohérence, mais consiste seulement à rappeler que la connaissance est en perpétuelle évolution et que le nombre d'informations disponibles s'accroît en proportion de l'expérience accumulée et des moyens de communication et de stockage. Par la négative et selon la position socratique bien connue, l'idée est relativement facile à défendre, bien que difficile à admettre : il est impossible de tout savoir, et quelqu'un qui prétendrait le contraire est un ignorant qui s'ignore ou un charlatan. La réalité dépasse les limites de notre esprit, d'où, nécessairement un certain nombre de questions sans réponses et de paradoxes [106].

D'une manière positive cette fois, l'idée de monde ouvert se défend par l'affirmation qu'il est nécessaire et inévitable que coexistent des points de vue différents, partiels, complémentaires [107]. C'est le sens de la fameuse parabole de l'éléphant, souvent reprise dans les discussions sur le web de données. Dans l'obs-

105. Cf. « L'agent de Lucy, qui a une *confiance totale* dans l'agent de Pete dans le contexte particulier de cette tâche », → 1 ; voir également le fameux schéma des couches : → 1 (fig.1). À propos du bouton "oh yeah ?", voir cette page de Berners-Lee : [Berners-Lee, 1997b].

106. « Les chercheurs du Web sémantique, au contraire, acceptent les paradoxes. Les questions sans réponse sont le prix à payer pour acquérir de la souplesse. » → 1.

107. Cette position, parfois appelée « perspectivisme », doit cependant être distinguée du simple « relativisme », qui met toutes les assertions au même niveau.

curité, celui qui touche une défense peut croire qu'il a affaire à une lance, celui qui touche une patte peut croire qu'il a affaire à un arbre, celui qui touche une oreille peut croire être en présence d'une feuille de bananier, etc. Chaque point de vue détient une partie de la vérité, seule la conclusion '*c'est* un arbre/une lance/une feuille' étant erronée : pour que les assertions soient vraies il suffirait de corriger en disant 'cela ressemble à'. Évidemment, cette position revient à abandonner l'idée que l'on puisse globalement décider de la vérité et de la fausseté d'une information qui se donne pour une connaissance, en cherchant la confirmation dans les recoupements que l'on pourrait faire avec d'autres sources : il peut toujours y avoir de nouveaux faits qui viennent contredire la conclusion d'une inférence [108]. Mais cela ne veut pas dire qu'il soit impossible de conclure quoi que ce soit : c'est plutôt l'idée d'un système global qui doit être abandonnée. Le problème est donc de savoir comment déterminer dans quel contexte une assertion est valable.

Il s'agit donc de pouvoir rajouter une couche de description déterminant le champ d'application des vocabulaires de description, et ce, en évitant le risque d'une régression infinie de la description, qu'illustre ce paradoxe de L. Carroll :

> "Le nom de la chanson est appelé Yeux de morue.
>
> - Oh, c'est le nom de la chanson ?" dit Alice.
>
> - "Non, vous ne comprenez pas dit le cavalier. C'est ce que le nom est appelé. Le vrai nom est : le Vieil, vieil homme.
>
> - Alors j'aurais dû dire : est-ce ainsi que la chanson est appelée ?" corrigea Alice.
>
> - "Non, vous n'auriez pas dû : c'est tout autre chose. La chanson est appelée Voies et moyens ; mais c'est seulement ce qu'elle est appelée, vous comprenez ?
>
> - Mais alors, qu'est-ce qu'elle est ?
>
> - J'y viens, dit le cavalier, la chanson est en réalité Assis sur une barrière [109].

108. Voir M. Bergman sur cette métaphore et l'opposition entre la logique du monde fermé des bases de données relationnelles et la logique de l'hypothèse du monde ouvert ouvert : [Bergman, 2009b].

109. Texte de Carroll traduit par Deleuze ; cf. son commentaire en annexe : → 6.6.

De nombreux travaux se proposent de résoudre ce problème. Il est à la fois non souhaitable et impossible de prévoir quelles réutilisations seront faites de tel ou tel vocabulaire, mais il est possible de spécifier de manière rigoureuse dans quelle intention il a été établi, le sens des classes qu'il introduit, de préciser le domaine et la portée des propriétés, etc. C'est ainsi que K. Coyle concluait son billet sur la différence entre la logique de FOAF et celle des bibliothèques.

VoID

Une première piste consiste à utiliser un vocabulaire dédié à... la description des vocabulaires de description et des jeux de données : ainsi a été proposé le jeu de métadonnées VoID [110], qui a pour but de faciliter la réexploitation en décrivant précisément les conditions d'utilisation d'un ensemble de données liées. Les métadonnées se répartissent en quatre champs : des métadonnées générales comme celles du Dublin Core, des métadonnées techniques spécifiant les différents protocoles utilisés, des précisions sur le modèle conceptuel et un ensemble d'éléments pour décrire les interconnexions de vocabulaires. La démarche peut donner l'impression d'être une couche supplémentaire d'abstraction peu utile, mais il ne faut pas oublier qu'un jeu de données n'est pas une réalité purement formelle, un simple ensemble de triplets (*i.e.* un graphe) : c'est aussi une réalité sociale, et le fait de pouvoir indiquer la date de la dernière modification, par exemple, évite un certain nombre d'incohérences et d'erreurs.

En effet, le fait que les termes du Dublin Core soient utilisés pour donner une traçabilité des changements permet d'automatiser la vérification des mises à jour. Mais plus généralement, cela donne un début de réponse au problème de la provenance des données et le bouton "ah vraiment ?" paraît moins farfelu. Ou encore, VoID donne la possibilité de préciser à quelle version de RDF et

110. Pour *Vocabulary of Interlinked Datasets*, Vocabulaire des jeux de données interconnectés : [W3C, 2011].

XML on se réfère : ce type d'information est fondamental pour les problématiques de pérennisation, puisque cela permet de gérer les problèmes de compatibilité à travers l'évolution des couches de logiciels. En ce qui concerne la structure conceptuelle, s'il n'est pas impossible pour l'homme de comprendre le fonctionnement d'un jeu de données en l'explorant, l'appropriation est beaucoup plus facile si l'on dispose d'exemples d'application, de la liste des emprunts à d'autres jeux de données, de l'explication des sous-ensembles (*subsets*) qui composent la structure, etc. Enfin et surtout, VoID fournit des classes et des propriétés pour caractériser les liens entre jeux de données : l'ensemble des correspondances (*linkset*) entre deux vocabulaires est intéressant à analyser pour étudier les relations de confiance entre les deux producteurs, pour savoir quels domaines sont partagés, et quelles inférences sont possibles d'un jeu à l'autre [111].

La recommandation sur la provenance

Autre piste très intéressante : la recommandation du W3C sur la provenance [112], qui propose une ontologie pour décrire les échanges et transformations que subissent les informations sur le web – mais le modèle, fondé sur le triangle agent–activité–entité fournit un cadre très large, qui rappelle les réflexions archivistiques. L'introduction présente le scénario suivant : lisant un article de journal sur la criminalité comprenant une liste de régions et une grille de statistiques gouvernementales, une blogueuse croit repérer une erreur dans grille. Grâce aux termes de l'ontologie, elle peut trouver l'origine de la grille (entité), mais aussi apprendre que la grille a été générée à partir de tel jeu de données avec tel logiciel (activité). On peut imaginer également que l'exécutant de l'extraction des données puisse être identifié, et même que les rôles respectifs de la grille et de la liste dans l'article

111. Signalons également le projet *Linked Open Vocabularies*, qui propose une cartographie des vocabulaires et de leurs réutilisations, des informations sur leur maintenance par les communautés, ainsi qu'une ontologie pour caractériser leurs liens : [LOV,].

112. Une introduction est donnée ici : [W3C, 2013].

soient précisés, de telle sorte qu'on puisse savoir si cet exécutant est aussi responsable de l'analyse. Le modèle fournit également des termes pour caractériser les différentes révisions effectuées sur les entités, la datation... Le scénario imagine même le cas où l'article est corrigé mais la blogueuse l'a auparavant cité : il serait possible de signaler dans les métadonnées de provenance que la dernière version n'est qu'une "spécialisation" de l'article, ce qui confirme la version de la blogueuse et évite au journal d'être accusé de falsification. Avec ce modèle, l'analyse de la provenance des données change de niveau de granularité, puisqu'il s'agit d'étiqueter les composants des pages web eux-mêmes et plus seulement les jeux de données.

L'approche est d'autant plus séduisante que nous sommes très souvent en présence de pages constituées par des agrégats d'images dont la source se trouve sur d'autres sites, de textes copiés depuis telle autre page, de vidéos incrustées ou d'applications interactives de cartographie fournies en réalité par des tiers dont nous ignorons tout. Cette manière de produire des pages web et des applications composites appelées *mashups* est de plus en plus fréquente et remet en question l'idée même de "page" web : cela invite à penser les représentations de ressources qui s'affichent dans nos navigateurs comme un ensemble de flux canalisés à la demande par une interface intégrant des paramètres personnalisés, et il est légitime de se demander si deux internautes ont affaire au même document, ou encore si d'une fois sur l'autre l'internaute retrouve les mêmes informations [113].

Les *Named Graphs*

Mais nous n'avons pas encore de réponse à la question évidente : "quelle garantie avons-nous que ces métadonnées de provenance sont fiables ?" Très évidemment, la réponse consiste à donner des informations de provenance sur les

113. Cf. L. Moreau sur le problème de la provenance dans les mashups : [Moreau, 2010], 3.3. Voir également le billet de M.-A. Chabin précité sur la nécessité d'une diplomatique numérique.

informations de provenance : le modèle fournit une solution technique consistant à envelopper ces métadonnées dans un paquet (*bundle*) qui fera lui-même l'objet d'une description.

Avant de reposer la question de la garantie au sujet de cette dernière couche de méta-métadonnées, il faut remarquer que le formalisme RDF permet de décrire un ensemble de triplets sans qu'il s'agisse d'un jeu de données complet (*dataset*), comme c'est le cas pour VoID. Nous avons déjà vu que RDF est un métalangage, et que les classes et propriétés d'un vocabulaire peuvent être définies par RDFS et OWL ; mais jusque là, les triplets étaient considérés soit comme des entités isolées, soit comme appartenant à un groupe et constituant un vocabulaire ou un *dataset*. Il est cependant possible de définir un ensemble de triplets à un autre niveau, c'est-à-dire de délimiter une région dans un territoire, un graphe dans un graphe. Cette délimitation suppose à la fois de marquer les frontières et d'identifier – éventuellement par une URI – le graphe ainsi découpé sur la carte : on parle alors de *named graph* (graphe nommé) [114].

Le fait de spécifier qu'un ensemble de triplets forme une unité crée alors un contexte bien déterminé en fonction duquel chacun des triplets peut être interprété de manière univoque : de cette manière, bien que faisant partie d'un monde ouvert, la région acquiert une autonomie où l'on peut appliquer la logique du monde fermé, à savoir raisonner sans qu'interviennent d'autres hypothèses et opérer des inférences avec certitude [115]. Plus concrètement, cela permet d'attribuer à un ensemble d'assertions un producteur – Untel est l'auteur de ces triplets –, ou encore de qualifier les triplets en précisant leur modalité : Untel suppose ou certifie que... Et surtout, ces graphes nommés peuvent être signés électroniquement, tout comme n'importe quel document numérique génère une empreinte unique si on lui ap-

114. Cf. [Gandon et Corby, 2009].

115. Cf. [Bizer et Carroll, 2004]. Les auteurs montrent notamment que la contextualisation par les *named graphs* permet de résoudre le paradoxe du Crétois "je mens".

plique un algorithme [116]. Les métadonnées de provenance formant un graphe, il suffit alors de le nommer et de le signer pour établir une garantie de la qualité de ces informations.

L'usage des *named graphs* ouvre donc des perspectives pour une « architecture de la confiance [117]. » au sein de l'architecture distribuée du web. Bizer et Oldakowski mentionnent trois directions de recherche complémentaires. La première est fondée sur la réputation et les mécanismes d'évaluation comme les notes attribuées par les internautes ou les sites spécialisés dans la comparaison de produits. La seconde consiste à utiliser les métadonnées précisant le contexte et permettant de savoir qui a dit quoi. Enfin, une dernière approche s'appuie sur l'analyse des contenus eux-mêmes pour recouper les informations avec d'autres sources. Il est en tout cas certain que la réflexion reste ouverte, puisqu'il s'agit de composer avec les contraintes du monde ouvert et de se passer d'une instance centrale de certification. Là encore, la manière dont se tissent les relations dans le web de données dépend des interactions sociales, et dont s'établissent des « chaînes de confiance » de proche en proche [118].

116. Cf. `http://www.w3.org/2011/rdf-wg/wiki/TF-Graphs-UC/Digital_Signatures`.

117. Cf. [Bizer et Oldakowski, 2004].

118. On peut rappeler le billet de T. Berners-Lee précité sur le web de données vu comme un réseau social (→ 2.3.3) et l'exemple du site Advogato. Sur cette notion de chaîne de confiance, voir [Bizer *et al.*, 2004].

4 Conclusion

> En vérité, tout, je vous assure, peut, absolument, répondre à tout : c'est le grand kaléidoscope des mots humains. Étant donnés la couleur et le ton d'un sujet dans l'esprit, n'importe quel vocable peut toujours s'y adapter en un sens quelconque, dans l'éternel à peu près de l'existence et des conversations humaines. – Il est tant de mots vagues, suggestifs, d'une élasticité intellectuelle si étrange ! et dont le charme et la profondeur dépendent, simplement, de ce à quoi ils répondent !
>
> Villiers de L'Isle Adam, *L'Ève future*

> Cette logique opère un peu à la façon du kaléidoscope : instrument qui contient aussi des bribes et des morceaux, au moyen desquels se réalisent des arrangements structuraux. Les fragments sont issus d'un procès de cassure et de destruction, en lui-même contingent, mais sous réserve que ses produits offrent entre eux certaines homologies : de taille, de vivacité de coloris, de transparence."
>
> C. Lévi-Strauss, *La pensée sauvage*.

Au terme de cette réflexion exploratoire, nous avons réuni des éléments montrant que le projet du web sémantique, redéfini par ou émergeant sous la forme du web de données, n'est pas une utopie abstraite mais une vision qui se construit sur des technologies et des standards. Nous posions deux questions : la première était de savoir si l'ambition de ses créateurs était de rendre les machines capables de comprendre le sens du discours humain ou seulement de le manipuler, et auquel cas, comment cette manipulation est possible. La seconde question consistait à se demander quelle est la structure du web de données et dans quelle mesure elle peut être à la fois compatible avec la logique participative des échanges du web 2.0 et avec celle des descriptions en usage dans les institutions patrimoniales.

Nous avons distingué la sémantique du web de données de celle du traitement automatique des langues, et montré que tout repose sur la représentation des connaissances au moyen de langages et de règles formelles. Si les machines sont capables de manipuler du sens sans le comprendre, c'est qu'elles ne font rien de plus que des opérations de tri et de liaison, comme nous sommes habitués à les

voir réaliser par les systèmes de gestion de bases de données. Toute l'astuce consiste en effet à caractériser de manière suffisamment univoque le contenu d'une ressource – c'est pourquoi on parle de retour des vocabulaires contrôlés – et à rendre cette description "lisible" par les programmes. Cette lisibilité repose seulement sur l'utilisation de langages formels fournissant des marqueurs, des balises, des repères qui délimitent un contenu et font connaître en même temps son type.

Mais la véritable innovation du *linked data* consiste à nous permettre d'utiliser les liens hypertexte plutôt que des relations entre tables ou schémas, le protocole HTTP plutôt que des systèmes très lourds de passerelles entre bases de données. La structuration des données étant au même niveau que celles-ci dans l'unité du triplet, n'importe quelle ressource, n'importe quel élément de description, peut être relié(e) à n'importe quel(le) autre, pourvu qu'on utilise des identifiants compatibles avec HTTP. Ces identifiants ne pointent plus alors seulement vers des pages web mais deviennent également des noms désignant des entités réelles et conceptuelles, ce qui constitue la base du système et un changement important – autant que difficile à comprendre. Le cadre RDF et les URIs permettent ainsi de construire autant de vocabulaires de description que souhaité, du plus simple au plus complexe – certaines ontologies ayant même un caractère encyclopédique. Grâce à ces vocabulaires, il devient possible de manipuler de très grandes collections de données, dispersées dans tout le web.

Cependant, cette mise en relation de ressources dispersées et de natures hétérogènes soulève des difficultés, à la fois du point de vue du web 2.0, où les collections de données sont librement "taggées", et du point de vue des organisations patrimoniales, où les collections sont décrites selon des normes et une division du travail très précises. Si le web de données suppose l'adoption d'une architecture et d'un certain nombre de standards, ce n'est pas incompatible avec les spécificités des institutions et des communautés : il ne s'agit pas de se plier à une logique

préétablie, puisque le web est un monde ouvert, où peuvent cohabiter les langages et les pratiques. Les vocabulaires peuvent proliférer, s'associer, se recouper. De très nombreuses combinaisons sont possibles, comme dans un kaléidoscope : c'est bien de cette manière d'ailleurs que les langues naturelles se mélangent et évoluent, en réutilisant les termes et les tournures les unes des autres.

Le problème est alors de savoir comment s'y retrouver dans ce grand mélange, et comment le web de données peut tenir sa promesse de nous fournir des informations de qualité. Une fois admis le principe que l'interconnexion des collections et des systèmes de description ne menace pas l'intégrité de chacun, on peut se demander comment conserver une cohérence dans ce monde ouvert où tout élément de description peut être réutilisé dans un autre contexte. Encore une fois, le formalisme de RDF et le fait qu'il soit un métalangage fournissent des solutions, ou au moins un certain nombre de pistes sérieuses. Si la difficulté est d'associer un monde ouvert et la nécessité de contextualiser l'information, de garder la trace de sa provenance et des transformations qu'elle subit, alors RDF doit pouvoir exprimer des données sur le contexte et la provenance dans de nouveaux vocabulaires et en traitant les graphes comme des ressources. Nous n'avons pu qu'évoquer ces perspectives, qui mériteraient une analyse détaillée, mais ce survol peut donner une idée des moyens d'établir dans le web des zones de confiance, sans rompre pour autant avec le principe de l'ouverture qui le caractérise.

Bien sûr, cette réflexion laisse de côté une dimension essentielle du problème de la faisabilité du web de données, à savoir celle des facteurs organisationnels, sociaux et économiques qui pourraient accélérer ou stopper son développement. De fait, si l'on compare son évolution à celle du web de documents, qui a été « virale », on peut rester sceptique sur l'avenir du projet. Ses défenseurs ont eux-mêmes dû admettre un relatif échec, et "revisiter" le concept de web sémantique (→ note [74]).

Un premier problème évident, c'est celui de l'absence criante d'applications populaires (*killerapps*) qui rendraient les technologies plus attractives et seraient un motif d'investissement. Il est vrai que la sémantique est discrète : par exemple les fils de syndication RSS sont connus mais ne laissent pas soupçonner le potentiel des technologies. Autre problème : les jeux de données disponibles ne sont pas forcément très intéressants en termes de réutilisation, ou sont trop peu visibles ; par ailleurs, il est légitime de s'interroger sur les dispositions de chacun à jeter ses données sur la place publique, même en admettant que tout ne soit pas gratuit. Ou encore, on peut se demander si le projet n'est pas condamné en l'absence d'organisations capables de pérenniser les identifiants, de gérer les changements de nom de domaine et autres problèmes d'adresses. Dans le cas des institutions patrimoniales, cette question de la pérennité est déterminante et implique une volonté politique.

Toutefois, il semble que ces dernières années nous assistions à un changement d'échelle qui pourrait être interprété comme la confirmation de la cohérence du projet.

Pour commencer, il faut rappeler que l'idée de *Linked Data* est presque indissociable de l'idée d'*Open Data*. Ainsi, la fameuse page de Berners-Lee énonçant les quatre principes du *Linked Data* propose une gradation qualitative entre plusieurs types de données liées, de une à cinq étoiles : la première étoile est attribuée à des données qui sont mises à disposition en libre accès, de manière structurée, sans qu'elles soient exprimées RDF. L'abondance de données en libre accès est donc la première condition à respecter pour que le web de données soit viable. De même, la conférence de 2009 précitée fut l'occasion de lancer un appel à l'exposition massive de données brutes (« *raw data now !* »). Or cet appel a été entendu, et Berners-Lee s'est impliqué dans la publication du jeu de données gouvernementales du Royaume-Uni (data.gov.uk), qui a initié un mouvement de mise à

disposition de la part des institutions politiques étatiques et locales. Une directive en ce sens avait été adoptée au niveau européen en 2003 et a été réactualisée cette année ; certains parlent de « mode » de l'open data [119]. L'ouverture de l'accès aux données des institutions donnant l'image d'une volonté de transparence, on peut supposer que le mouvement va se poursuivre.

Enfin, un dernier élément pourrait bien favoriser l'essor des technologies sémantiques : l'implication de majors du web. Un événement intéressant est l'association des trois moteurs de recherche Google Yahoo et Bing dans le projet schema.org [120], qui a pour but de développer des balises de description du contenu des pages web destinées à améliorer l'indexation par les robots, et qui fournit un vocabulaire RDF. Facebook s'est également lancé avec Open Graph Protocol [121], un projet du même ordre mais tourné vers la récupération des descriptions dans les applications du réseau social. Enfin, Google a lancé cette année le Knowledge Graph [122], qui utilise les données de DBpedia pour enrichir ses pages de résultats en proposant un encart (en haut et à droite de la page, lorsque des informations sont disponibles) qui incruste des images, des données de localisations, des éléments de biographies, propose des termes associés.

Mais la question reste ouverte de savoir si et comment les internautes vont s'emparer des outils du web de données comme ils l'ont fait pour ceux du web 2.0.

119. Cf. par exemple cet article : `http://www.lagazettedescommunes.com/172065/open-data-nouvelle-directive-adoptee-par-le-parlement-europeen/`.
120. → `http://schema.org/`..
121. → `http://opengraphprotocol.org/`
122. Présenté ici : → `http://www.google.com/insidesearch/features/search/knowledge.html`

5 Bibliographie

5.1 Présentations générales et manuels

[Bachimont, 2007] BACHIMONT, B. (2007). *Ingénierie des connaissances et des contenus - le numérique entre ontologies et documents*. Lavoisier, Paris.

[Bachimont et alii, 2011] BACHIMONT, B. et ALII (2011). Enjeux et technologies : des données au sens. *Documentaliste-Sciences de l'Information*, n°48.

[Bermès, 2011] BERMÈS, E. (2011). *Convergence et interopérabilité : l'apport du Web de données – IFLA*. `http://conference.ifla.org`, page consultée le 28/08/2013.

[Bermès et Martin, 2010] BERMÈS, E. et MARTIN, F. (2010). Le concept de collection numérique. *Bulletin des Bibliothèques de France*, n°55.

[Bertails et Poupeau, 2009] BERTAILS, A. et POUPEAU, G. (2009). *L'apport des technologies du Web sémantique à la gestion des données structurées*. `http://fr.slideshare.net/lespetitescases/lapport-des-technologies-du-web-smantique-la-gestion-des-donnes-structures`, page consultée le 28/08/2013.

[Brand *et al.*,] BRAND, A., DALY, F. et MEYERS, B. *Metadata Demystified : A Guide for Publishers – NISO*. `http://www.niso.org`, page consultée le 28/08/2013.

[Calderan *et al.*, 2008] CALDERAN, L., HIDOINE, B., MILLET, J. et ALII (2008). *Métadonnées mutations et perspectives : séminaire INRIA, Dijon, 29 septembre - 3 octobre 2008*. ADBS éditions, Paris.

[Calderan *et al.*, 2012] CALDERAN, L., LAURENT, P. et ALII (2012). *Le document numérique à l'heure du Web de données séminaire INRIA, Carnac, 1er - 5 octobre 2012*. ADBS éditions, Paris.

[Chaudiron, 2007] CHAUDIRON, S. (2007). Technologies linguistiques et modes de représentation de l'information textuelle. *Documentaliste-Sciences de l'Information*, n°44.

[Gandon, 2012] GANDON, F. (2012). *Le Web sémantique - comment lier les données et les schémas sur le web ?* Dunod Infopro, Paris.

[linkeddata.org,] LINKEDDATA.ORG. *Guides and Tutorials | Linked Data - Connect Distributed Data across the Web*. `http://linkeddata.org/guides-and-tutorials`, page consultée le 28/08/2013.

[linkeddatatools.com, 2010] LINKEDDATATOOLS.COM (2010). *Introducing Linked Data And The Semantic Web*. `http://www.linkeddatatools.com/semantic-web-basics`, page consultée le 28/08/2013.

[OKFN, 2009] OKFN (2009). *The Open Data Handbook — Open Data Handbook*. http://opendatahandbook.org/, page consultée le 28/08/2013.

[Timimi *et al.*, 2005] TIMIMI, I., KOVACS, S. et ALII (2005). *Indice, index, indexation : actes du colloque CERSATES-GERIICO*. ADBS éditions, Paris.

5.2 Textes fondateurs et recommandations

[Berners-Lee, 1989] BERNERS-LEE, T. (1989). *The original proposal of the WWW, HTMLized*. http://www.w3.org/History/1989/proposal.html, page consultée le 28/08/2013.

[Berners-Lee, 1994] BERNERS-LEE, T. (1994). *Plenary talk by Tim BL at WWWF94 : Overview*. http://www.w3.org/Talks/WWW94Tim/, page consultée le 28/08/2013.

[Berners-Lee, 1997a] BERNERS-LEE, T. (1997a). *Realising the Full Potential of the Web*. http://www.w3.org/1998/02/Potential.html, page consultée le 28/08/2013.

[Berners-Lee, 1997b] BERNERS-LEE, T. (1997b). *Tim Berners-Lee - Consistent User Interface*. http://www.w3.org/DesignIssues/UI.html, page consultée le 28/08/2013.

[Berners-Lee, 1998a] BERNERS-LEE, T. (1998a). *Hypertext Style : Cool URIs don't change.* http://www.w3.org/Provider/Style/URI, page consultée le 28/08/2013.

[Berners-Lee, 1998b] BERNERS-LEE, T. (1998b). *Web design issues ; What a semantic can represent*. http://www.w3.org/DesignIssues/RDFnot.html, page consultée le 28/08/2013.

[Berners-Lee, 2001] BERNERS-LEE, T. (2001). *Le web sémantique – traduction de l'article de Berners-Lee et alii paru en 2001*. http://www.urfist.cict.fr/archive/lettres/lettre28/lettre28-22.html, page consultée le 28/08/2013.

[Berners-Lee, 2006] BERNERS-LEE, T. (2006). *Linked Data - Design Issues*. http://www.w3.org/DesignIssues/LinkedData.html, page consultée le 28/08/2013.

[Berners-Lee, 2007] BERNERS-LEE, T. (2007). *Giant Global Graph | Decentralized Information Group (DIG) Breadcrumbs*. http://dig.csail.mit.edu/breadcrumbs/node/215, page consultée le 28/08/2013.

[Berners-Lee, 2009] BERNERS-LEE, T. (2009). *On the next web – TED talks*. http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html, page consultée le 28/08/2013.

[Berners-Lee *et al.*, 2006] BERNERS-LEE, T., HALL, W. et SHADBOLT, N. (2006). *The Semantic Web Revisited*. http://eprints.soton.ac.uk/262614/, page consultée le 28/08/2012.

[Berners-Lee, 1998c] BERNERS-LEE, T. a. (1998c). *Semantic Web roadmap*. http://www.w3.org/DesignIssues/Semantic.html, page consultée le 28/08/2013.

[Bizer et Heath, 2011] BIZER, C. et HEATH, T. (2011). *Linked Data : Evolving the Web into a Global Data Space*. http://linkeddatabook.com/editions/1.0/, page consultée le 28/08/2013.

[Bizer *et al.*, 2009] BIZER, C., HEATH, T. et BERNERS-LEE, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, Special Issue on Linked Data.

[Gruber, 1992] GRUBER, T. (1992). *What is an Ontology ?* http://www-ksl.stanford.edu/kst/what-is-an-ontology.html,page consultée le 28/08/2013.

[W3C,] W3C. *LinkingOpenData - W3C Wiki*. http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData, page consultée le 28/08/2013.

[W3C, 2004a] W3C (2004a). *OWL Web Ontology Language Guide – mapping*. http://www.w3.org/TR/2004/REC-owl-guide-20040210/#OntologyMapping, page consultée le 28/08/2013.

[W3C, 2004b] W3C (2004b). *RDF Primer*. http://www.w3.org/TR/rdf-primer/, page consultée le 28/08/2013.

[W3C, 2004c] W3C (2004c). *RDF Semantics*. http://www.w3.org/TR/rdf-mt/, page consultée le 28/08/2013.

[W3C, 2004d] W3C (2004d). *RDF Vocabulary Description Language 1.0 : RDF Schema*. http://www.w3.org/TR/rdf-schema/, page consultée le 28/08/2013.

[W3C, 2004e] W3C (2004e). *RDF/XML Syntax Specification (Revised)*. http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/#section-Namespace, page consultée le 28/08/2013.

[W3C, 2004f] W3C (2004f). *Resource Description Framework (RDF) : Concepts and Abstract Syntax*. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, page consultée le 28/08/2013.

[W3C, 2009] W3C (2009). *SKOS Simple Knowledge Organization System Primer*. http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/, page consultée le 28/08/2013.

[W3C, 2011] W3C (2011). *Describing Linked Datasets with the VoID Vocabulary*. http://www.w3.org/TR/void/, page consultée le 28/08/2013.

[W3C, 2011] W3C (2011). *Rapport final du groupe d'incubation "Bibliothèques et web de données"*. http://mediatheque.cite-musique.fr/MediaComposite/ARTICLES/W3C/XGR-lld-fr.html, page consultée le 28/08/2013.

[W3C, 2013] W3C (2013). *PROV Model Primer*. http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/, page consultée le 28/08/2013.

5.3 Aspects scientifiques et techniques

[Abiteboul, 2012] ABITEBOUL, S. (2012). *Sciences des données - S. Abiteboul - Collège de France*. http://lecons-cdf.revues.org/529, page consultée le 28/08/2013.

[Aussenac-Gilles, 2012] AUSSENAC-GILLES, N. (2012). *Donner du sens à des documents semi- structurés : de la construction d'ontologies à l'annotation sémantique*. http://www.inria.fr/medias/agenda/general/documents-pdf/ist-2012-aussenac, page consultée le 28/08/2013.

[Bachimont, 2000] BACHIMONT, B. (2000). *Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en Ingénierie des connaissances*. http://www.utc.fr/~bachimon/Publications_attachments/Ontologie-ICBook.pdf.

[Bao *et al.*, 2010] BAO, J., DING, L. et MC GUINNESS, D. (2010). *Contexts and Importing in RDF*. http://fr.slideshare.net/baojie_iowa/2010-0624-rdfcontext, page consultée le 28/08/2013.

[Bermès, 2005] BERMÈS, E. (2005). *Petite histoire des classifications*. http://www.figoblog.org/document803.php, page consultée le 28/08/2013.

[Bizer et Carroll, 2004] BIZER, C. et CARROLL, J. (2004). *Modelling context using Named Graphs*. http://lists.w3.org/Archives/Public/www-archive/2004Feb/att-0072/swig-bizer-carroll.pdf, page consultée le 28/08/2013.

[Bizer *et al.*, 2004] BIZER, C., CARROLL, J., HAYES, P. et STICKLER, P. (2004). Named graphs, provenance and trust. *HP Laboratories Bristol*.

[Bizer et Oldakowski, 2004] BIZER, C. et OLDAKOWSKI, R. (2004). *Using context and content based trust policies on the Semantic Web*. http://wissensnetze.ag-nbi.de/publ/Using%20Context-

%20and%20Content-Based%20Trust%20Policies%20on%20the%20Semantic%20Web.pdf, page consultée le 28/08/2013.

[BnF, 2008] BNF (2008). *BnF - Référentiels, données d'autorité, thésaurus, ontologies, taxonomies* http://www.bnf.fr/fr/professionnels/autres_journees_professionnelles/a.referentiels_afnor_2008.html, page consultée le 28/08/2013.

[BnF, 2011] BNF (2011). *BnF - Référentiels et données d'autorité à l'heure du Web sémantique*. http://www.bnf.fr/fr/professionnels/autres_journees_professionnelles/a.referentiel_donnees_autorites_110527.html, page consultée le 28/08/2013.

[Boulet, 2011] BOULET, V. (2011). *VIAF, une brique importante pour le Web sémantique*. http://www.bnf.fr/documents/afnor2011_viaf_boulet.pdf, page consultée le 28/08/2013.

[Capelli *et al.*, 2012] CAPELLI, L., KILOUCHI, S., MINEL, J.-L., POUPEAU, G. et POUYLLAU, S. (2012). *Comment contribuer, avec ses données numériques, à ISIDORE ?* http://www.huma-num.fr/sites/default/files/ressourcesdoc/guide_isidore_2012.pdf, page consultée le 28/08/2013.

[Cavalié, 2010] CAVALIÉ, . (2010). *RDF vs XML : illustration avec SKOS vs MarcXML | Bibliothèques [reloaded]*. http://bibliotheques.wordpress.com/2010/10/22/rdf-vs-xml-illustration-avec-skos-vs-marcxml/, page consultée le 28/08/2013.

[Charlet et Declerck, 2011] CHARLET, J. et DECLERCK, G. (2011). *Intelligence Artificielle, ontologies et connaissances en médecine. Les limites de la mécanisation de la pensée*. http://www.academia.edu/1038158/Intelligence_Artificielle_ontologies_et_connaissances_en_medecine._Les_limites_de_la_mecanisation_de_la_pensee, page consultée le 28/08/2013.

[Eiter *et al.*, 2008] EITER, T., IANNI, G., KRENNWALLNER, T. et POLLERES, A. (2008). *Rules and Ontologies for the Semantic Web*. http://axel.deri.ie/publications/eite-etal-2008.pdf, page consultée le 28/08/2013.

[Gandon et Corby, 2009] GANDON, F. et CORBY, O. (2009). *Name That Graph*. http://www.w3.org/2009/12/rdf-ws/papers/ws06/, page consultée le 28/08/2013.

[GRIHO,] GRIHO. *Rhizomik – Living SW*. http://rhizomik.net/html/livingsw/, page consultée le 28/08/2013.

[Halpin *et al.*, 2010] HALPIN, H., THOMPSON, H., HAYES, P., MCGUINNESS, D. et MCCUSKER, J. (2010). *When owl :sameAs isn't the Same*.

http://iswc2010.semanticweb.org/pdf/261.pdf, page consultée le 28/08/2013.

[Hayes, 2009] HAYES, P. (2009). *BLOGIC. (ISWC 2009 Invited Talk)*. http://fr.slideshare.net/PatHayes/blogic-iswc-2009-invited-talk, page consultée le 28/08/2013.

[Hayes, 2012] HAYES, P. (2012). *Rdf with contexts*. http://fr.slideshare.net/PatHayes/rdf-with-contexts, page consultée le 28/08/2013.

[ICOM et IFLA, 2013] ICOM et IFLA (2013). *FRBR – object-oriented definition and mapping from FRBERer, FRAD and FRSAD*. http://www.cidoc-crm.org/docs/frbr_oo//frbr_docs/FRBRoo_V2.0_draft_2013May.pdf, page consultée le 28/08/2013.

[InterPARES, 2002] INTERPARES (2002). *Conditions requises pour évaluer et maintenir l'authenticité des documents d'archives électroniques*. http://www.cdncouncilarchives.ca/ATF_Requirements_FR.pdf, page consultée le 28/08/2013.

[Johnston, 2011a] JOHNSTON, P. (2011a). *Things and their conceptualisations : SKOS, foaf :focus & modelling choices*. http://efoundations.typepad.com/efoundations/2011/09/things-their-conceptualisations-skos-foaffocus-modelling-choices.html, page consultée le 28/08/2013.

[Johnston, 2011b] JOHNSTON, P. (2011b). *Two changes to the model and some definitions | LOCAH Project*. http://archiveshub.ac.uk/locah/2011/02/two-changes-to-the-model-and-some-definitions/, page consultée le 28/08/2013.

[Lahary, 2003] LAHARY, D. (2003). *Z39.50 / Dominique Lahary*. http://www.lahary.fr/pro/z3950/, page consultée le 28/08/2013.

[Lallich-Boidin et Maret, 2005] LALLICH-BOIDIN, G. et MARET, D. (2005). *Recherche d'information et traitement de la langue*. Presses de l'Enssib, Villeurbanne.

[Mayr et Petras, 2008] MAYR, P. et PETRAS, V. (2008). *Cross-concordances : terminology mapping and its effectiveness for information retrieval*. http://www.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf, page consultée le 28/08/2013.

[McCusker et McGuinness, 2010] MCCUSKER, J. et MCGUINNESS, D. L. (2010). *owl :sameAs Considered Harmful to Provenance*. http://fr.slideshare.net/jpmccusker/owlsameas-considered-harmful-to-provenance, page consultée le 28/08/2013

[Mkadmi et Saleh, 2008] MKADMI, A. et SALEH, I. (2008). *Bibliothèque numérique et recherche d'informations*. Lavoisier, Paris.

[Moreau, 2010] MOREAU, L. (2010). *The Foundations for Provenance on the Web*. http://eprints.soton.ac.uk/271691/1/survey.pdf, page consultée le 28/08/2013.

[Nicolas, 2012] NICOLAS, Y. (2012). *IdRef dans VIAF et après ... Passer d'un identifiant à l'autre (VIAF, IdRef, LC, BnF, Wikipedia, ...) | Punktokomo*. http://punktokomo.abes.fr/2012/05/11/idref-dans-viaf-identifiants/, page consultée le 28/08/2013.

[Pouyllau, 2010] POUYLLAU, S. (2010). *Construire le web de données pour les sciences humaines et sociales*. http://blog.stephanepouyllau.org/401.

[Rubinstein,] RUBINSTEIN, A. *OWLDoc – Archival Ontology*. http://gslis.simmons.edu/archival/arch/index.html, page consultée le 28/08/2013.

[SIAF, 2011] SIAF (2011). *Bulletin sur les ressources archivistiques numériques*. http://www.archivesdefrance.culture.gouv.fr/static/4997, page consultée le 28/08/2013.

[SIAF, 2012] SIAF (2012). *Bulletin sur les ressources archivistiques numériques*. n°44, page consultée le 28/08/2013.

[Sowa,] SOWA, J. *Conceptual Graphs*. http://conceptualgraphs.org, page consultée le 28/08/2013.

[Tennison, 2011] TENNISON, J. (2011). *What Do URIs Mean Anyway ? | Jeni's Musings*. http://www.jenitennison.com/blog/node/159, page consultée le 28/08/2013.

[Théreaux, 2006] THÉREAUX, O. (2006). *Content Negotiation : why it is useful, and how to make it work - W3C Blog*. http://www.w3.org/QA/2006/02/content_negotiation.html, page consultée le 28/08/2013.

[Tosca-Consultants, 2008] TOSCA-CONSULTANTS (2008). *Le catalogue de la bibliothèque à l'heure du Web 2.0*. ADBS éditions, Paris.

[Tricot, 2006] TRICOT, C. (2006). Cartographie sémantique de fonds numériques et techniques. *Document numérique*, n°9.

[Van Gendt *et al.*, 2006] VAN GENDT, M., ISAAC, A. et Van der MEIJ, L. (2006). *Semantic Web Techniques for Multiple Views on Heterogeneous Collections : a Case Study*. http://www.few.vu.nl/~aisaac/papers/STITCH-ECDL06.pdf, page consultée le 28/08/2013.

[Weller, 2010] WELLER, K. (2010). *Knowledge representation in the social semantic web*. De Gruyter, Berlin.

[Zacklad, 2007] ZACKLAD, M. (2007). *Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information*. Association Canadienne des Sciences de l'Information, Montréal.

5.4 Applications

[Angjeli et Isaac, 2008] ANGJELI, A. et ISAAC, A. (2008). *Web sémantique et interopérabilité des vocabulaires : une expérimentation dans le domaine des enluminures*. IFLA.

[Antidot, 2012] ANTIDOT (2012). *Découvrez les Monuments Historiques grâce à l'Open Data ! « Blog Antidot*. `http://blog.antidot.net/2011/12/19/decouvrez-les-monuments-historiques-grace-a-lopen-data/`, page consultée le 28/08/2013.

[CATCH,] CATCH. *STITCH @ CATCH Vocabulary and Alignment Repository*. `http://www.cs.vu.nl/STITCH/repository/`, page consultée le 28/08/2013.

[Centre Pompidou,] CENTRE POMPIDOU. *Centre Pompidou Virtuel | Accueil*. `http://www.centrepompidou.fr/`, page consultée le 28/08/2013.

[data.bnf.fr, 2012] DATA.BNF.FR (2012). *Web sémantique et modèle de données (data.bnf.fr)*. `http://data.bnf.fr/semanticweb`, page consultée le 28/08/2013.

[DERI,] DERI. *Sindice - The semantic web index*. `http://sindice.com/`, page consultée le 28/08/2013.

[Europeana – LOD, a] EUROPEANA – LOD. *Europeana Professional - Linked Open Data*. `http://pro.europeana.eu/linked-open-data`, page consultée le 28/08/2013.

[Europeana – LOD, b] EUROPEANA – LOD. *Search - Europeana*. `http://eculture.cs.vu.nl/europeana/session/search?`, page consultée le 28/08/2013.

[FOAF, 2009] FOAF (2009). *FoafExamples - FOAF Wiki*. `http://wiki.foaf-project.org/w/FoafExamples`, page consultée le 28/08/2013.

[ICOM-CIDOC,] ICOM-CIDOC. *The CIDOC CRM*. `http://www.cidoc-crm.org/frbr_inro.html`, page consultée le 28/08/2013.

[LODLAM,] LODLAM. *LODLAM - Linked Open Data in Libraries, Archives & Museums*. `http://lodlam.net/`, page consultée le 28/08/2013.

[LOV,] LOV. *(LOV) Linked Open Vocabularies*. `http://lov.okfn.org/dataset/lov/`, page consultée le 28/08/2013.

[OCLC,] OCLC. *VIAF – Virtual International Authority File*. http://viaf.org/, page consultée le 28/08/2013.

[OKFN,] OKFN. *Welcome - the Data Hub*. http://datahub.io/, page consultée le 28/08/2013.

[Schema.org,] SCHEMA.ORG. *Home - schema.org*. http://schema.org/, page consultée le 28/08/2013.

[SIAF,] SIAF. *Le thésaurus W et le Web de données | Le thésaurus W dans le Web de données*. http://web.archive.org/web/20130324193310/http://www.archivesdefrance.culture.gouv.fr/thesaurus/thesaurus-w-web-de-donnees.html, page consultée le 28/08/2013.

[SIOC, 2009] SIOC (2009). *SIOC Applications | sioc-project.org*. http://sioc-project.org/applications, page consultée le 28/08/2013.

[SNAC Project, 2010] SNAC PROJECT (2010). *SNAC : The Social Networks and Archival Context*. http://socialarchive.iath.virginia.edu/index.html, page consultée le 28/08/2013.

[Stevenson, 2011] STEVENSON, A. (2011). *Archivalia : How to Make Bibliographic and Archival Linked Data*. http://archiv.twoday.net/stories/14648504/, page consultée le 28/08/2013.

[TGE-Adonis,] TGE-ADONIS. *ISIDORE - Accès aux données et services numériques de SHS*. http://www.rechercheisidore.fr/, page consultée le 28/08/2013.

[W3C, 2011] W3C (2011). *Use Cases – Library Linked Data*. http://www.w3.org/2005/Incubator/lld/wiki/Use_Cases, page consultée le 28/08/2013.

[Wimmics *et al.*,] WIMMICS, WIKIMEDIA et MINISTÈRE DE LA CULTURE ET DE LA COMMUNICATION. *Accueil - DBpediaFr*. http://wimmics.inria.fr/projects/dbpedia/doc/index.php/Accueil, page consultée le 28/08/2013.

5.5 Commentaires et discussions

[Bergman, 2009a] BERGMAN, M. (2009a). *Advantages and Myths of RDF*. http://www.mkbergman.com/483/advantages-and-myths-of-rdf/, page consultée le 28/08/2013.

[Bergman, 2009b] BERGMAN, M. (2009b). *The Open World Assumption : Elephant in the Room*. http://www.mkbergman.com/852/the-open-world-assumption-elephant-in-the-room/, page consultée le 28/08/2013.

[Chabin, 2011] CHABIN, M.-A. (2011). *Peut-on parler de diplomatique numérique ?*. http://www.marieannechabin.fr/diplomatique-numerique/, page consultée le 28/08/2013.

[Coyle, 2010] COYLE, K. (2010). *Coyle's InFormation : Libraries, FOAF, and community*. http://kcoyle.blogspot.fr/2010/09/libraries-foaf-and-community.html, page consultée le 28/08/2013.

[Deleuze, 1969] DELEUZE, G. (1969). *Logique du sens*. Éditions de Minuit, Paris.

[Gandon, 2006] GANDON, F. (2006). *Le web sémantique n'est pas antisocial*. http://fr.slideshare.net/fabien_gandon/le-web-smantique-nest-pas-antisocial-version-de-2006, page consultée le 28/08/2013.

[Halpin, 2006] HALPIN, H. (2006). *Identity, Reference, and Meaning on the Web*. http://wwwconference.org/proceedings/www2006/www.ibiblio.org/hhalpin/irw2006/hhalpin.html, page consultée le 28/08/2013.

[Hendler, 2007] HENDLER, J. (2007). *The Dark Side of the Semantic Web*. http://www.computer.org/csdl/mags/ex/2007/01/x1002.html, page consultée le 28/08/2013.

[Iskold, 2007] ISKOLD, A. (2007). *The Attention Economy : An Overview – ReadWrite*. http://readwrite.com/2007/03/01/attention_economy_overview#awesm=~odmyfgBKNZ2LPd, page consultée le 28/08/2013.

[Maignien, 2009] MAIGNIEN, Y. (2009). *l'éditorial du 9 Décembre 2009 du Très Grand Equipement Adonis : Les nouvelles frontières numériques des sciences*. http://www.tge-adonis.fr/editorial/les-nouvelles-frontieres-numeriques-des-sciences, page consultée le 28/08/2013.

[Monnin, 2012a] MONNIN, A. (2012a). *L'ingénierie philosophique comme design ontologique : retour sur l'émergence de la ressource*. Revue Réel-virtuel : enjeux du numérique. http://reelvirtuel.univ-paris1.fr/index.php?/revue-en-ligne/3-monnin/, page consultée le 28/08/2013.

[Monnin, 2012b] MONNIN, A. (2012b). *Philosophie et ingénierie du Web | Implications philosophiques*. http://www.implications-philosophiques.org/actualite/une/philosophie-et-ingenierie-du-web/, page consultée le 28/08/2013.

[Monnin, 2013] MONNIN, A. (2013). *Les ressources des ombres recalcitrantes*. http://hal.archives-ouvertes.fr/docs/00/83/85/

35/PDF/sociologies-4334-les-ressources-des-ombres-recalcitrantes.pdf, page consultée le 28/08/2013.

[Petitot et Rosentiehl, 1974] PETITOT, J. et ROSENTIEHL, P. (1974). Automate asocial et systèmes acentrés.

[Poupeau, 2006] POUPEAU, G. (2006). *L'ontologie est-elle vraiment surfaite ?* | *Les petites cases*. http://www.lespetitescases.net/l-ontologie-est-elle-vraiment-surfaite-y, page consultée le 28/08/2013.

[Poupeau, 2008a] POUPEAU, G. a. (2008a). *Du Web sémantique au web de données, 1ère partie* | *Les petites cases*. http://www.lespetitescases.net/Du-Web-semantique-au-web-de-donnees-1, page consultée le 28/08/2013.

[Poupeau, 2008b] POUPEAU, G. b. (2008b). *Du Web sémantique au web de données, 2ème partie : retour sur un des articles de Roger T. Pédauque* | *Les petites cases*. http://www.lespetitescases.net/Du-Web-semantique-au-web-de-donnees-2, page consultée le 28/08/2013.

[Pédauque, 2011] PÉDAUQUE, R. I. (2011). *Le web sous tensions (V1) - Espaces Temps*. https://espacestemps.co-ment.com/text/vsMAqHUTfIi/history-version/gzfnfn9JF7H/, page consultée le 28/08/2013.

[Pédauque, 2005] PÉDAUQUE, R. T. (2005). *Le texte en jeu : permanence et transformations du document*. http://archivesic.ccsd.cnrs.fr/sic_00001401.

[Shirky, 2003] SHIRKY, C. (2003). *The Semantic Web, Syllogism, and Worldview*. http://www.shirky.com/writings/herecomeseverybody/semantic_syllogism.html, page consultée le 28/08/2013.

[Shirky, 2005] SHIRKY, C. (2005). *Ontology is Overrated – Categories, Links, and Tags*. http://www.shirky.com/writings/ontology_overrated.html, page consultée le 28/08/2013.

[Sparck-Jones, 2004] SPARCK-JONES, K. (2004). *What's new about the Semantic Web ? Some questions*. http://www.sigir.org/forum/2004D/sparck_jones_sigirforum_2004d.pdf, page consultée le 28/08/2013.

6 Annexe

6.1 The Semantic Web – A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities

Depuis le cabinet du médecin, Lucie donna des instructions à son agent du Web sémantique par le biais de son navigateur. L'agent trouva rapidement l'information concernant le *traitement prescrit* à Maman à partir de l'agent du médecin, parcourut plusieurs listes de *fournisseurs* de services, vérifia la *couverture* d'assurance pour la mère pour trouver un thérapeute dans un rayon de 20 miles de son domicile et prit en considération les *taux de satisfaction* "*excellent*" ou "*très bon*" attribués par des services d'évaluation fiables. Il fallut ensuite essayer de faire coïncider les *heures de rendez-vous* possibles (fournies par les agents de fournisseurs individuels à travers leur site Web) avec les emplois du temps chargés de Pete et Lucy. (Les mots clés en italique indiquent les termes dont la sémantique ou la signification ont été définies pour l'agent à travers le Web sémantique). L'agent leur fournit un plan en quelques minutes. Pete ne fut pas d'accord : l'hôpital de l'université se trouvait de l'autre côté de la ville par rapport au domicile de sa mère, et il lui faudrait la ramener à l'heure de pointe. Il fit refaire la recherche par son propre agent en ajoutant des critères de choix plus précis comme le *lieu* et l'*horaire*. L'agent de Lucy, qui a une *confiance totale* dans l'agent de Pete dans le contexte particulier de cette tâche lui a apporté automatiquement de l'aide en lui fournissant des codes d'accès et des raccourcis à partir des données qu'il avait déjà triées. Un nouveau plan fut présenté instantanément : il y avait une clinique beaucoup plus proche avec des horaires plus matinaux, mais il y avait aussi deux avertissements. D'abord, Pete devrait reprogrammer deux de ses rendez-vous (parmi les moins importants). Il vérifia de quoi il s'agissait : ce n'était pas un problème. L'autre remarque concernait la liste des compagnies d'assurance qui avait oublié d'inclure ce fournisseur (NDLR : la clinique) dans la liste des *thérapeutes médicaux*. "Le type de service ainsi que le statut du plan d'assurance ont été vérifiés de manière sûre par d'autres moyens "le rassura l'agent."Détails ?".

Traduction par Elisabeth Lacombe et Jo Link-Pezet.

6.2 Manifeste d'Ars Industrialis

On n'y comprend rien si l'on ne pose pas que l'esprit, c'est ce qui, toujours déjà, s'est mis en garde contre le vivant, s'est mis hors du vivant pour pouvoir se garder, pour l'avenir de la vie, pour les vivants à venir, et peut-être en se protégeant des vivants présents, du présent vivant. Il y a une extériorisation originaire

de la mémoire de la vie pour l'au-delà de la vie présente, il y a, dans le vécu, du non-vécu, et c'est pourquoi l'esprit est ce qui se transmet, ce dont on hérite, et ce qui nous a toujours déjà précédés. La bibliothèque est l'un des milieux par excellence de cette pré-cédence, et ce milieu est spirituel par excellence - avec ceci cependant qu'il passe dans le réseau, comme milieu associé précisément, et qu'il serait temps de se donner les moyens de penser les conséquences immenses, précisément sur le plan spirituel, de ce devenir à la fois exaltant et effrayant : exaltant comme nouvelle perspective d'action d'une nouvelle forme de puissance publique, effrayant comme risque de voir cette puissance publique ne rien comprendre à ce qui se passe et se joue ici, ne rien comprendre, autrement dit, aux enjeux de ce que nous appelons donc, dans Ars Industrialis, les technologies de l'esprit.

6.3 *Giant Global Graph*

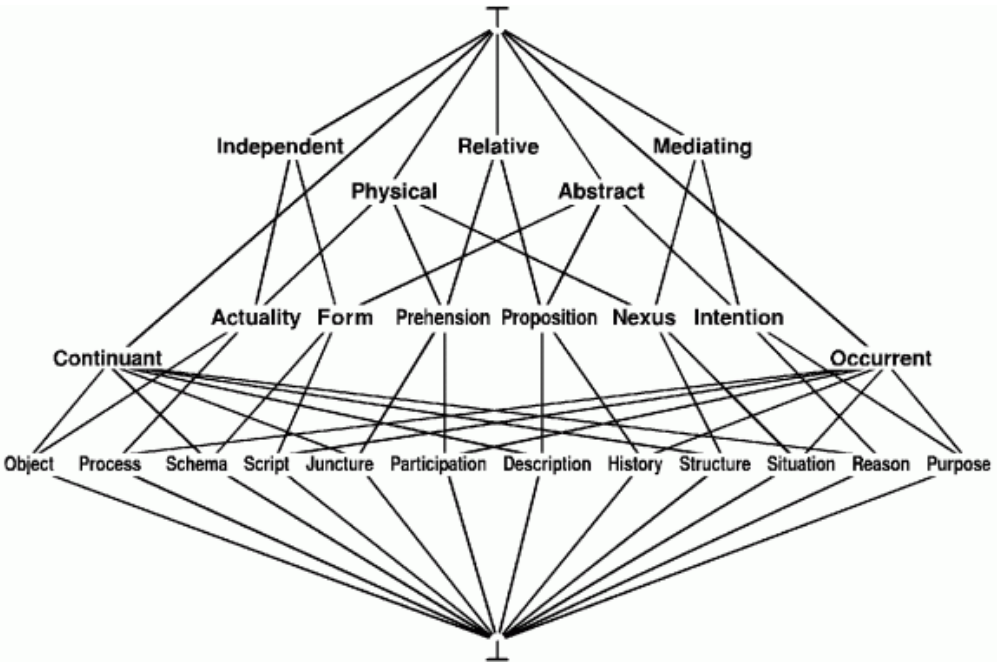
Biologists are interested in proteins, drugs, genes. Businesspeople are interested in customers, products, sales. We are all interested in friends, family, colleagues, and acquaintances. There is a lot of blogging about the strain, and total frustration that, while you have a set of friends, the Web is providing you with separate documents about your friends. One in facebook, one on linkedin, one in livejournal, one on advogato, and so on. The frustration that, when you join a photo site or a movie site or a travel site, you name it, you have to tell it who your friends are all over again. The separate Web sites, separate documents, are in fact about the same thing – but the system doesn't know it.

There are cries from the heart (e.g The Open Social Web Bill of Rights) for my friendship, that relationship to another person, to transcend documents and sites. There is a "Social Network Portability" community. It's not the Social Network Sites that are interesting – it is the Social Network itself. The Social Graph. The way I am connected, not the way my Web pages are connected.

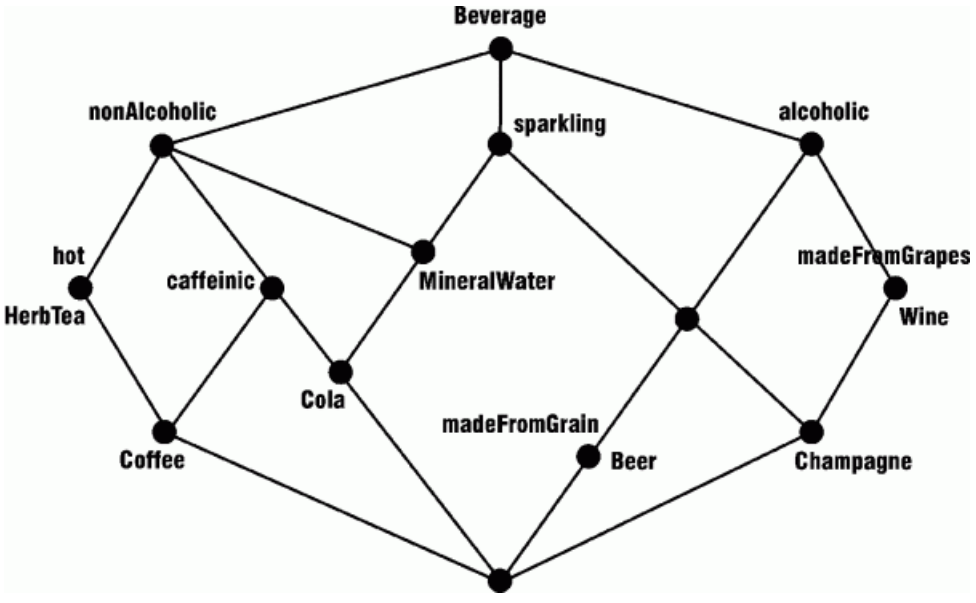
We can use the word Graph, now, to distinguish from Web.

I called this graph the Semantic Web, but maybe it should have been Giant Global Graph ! Any worse than WWWW ? ;-) Not the "Semantic Web" term has been established for a long time, I'm not proposing to change it. But let's think about the graph which it is. (Footnote : "Graph" also happens to be the word the RDF specifications use, but that is by the way. While an XML parser creates a DOM tree, an RDF parser creates an RDF graph in memory.)

6.4 Graphes conceptuels de Sowa



Une ontologie de « haut niveau »



Une ontologie « spécifique à un domaine »

6.5 LOD Cloud

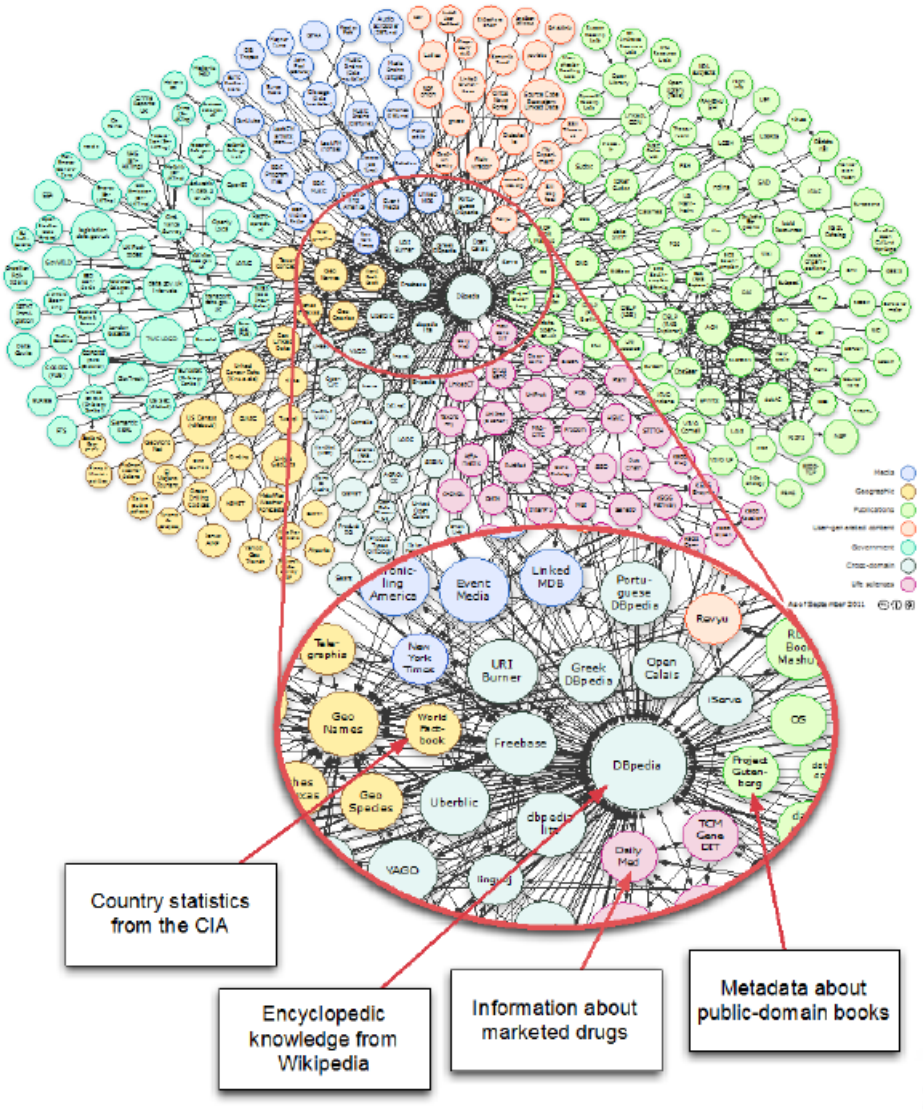


Figure 1.4 The Linked Open Data (LOD) cloud in late 2011: The circles represent freely available datasets and the arrows represent links between them

http://semanticommunity.info/3_Round_Stones

6.6 Logique du sens

Le sens est toujours présupposé dès que *je* commence à parler ; je ne pourrais pas commencer sans cette présupposition. En d'autres termes, je ne dis jamais le sens de ce que je dis. Mais en revanche, je peux toujours prendre le sens de ce que je dis comme l'objet d'une autre proposition dont, à son tour, je ne dis pas le sens (...)

Il y a bien quatre noms dans la classification de Carroll : le nom comme réalité de la chanson ; le nom qui désigne cette réalité, qui désigne donc la chanson, ou qui représente ce que la chanson est appelée ; le sens de ce nom, qui forme un nouveau nom ou une nouvelle réalité ; le nom qui désigne cette nouvelle réalité, qui désigne donc le sens du nom de la chanson, ou qui représente ce que le nom de la chanson est appelé.

G. Deleuze, *Logique du sens* pp. 42-43.

6.7 La pensée sauvage

Cette logique opère un peu à la façon du kaléidoscope : instrument qui contient aussi des bribes et des morceaux, au moyen desquels se réalisent des arrangements structuraux. Les fragments sont issus d'un procès de cassure et de destruction, en lui-même contingent, mais sous réserve que ses produits offrent entre eux certaines homologies : de taille, de vivacité de coloris, de transparence. Ils n'ont plus d'être propre, par rapport aux objets manufacturés qui parlaient un « discours » dont ils sont devenus les indéfinissables débris ; mais, sous un autre rapport, ils doivent en avoir suffisamment pour participer utilement à la formation d'un être d'un nouveau type : cet être consiste en arrangements où, par le jeu des miroirs, des reflets équivalent à des objets, c'est-à-dire où des signes prennent rang de choses signifiées ; ces arrangements actualisent des possibles, dont le nombre, même très élevé, n'est tout de même pas illimité puisqu'il est fonction des dispositions et des équilibres réalisables entre des corps dont le nombre est lui-même fini ; enfin et surtout, ces arrangements, engendrés par la rencontre d'événements contingents (la giration de l'instrument par l'observateur) et d'une loi (celle présidant à la construction du kaléidoscope, qui correspond à l'élément invariant des contraintes dont nous parlions tout à l'heure), projette des modèles d'intelligibilité en quelque sorte provisionnels, puisque chaque arrangement est exprimable sous forme de relations rigoureuses entre ses parties, et que ces relations n'ont d'autre

contenu que l'arrangement lui-même, auquel, dans l'expérience de l'observateur, ne correspond aucun objet (bien qu'il se puisse que, par ce biais, certaines structures objectives soient révélées avant leur support empirique, comme, par exemple, celles des cristaux de neige ou de certains types de radiolaires et de diatomées, à l'observateur qui n'en aurait encore jamais vu).

C. Lévi-Strauss, *La pensée sauvage* p. 51.