



Université Paris IV - Sorbonne
École doctorale V - Concepts et Langages

Technologies du Web Sémantique pour l'Entreprise 2.0

Thèse

Pour l'obtention du grade de

Docteur de l'Université Paris IV - Sorbonne

Discipline: Informatique

Présentée et soutenue publiquement

Le 9 Juin 2009 par

Alexandre Passant

Défendue devant un jury composé de:

- Fabien L. Gandon, INRIA Sophia-Antipolis, *Rapporteur*
- Gilles Kassel, Université de Picardie, *Rapporteur*
- Jean-Pierre Desclés, Université Paris IV - Sorbonne, *Directeur*
- Philippe Laublet, Université Paris IV - Sorbonne, *Co-directeur*
- Ivan Herman, CWI Amsterdam / W3C, *Examinateur*
- François-Xavier Testard-Vaillant, Électricité de France, *Examinateur*

|_|_|_|_|_|_|_|_|_|_|_|_|
(Numéro d'enregistrement attribué par la bibliothèque)



Ce mémoire est mis à disposition sous un contrat Creative Commons "Paternité-Pas d'Utilisation Commerciale-Pas de Modification 2.0 France". Les détails de ce contrat sont disponibles à l'adresse suivante : <http://creativecommons.org/licenses/by-nc-nd/2.0/fr/>

À Julie

Résumé

Cette thèse s'inscrit dans le cadre des récents travaux relatifs à la complémentarité entre Web Sémantique et Web 2.0, deux visions du Web qui ont souvent été considérées, à tort, comme disjointes. Plus particulièrement, nous nous intéressons à l'utilisation des technologies du Web Sémantique (*i.e.* langages, modèles, outils et protocoles) dans le contexte de l'Entreprise 2.0, vision où les outils de plus en plus courants du Web 2.0 (blogs, wikis, services de partage de contenus, pratiques de tagging ...) font leur apparition dans les systèmes d'information organisationnels. Si ces outils facilitent le partage et la collaboration entre individus, dans l'objectif de faire émerger une Intelligence Collective au sein de telles structures, ils introduisent de nouvelles problématiques en termes d'exploitation pertinente des informations produites. D'une part, la diversité des outils utilisés complexifie l'intégration d'informations provenant de diverses sources (blogs, wikis, flux RSS ...) fragmentées au sein du réseau d'entreprise. D'autre part, la nature plein-texte des outils utilisés rend délicate la réutilisation de manière autonome des connaissances ainsi produites, notamment au sein des wikis qui permettent pourtant l'élaboration de bases de connaissances pérennes. Enfin, les pratiques de tagging soulèvent différents problèmes en terme de recherche d'informations, dus notamment à l'ambiguïté et l'hétérogénéité des mots-clés utilisés, ainsi qu'à leur manque d'organisation.

Afin de répondre à ces différentes problèmes et en reprenant l'acronyme *SLATES* (*Search, Links, Authoring, Tags, Extension, Signals*) utilisé pour identifier l'Entreprise 2.0, nous définissons le paradigme *SemSLATES*, proposant la mise en place d'une architecture de médiation sociale et sémantique venant en support d'un ensemble d'outils existants. Cette évolution implique la définition et l'implémentation de différents composants, aussi bien en termes de représentation des connaissances que d'architecture logicielle, composants que nous avons mis en place dans le cadre de cette thèse, en s'appuyant essentiellement sur les technologies du Web Sémantique via les standards du W3C.

Ainsi, nos travaux ont consisté d'une part en la mise en place d'ontologies formelles, aussi bien en terme de métadonnées socio-structurelles (afin de représenter les interactions sociales produites au sein des différents applications utilisées et les contenus issus de ces interactions) que de métadonnées métier (afin d'annoter les contenus eux-mêmes). En ce qui concerne le premier type, nous avons participé activement au projet SIOC – *Semantically-Interlinked Online Communities* –, définissant une ontologie permettant de représenter les activités des communautés en ligne et les contributions associées. En rapport au second point, nous avons défini un certain nombre d'ontologies de domaine, légères et extensibles,

reposant sur des modèles déjà existants et adoptés sur le Web, proposant ainsi certaines bonnes pratiques relatives à la modélisation de telles ontologies. Enfin, afin d'établir un lien entre ces deux niveaux de représentation, nous avons mis en place le modèle MOAT – *Meaning Of A Tag* – permettant de faire le lien entre tags et ressources du Web Sémantique (classes et instances d'ontologies), dans l'objectif de coupler la souplesse des folksonomies et la puissance de l'indexation sémantique basée sur des ontologies. Bien qu'indépendants, l'ensemble de ces modèles s'articule ainsi de manière cohérente afin de prendre en compte les différentes strates de représentations des connaissances nécessaires à de tels écosystèmes sémantiques.

Nous avons également mis en place différents composants logiciels permettant la production et l'exploitation d'annotations sémantiques de manière intuitive pour les utilisateurs finals et communiquant au travers d'un ensemble de protocoles dédiés. En termes de production d'annotations, nous avons développé différents services permettant l'export automatique d'annotations représentées avec SIOC depuis des outils de blogs, wikis et flux RSS dans ce contexte d'entreprise. Nos travaux se sont également concentrés sur la définition d'un service de wiki sémantique afin de permettre une constitution collaborative, ouverte et incrémentale de bases de connaissances formelles reposant sur des ontologies, sans pour autant confronter les utilisateurs à la complexité des modèles sous-jacents. Nous avons également proposé différents services innovants venant tirer parti des graphes d'annotation produits. C'est ainsi le cas d'un moteur de recherche sémantique que nous avons mis en place et qui permet de visualiser des informations (agrégées depuis différents outils d'entreprise) au sujet des instances d'ontologies peuplées depuis les wikis, tout en proposant d'étendre la recherche en considérant l'ensemble des différents graphes d'annotations disponibles au sein du système. Nous avons également proposé de nouvelles manières de visualiser ces informations, notamment au travers d'un système de *mash-up* combinant données internes au système organisationnel et données RDF publiques et reposant sur une interface à facettes.

Alors que l'ensemble de nos recherches ont été validés dans un contexte industriel, la portée de certaines de nos propositions est plus large que ce cadre d'entreprise, et plus généralement que ce contexte d'Entreprise 2.0. Différents travaux ont ainsi été publiés sous forme d'ontologies publiques ou de logiciels libres, permettant leur utilisation à grande échelle sur le Web. Ainsi, ce manuscrit propose, plus globalement, différentes réflexions sur la complémentarité, selon nous nécessaire, entre Web 2.0 et Web Sémantique, pour conduire le Web à son plus haut potentiel.

Mots-clés :

Web 2.0, Entreprise 2.0, Web Sémantique, Ontologies, Folksonomies, Wikis, SIOC, MOAT, Linked Data

Abstract

This Ph.D. thesis is part of some recent works regarding the complementarity between the Semantic Web and the Web 2.0, two visions of the Web that have often been considered, wrongly, as disjoints. Especially, our focus is the use of Semantic Web technologies (*i.e.* languages, models, tools and protocols) in Enterprise 2.0 contexts, a vision in which most of the commonly used Web 2.0 tools (such as blogs, wikis, content-sharing services, tagging practices ...) became popular in corporate information systems.

Yet, while these tools can ease the process of information sharing and collaborations between individuals, with the global aim to create a Collective Intelligence within such structures, they introduce new issues regarding how to efficiently use the information they helped to produce. On the one hand, the nature and diversity of the services used makes the information integration process a complex task, from various sources fragmented in the corporate network (blogs, wikis, RSS feeds ...). On the other hand, the plain-text nature of these tools makes also difficult to reuse the created knowledge, especially regarding wikis, generally used as valuable knowledge bases. Finally, the practice of tagging raises several problems in terms of information retrieval, especially due to the ambiguity and heterogeneity of the tags used, as well as their lack of organization.

In order to solve these different issues and considering the *SLATES* acronym (*Search, Links, Authoring, Tags, Extension, Signals*) used to define the Enterprise 2.0 vision, we defined the *SemSLATES* paradigm, proposing a social semantic middleware architecture on the top of existing services. This proposal implies to define and implement various components, both in terms of knowledge engineering and software architecture, components that we have developed in the context of this Ph.D., relying essentially on Semantic Web technologies via W3C standards.

Hence, our research consisted in defining formal ontologies, in order to model both the socio-structural meta-data (in order to model community interactions happening in these applications as well as the content emerging from these interactions) and business data (in order to annotate the data contained in the application). Regarding the first type of ontologies, we actively participated in the SIOC project – *Semantically-Interlinked Online Communities* – that defines a model to represent activities of online communities and their related contributions. Regarding the second one, we defined several domain ontologies, lightweight, extensible and based on existing and Web-used models, hence defining some good practices regarding lightweight ontologies modeling in such contexte. Finally, in order to provide some relationships between these two levels of knowledge representation, we defi-

ned MOAT – *Meaning Of A Tag* – that allows to create a bridge between tags and Semantic Web resources (*i.e.* aclasses and instances from ontologies) in order to take into account both the flexibility of folksonomies and the power of semantic indexing based on ontologies. While being independant, all these models articulate themselves in a consistent manner in order to take into account the different layers of knowledge representation for such semantic ecosystems.

We also developed several software components (communicating thanks to a set of dedicated protocols) in order to produce and use semantic annotations in a user-friendly way for end-users. In the context of producing semantic annotations, we wrote different services that automatically export SIOC-based annotations from blogs, wikis and RSS feeds in this enterprise context. We also defined a semantic wiki service in order to let end-users participate in a collaborative, open and incremental process to define formal knowledge bases driven by ontologies, without letting these users face the complexity of the underlying models. Moreover, we also designed several innovative services using the produced annotations. We wrote a dedicated semantic search engine allowing to browse information (aggregated from various enterprise sources) related to ontologies instances populated via the wikis and also offering a search extension system by considering the whole graphs of semantic annotations available in the system. We also proposed new ways to browse these information, building a dedicated mash-up system combining internal information and public RDF data and using a faceted browsing interface.

While our research has been done in an industrial context, the scope of our proposals goes further than this corporate context and more generally than the Enterprise 2.0 context. Hence, various works have then been published as public ontologies or free software, allowing to be used at a Web scale. Thus, this thesis suggests, more broadly, different ideas and thoughts regarding the complementarity, in our opinion needed, between Web 2.0 and the Semantic Web, leading the Web to its highest potential.

Keywords :

Web 2.0, Enterprise 2.0, Semantic Web, Ontologies, Folksonomies, Wikis, SIOC, MOAT, Linked Data

Table des matières

Résumé	i
Abstract	iii
Table des matières	v
Tables figures	ix
Liste des tableaux	xiii
Listings	xv
Introduction	1
Contexte et problématique scientifique	1
Contexte de la thèse	1
Motivations et axes de recherche	2
Principaux résultats	4
Organisation du mémoire	6
Plan du mémoire	6
Guide de lecture	8
1 Vers une convergence entre Web Sémantique et Web 2.0	11
Introduction	11
1.1 Formalismes et structures de données avec le Web Sémantique	12
1.1.1 Vers un Web interprétable par les machines	12
1.1.2 Représentation des connaissances avec RDF(S) et OWL	16
1.1.3 Interrogation de données avec SPARQL	25
1.1.4 Web Sémantique et <i>Web of Data</i>	27
1.2 Du consommateur au producteur avec le Web 2.0	30
1.2.1 Une vision participative du Web	30
1.2.2 Blogs, wikis, réseaux sociaux et syndication de contenu	33
1.2.3 Métadonnées sociales : tags et folksonomies	38
1.3 Complémentarité entre les deux domaines	42
1.3.1 Synthèse des deux visions	42

1.3.2	Apports du Web 2.0 pour le Web Sémantique	43
1.3.3	Apports du Web Sémantique pour le Web 2.0	44
Conclusion		46
2	<i>SemSLATES : Une approche sémantique pour l'Entreprise 2.0</i>	49
Introduction		49
2.1	Web collaboratif en entreprise : le projet Athéna	50
2.1.1	Origine et objectifs du projet	50
2.1.2	Répondre efficacement aux différents besoins	53
2.1.3	Complémentarité générale des outils	57
2.1.4	Retour sur expérience	59
2.2	Limites de l'approche classique	62
2.2.1	Fragmentation de l'information et hétérogénéité des formats	62
2.2.2	Capitalisation des connaissances	63
2.2.3	Tags et recherche d'information	63
2.2.4	Synthèse des problèmes rencontrés	68
2.3	Écosystème sémantique pour l'Entreprise 2.0	69
2.3.1	Web Sémantique et méthodologie <i>SemSLATES</i>	69
2.3.2	Définition d'une architecture sociale de médiation sémantique	71
2.3.3	Modèles, adaptateurs et services	73
2.3.4	Situation de l'approche vis-à-vis de l'état de l'art	77
Conclusion		81
3	Rôle et définition d'un ensemble d'ontologies pour l'Entreprise 2.0	83
Introduction		83
3.1	Métadonnées socio-structurelles pour le Web 2.0 avec SIOC	84
3.1.1	Identification des Besoins	84
3.1.2	Positionnement par rapport à de l'art	86
3.1.3	Présentation du modèle de représentation SIOC	89
3.1.4	Alignement avec des vocabulaires existants	94
3.1.5	SIOC, FOAF et la portabilité des données Web 2.0	96
3.1.6	Adoption du modèle et évaluation	101
3.2	Modélisation des ontologies métier	104
3.2.1	Besoins en termes de représentation métier	104
3.2.2	FOAF pour la représentation des personnes physiques et morales	104
3.2.3	Localisation avec Geonames	107
3.2.4	Ontologies des rôles et utilisation de SKOS	109
3.2.5	Articulation globale des différentes ontologies métier	117
3.3	MOAT pour lier tags et ontologies	120
3.3.1	Tags, folksonomies et ontologies : un état de l'art	120
3.3.2	Représentation de la signification des tags avec MOAT	127
3.3.3	Modèle de représentation MOAT	128
3.3.4	Positionnement de MOAT par rapport à l'état de l'art	134
Conclusion		135

4 Annotations sémantiques et peuplement collaboratif d'ontologies	137
Introduction	137
4.1 Annotation sémantique de documents Web 2.0	138
4.1.1 Une approche automatisée pour l'annotation socio-structurelle	138
4.1.2 Implémentation au sein de la plate-forme Hermès	139
4.1.3 API SIOC et passage à l'échelle de l'annotation socio-structurelle de documents Web 2.0	143
4.2 UfoWiki pour le peuplement d'ontologies métier	148
4.2.1 Wikis sémantiques et peuplement d'ontologies : intérêt et état de l'art	148
4.2.2 Objectifs, principes et architecture d'UfoWiki	154
4.2.3 Architecture logicielle	156
4.2.4 Utilisation d'UfoWiki et peuplement collaboratif d'ontologies . .	160
4.2.5 Evaluation de l'outil et statistiques d'utilisation	166
4.3 Du tagging à l'indexation sémantique	170
4.3.1 Processus d'indexation sémantique associé à MOAT	170
4.3.2 Implémentations logicielles	174
4.4 Retour sur l'utilisation de MOAT dans notre contexte d'Entreprise 2.0 . .	182
Conclusion	183
5 Intégration et utilisation d'annotations sémantiques distribuées	185
Introduction	185
5.1 Stockage des données et protocoles associés	186
5.1.1 De la nécessité d'un entrepôt de données	186
5.1.2 Besoins et choix de l'entrepôt	190
5.1.3 Protocoles de communication	192
5.2 Enrichissement des fonctionnalités des wikis	196
5.2.1 Utilisation de macros sémantiques pour l'utilisation d'annotations .	196
5.2.2 Contextualisation des macros pour augmenter le potentiel de veille .	201
5.2.3 Interfaces avancées de visualisation et <i>mash-ups</i> sémantiques . .	203
5.3 Interopérabilité entre applications via les annotations	207
5.3.1 Intégration des contenus des blogs au sein des wikis	207
5.3.2 Indexation de flux RSS guidée par les annotations	209
5.3.3 Projection de connaissances pour l'aide à la veille technologique .	211
5.4 Recherche sémantique pour l'Entreprise 2.0	212
5.4.1 Recherche d'information et Web Sémantique	212
5.4.2 Mise en place d'un moteur de recherche exploitant ontologies et annotations	213
5.4.3 Suggestion de concepts et de contenus proches	216
Conclusion	221
Conclusion générale	223
Retour sur les impacts de la thèse	223
Perspectives et réflexions	226

A Requête SPARQL pour la traduction de données RSS vers SIOC	229
B Ontologie des rôles	231
C Analyse de propriétés DBpedia	233
D Exemple d'annotations métier produites avec UfoWiki	235
E Exemple d'annotations socio-structurelles produites avec UfoWiki	239
Bibliographie	243

Table des figures

0.1	Organisation des chapitres	9
1.1	Proposition d'architecture distribuée qui conduira au <i>World Wide Web</i>	13
1.2	Pile du Web Sémantique, Février 2008	15
1.3	Représentation graphique de triplets RDF	18
1.4	Graphes nommés et identification de l'auteur d'un ensemble de triplets	20
1.5	Nuage de données du projet Linking Open Data	28
1.6	Le document en tant que support de données pour le Web Sémantique	29
1.7	L'écosystème Web 2.0	31
1.8	Etat de la blogosphère, Avril 2007	34
1.9	Le Web en tant que plate-forme, l'exemple de RSS	38
1.10	Actions de <i>tagging</i> combinées autour d'une même photo	40
1.11	Exemple de nuage de tags (Delicious)	41
1.12	Web 2.0 pour le Web Sémantique	44
1.13	Web Sémantique pour le Web 2.0	45
1.14	Convergence entre Web Sémantique et Web 2.0	46
2.1	Utilisation de Twitter par le service Web 2.0 Slideshare pour communiquer avec ses utilisateurs	52
2.2	Interface personnelle de visualisation de flux RSS au sein d'Hermès	54
2.3	Coconstruction de connaissances avec les wikis	56
2.4	Scénario idéal d'utilisation des différents éléments de publication de la plate-forme	58
2.5	Évolution des billets et des commentaires sur la plate-forme	60
2.6	Résultats d'une recherche associée au tag <i>apple</i> sur Flickr	64
2.7	Tags suggérés par cooccurrence sur Delicious	66
2.8	Distribution des tags au sein de notre folksonomie	67
2.9	Annotations sémantiques en support d'un système d'Entreprise 2.0 existant selon trois niveaux d'annotations	69
2.10	Architecture de médiation sémantique pour l'Entreprise 2.0	72
2.11	Représentation unifiée des métadonnées documentaires avec SIOC	75
2.12	Architecture RDF Bus	80
3.1	Intégration de données hétérogènes réparties avec SIOC	86

3.2	Le modèle de classes et propriétés de SIOC	91
3.3	Comptes utilisateur et personne physique avec SIOC et FOAF	95
3.4	Interopérabilité entre données sociales avec SIOC et FOAF	98
3.5	Unification de réseaux sociaux distribués avec owl :sameAS	99
3.6	Visualisation uniforme de réseaux sociaux distribués	100
3.7	Utilisation combinée de FOAF et OpenID pour identifier un profil utilisateur avec SparqlPress	100
3.8	Statistiques de production de données SIOC sur le Web	102
3.9	Taxonomie des sous-classes d'Agent dans Proton	106
3.10	Relations géographiques entre entités et transitivité de la propriété locatedIn de Geonames	110
3.11	Distinction entre taxonomies et ontologies	114
3.12	Taxonomies de domaines en OWL-Full	114
3.13	Taxonomies de domaines en OWL-Lite	115
3.14	Taxonomies de domaines avec SKOS	117
3.15	Combinaison d'ontologies et base de connaissance associée pour définir des assertions au sujet d'EDF	119
3.16	Tags et actions de <i>tagging</i> avec la <i>Tag Ontology</i>	123
3.17	Modélisation quadripartite de deux relations de <i>tagging</i> au sein d'une folksonomie	129
3.18	Significations globales du tag <i>apple</i> avec MOAT	130
3.19	Représentation de la signification locale du tag <i>apple</i> avec MOAT et DBpedia	132
3.20	Modèle de représentation MOAT	133
3.21	Articulation d'ontologies pour l'Entreprise 2.0	136
4.1	Processus générique de production de données RDF depuis des services Web 2.0	139
4.2	Processus de traduction RSS / Atom vers SIOC	141
4.3	Processus de traduction des données de blogs et wikis vers SIOC	144
4.4	Exemple de traduction d'un billet de blog vers SIOC	144
4.5	Représentation de liens rdfs :seeAlso entre documents RDF avec l'API SIOC .	146
4.6	Cartographie de réseaux sociaux avec FOAFMap	148
4.7	Du wiki au Web Sémantique	149
4.8	Interactions entre annotations documentaires et annotations métier dans Ufo-Wiki	157
4.9	Association d'un type de page à une classe avec UfoWiki	158
4.10	Création de formulaire pour une classe donnée avec UfoWiki	159
4.11	Architecture d'un wiki au sein d'UfoWiki	160
4.12	Sélection d'un type de contenu avec UfoWiki	161
4.13	Édition d'une page wiki pour la création d'instance via UfoWiki	162
4.14	Gestion d'une taxonomie de domaines avec UfoWiki	163
4.15	Production d'annotations basées sur Geonames avec UfoWiki	165
4.16	Statistiques d'utilisation d'UfoWiki : Pages et instances	168

4.17	Statistiques d'utilisation d'UfoWiki : Pages, instances et triplets	169
4.18	<i>Framework</i> utilisateur MOAT	171
4.19	<i>Workflow</i> client / serveur et processus MOAT	173
4.20	Interface utilisateur du module MOAT pour Drupal couplée au <i>widget</i> Sindice	174
4.21	Choix d'un concept pour désambiguïser un tag au sein du client MOAT Athéna	176
4.22	Parcours de la taxonomie des classes pour définir une nouvelle signification .	177
4.23	Création d'une nouvelle instance et association d'un tag via le client MOAT .	177
4.24	Visualisation des différents tags associés à un concept	178
4.25	Architecture de LODr	178
4.26	Assignation d'une URI à un tag particulier avec LODr	179
4.27	Nuage de concepts avec LODr	180
5.1	Vision globale des actions, annotations et ontologies d'un écosystème sémantique pour l'Entreprise 2.0	187
5.2	Répartition des ontologies et annotations au sein du système	188
5.3	Architecture associée à PTSW pour l'indexation et la découverte de documents RDF sur le Web Sémantique	194
5.4	doap :store : Annuaire et interface de visualisation de projets logiciels modélisés avec DOAP	195
5.5	Protocoles d'abstraction au-dessus de l'entrepôt de données du médiateur .	196
5.6	Processus d'interprétation des macros au sein d'UfoWiki	198
5.7	Résultat d'une macro sémantique listant l'ensemble des associations recensées au sein d'un wiki	201
5.8	Résultat d'une macro contextualisée	202
5.9	URIs partagées entre graphes d'annotations	203
5.10	Sélection de facettes à partir de différentes ontologies	204
5.11	Visualisation à facettes d'un wiki avec Exhibit	205
5.12	Interface à facettes pour visualiser des données SIOC avec SMOB	205
5.13	Géolocalisation d'un ensemble d'acteurs avec Exhibit et Geonames	206
5.14	Géolocalisation au sein d'une macro contextualisée	207
5.15	Interopérabilité entre applications via l'utilisation d'annotations sémantiques	208
5.16	Projection de connaissances sur des contenus internes	211
5.17	Choix d'un concept à partir d'un terme de recherche	214
5.18	Rendu du moteur de recherche sémantique au sein d'Hermès	215
5.19	Accès au moteur de recherche via les concepts identifiés avec MOAT	216
5.20	Identification de contenus proches via des relations entre concepts associés .	217
5.21	Identification des domaines plus spécifiques qu' <i>'énergie solaire</i>	218
5.22	Identification d'acteurs proches de Gaz de France selon une règle prédéfinie .	219
5.23	Suggestion de concepts proches au sein de LODr	220
5.24	Système de recommandations musicales basées sur DBpedia	221
5.25	Vision du Web axée sur une convergence <i>humain-machine-humain</i>	225

Liste des tableaux

1.1	Règles d'inférence RDFS	23
1.2	Caractéristiques comparées du Web Sémantique et du Web 2.0	42
2.1	<i>SLATES</i> et la plate-forme Hermès	57
2.2	Utilisateurs et contributeurs au sein d'Hermès	59
2.3	Statistiques des flux RSS au sein d'Hermès	59
2.4	Statistiques des contributions utilisateur au sein d'Hermès	60
2.5	Tags utilisés pour le concept de Web Sémantique sur Delicious	65
2.6	Distribution des tags au sein de la plate-forme Hermès	67
2.7	Problématiques soulevés par l'approche <i>SLATES</i> classique au sein d'Hermès	68
2.8	Fonctionnalités comparées de <i>SLATES</i> et <i>SemSLATES</i>	70
3.1	Eléments du module Types de SIOC	93
3.2	Comparaison de différentes ontologies pour la représentation des tags et des objets associés	125
3.3	Situation de MOAT par rapport à l'état de l'art	135
4.1	Positionnement d'UfoWiki par rapport à d'autres wikis sémantiques	167
4.2	Distribution des tags au sein de la plate-forme Hermès	183
5.1	Associations entre URIs et termes contrôlées par les utilisateurs	210

Listings

1.1	Représentation Turtle de triplets RDF	17
1.2	Représentation RDF/XML de triplets RDF	17
1.3	Exemple d'assertions modélisées avec RDFA	18
1.4	Exemple de base de connaissances associée à une ontologie	22
1.5	Exemple d'ontologie représentée en RDFS et sérialisée en Turtle	23
1.6	Exemple de requête SPARQL SELECT	25
1.7	Exemple de requête SPARQL CONSTRUCT	26
1.8	Exemple de requête SPARQL ASK	26
1.9	Exemple de requête SPARQL DESCRIBE	26
1.10	Exemple de flux RSS 2.0	37
2.1	Représentation d'assertions au sujet d'EDF	75
3.1	Exemple de contenu Web 2.0 avec SIOC	91
3.2	Exemple de requête SPARQL dédiée à SIOC	92
3.3	Exemple de billet de blog avec SIOC et son module Types	93
3.4	Utilisation de propriétés issues du DublinCore avec SIOC	94
3.5	Règle d'inférence pour lier SIOC et FOAF, représentée en N3	95
3.6	Extension de FOAF pour la gestion de différents types d'agents	107
3.7	Modélisation de partenariats entre agents	107
3.8	Localisation d'une entreprise avec FOAF et le Geo Vocabulary	108
3.9	Définition de la propriété locatedIn de Geonames	109
3.10	Modèle simple pour la représentation des rôles	111
3.11	Modèle pour la représentation des rôles avec prise en compte du métier et du domaine	112
3.12	Association d'un rôle à un agent	112
3.13	Modèle complet pour la représentation des rôles	116
3.14	Ensemble d'assertions au sujet d'EDF à l'aide de différents modèles	118
3.15	Significations globales du tag "apple" avec MOAT	131
3.16	Signification locale du tag "apple" avec MOAT	131
3.17	Règle d'inférence pour MOAT, représentée en N3	132
4.1	Utilisation de Jena pour représenter des données RDF	145
4.2	Requête interne au sein de MediaWiki	152
5.1	Requête SPARQL pour l'interrogation de données SIOC via un moteur supportant les principes d'inférence RDFS	192

5.2	Restriction d'une requête SPARQL aux graphes produits par un wiki donné	199
5.3	Fonction PHP et requête SPARQL associées à une macro UfoWiki	200
5.4	Requête SPARQL avec contextualisation des macros	202
5.5	Requête SPARQL pour identifier des billets annotés avec un concept particulier	209
5.6	Identification de pages associées à un concept proche	215
5.7	Règle d'inférence pour identifier deux contenus proches en utilisant MOAT, SIOC et des relations entre URIs	217
5.8	Règle d'inférence basée sur SKOS pour l'identification de concepts proches	218
5.9	Règle d'inférence pour l'identification de concepts proches à partir de relations entre domaines	219

To a computer, the Web is a flat, boring world, devoid of meaning. This is a pity, as in fact documents on the Web describe real objects and imaginary concepts, and give particular relationships between them. For example, a document might describe a person. The title document to a house describes a house and also the ownership relation with a person. 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. Only when we have this extra level of semantics will we be able to use computer power to help us exploit the information to a greater extent than our own reading.

*Tim Berners-Lee, Présentation "W3 future directions"
1st World Wide Web Conference, Genève, Mai 1994*

Introduction

CONTEXTE ET PROBLÉMATIQUE SCIENTIFIQUE

Contexte de la thèse

Les travaux présentés dans ce mémoire s'inscrivent dans le cadre d'une thèse effectuée en contrat CIFRE¹ en collaboration entre le LaLIC², Université Paris-Sorbonne (Paris IV) et le centre de Recherche et Développement d'Electricité de France (EDF R&D par la suite) à Clamart⁴. Nous avons ainsi été rattachés à EDF R&D de Février 2005 à Mai 2008, au sein de trois services successifs, poursuivant ensuite nos travaux à part entière au LaLIC puis au DERI⁵, *National University of Ireland, Galway*, à partir de Septembre 2008.

Si ce contexte industriel nous a parfois amené à chercher un compromis entre impératifs à court ou moyen terme et recherche scientifique, il nous a cependant permis de confronter nos travaux à des situations réelles. Ainsi, nous avons pu tester nos différentes hypothèses et les outils associés au sein d'un système déployé en grandeur nature, nous permettant de prendre en compte les retours utilisateur pour affiner certains choix. Ceci nous a en outre conduit à une certaine rigueur et à essayer le plus souvent possible d'envisager des solutions évolutives et adaptées à un nombre croissant d'utilisateurs. Si cette composante appliquée nous a conduits dans certains cas à développer des solutions *ad hoc* pour l'entreprise, nous avons fait en sorte de toujours garder à l'esprit une problématique de recherche plus large de manière à généraliser nos résultats à l'échelle du Web, comme nous le verrons tout au long de ce mémoire. Ainsi, si la plupart des travaux présentés ici trouvent leur motivation et s'articulent globalement dans un contexte d'Entreprise 2.0, la portée de certains d'entre eux s'avère plus large que ce cadre industriel. Il nous a en effet semblé pertinent de considérer cette thèse CIFRE non pas comme un vase clos, mais comme un contexte d'expérimentation de ce qu'il est possible de réaliser à plus grande échelle sur le Web Sémantique, notamment en faisant le choix dès le début de nous baser sur les différents langages et recommandations du W3C⁶.

¹Conventions Industrielles de Formation par la Recherche

²Langages, Logique, Informatique et Cognition – <http://www.lalic.paris4.sorbonne.fr/>

³L'ensemble des liens hypertexte de cette thèse ont été vérifiés à la date du 26 Janvier 2009.

⁴EDF R&D dispose de trois sites sur le territoire français, rassemblant plus de 2000 chercheurs. Plus d'un millier d'entre eux sont situés sur le site de Clamart, sur des thématiques aussi diverses que les énergies renouvelables ou la sécurité informatique au sein des centrales nucléaires. – <http://retd.edf.fr>

⁵Digital Enterprise Research Institute – <http://deri.org>

⁶World Wide Web Consortium – <http://w3c.org>

Enfin, d'un point de vue plus général, il est important de mentionner que nous sommes arrivés au Web Sémantique (et aux travaux de recherche présentés dans cette thèse) par attrait pour le Web et par volonté de participer, à notre échelle, à l'évolution de ce formidable médium. C'est d'ailleurs à la suite d'un IUP Génie Mathématiques et Informatique et d'un DESS Technologies de l'Internet pour les Organisations, accompagnés en parallèle de plusieurs années d'expérience en tant qu'ingénieur développement Web que nous avons décidé de reprendre le chemin des études pour mener une thèse sur le sujet. Un DEA Informatique et Systèmes Intelligents⁷ nous a ainsi amené à découvrir la notion d'ontologies dédiées à la modélisation de données sur le Web avant de poursuivre sur un stage relatif à l'annotation sur le Web Sémantique au LaLIC, point de départ de nos travaux. Notre expérience passée autour des technologies du Web et notre passion pour celui-ci nous semblent importants à signaler dans la mesure où ils permettent de comprendre certains choix relatifs à nos travaux. Nous défendons ainsi dans ce mémoire une vision assez pragmatique du Web Sémantique, et plus généralement une vision appliquée de la recherche. C'est en effet selon nous en combinant recherche et standardisation autour de technologies clés associées à un contexte applicatif fort que l'on parviendra à mener le Web à son plein potentiel.

Motivations et axes de recherche

Les travaux présentés dans ce mémoire s'inscrivent dans la lignée des recherches autour du Web Sémantique et du Web 2.0, deux visions récentes d'une certaine évolution du Web. Plus particulièrement, nous nous intéressons à la manière dont celles-ci peuvent co-habiter et bénéficier chacune des apports de l'autre. Alors qu'elles ont souvent, à tort, été considérées comme disjointes, il nous semble au contraire pertinent d'étudier en quoi leur complémentarité permettra de conduire à un Web basé sur un ensemble d'interactions sociales entre internautes et aux données interprétables sans ambiguïté par des agents logiciels autonomes.

C'est en envisageant cette complémentarité que l'on pourra à terme proposer de nouveaux services innovants en termes d'intégration, de visualisation et de recherche d'information sur le Web, alors considéré comme une immense base de données sociale et distribuée. Plus particulièrement, l'étude de cette convergence nous a amené à approfondir nos travaux en fonction de trois thématiques principales, dont nous présenterons de manière succincte différents résultats dans la seconde partie de cette introduction.

La modélisation des métadonnées socio-structurelles associées aux outils Web 2.0

Si le Web 2.0 a introduit de nouvelles pratiques sociales en termes d'échange d'informations et d'émergence de communautés en ligne, la diversité des applications et des services introduits nous confronte inévitablement à une hétérogénéité des formats de modélisation. Chaque outil ou service dispose en effet de ses propres modèles de données, rendant de ce fait complexes l'intégration, l'échange et la recherche d'information à partir de sources multiples. Si cette diversité est problématique dans un contexte comme celui du Web, elle l'est également dans des environnements plus restreints utilisant ces mêmes outils, tels que

⁷Celui-ci, tout comme l'IUP et le DESS évoqués précédemment, a été suivi à L'Université Paris-Dauphine (Paris IX).

les systèmes d'informations d'Entreprise 2.0 où un accès pertinent à l'information est nécessaire. Ainsi, une partie de nos travaux a consisté en la définition de modèles pour permettre la représentation commune des métadonnées socio-structurelles associées aux outils Web 2.0 via l'utilisation de technologies du Web Sémantique. Par représentation des métadonnées socio-structurelles, nous entendons à la fois la modélisation de notions documentaires et structurelles (distinguer par exemple un billet de blog d'une page wiki, identifier le lien entre une page wiki et le wiki associé, etc.) et celle des interactions sociales qui s'y rapportent (commentaire sur un blog, édition d'une page wiki, etc.). De tels modèles permettent de disposer d'annotations sémantiques partagées depuis des systèmes hétérogènes, facilitant ainsi l'intégration de contenus depuis différentes plates-formes et en conséquence la recherche d'information associée.

La représentation de connaissances termino-ontologiques et le peuplement d'ontologies de domaine à partir d'outils Web 2.0

Alors que le point précédent se concentre sur des aspects documentaires et sociaux, il est également important de prendre en compte le contenu même de ces documents Web 2.0. Si l'on se réfère aux définitions actuelles du Web Sémantique telles que mises en avant par le W3C – "The Semantic Web is a Web of Data"⁸ – , il s'agit donc de passer de documents aux représentations des données du monde réel qu'ils contiennent. Par exemple, nous souhaitons modéliser à partir d'une page wiki intitulé LaLIC qu'il s'agit d'un laboratoire de recherche basé à Paris, *i.e.* passer du document et du terme à la représentation du concept associé. S'il s'agit ici de thématiques connues de peuplement d'ontologies, ou de manière plus large de représentations de connaissances termino-ontologiques, la problématique qui nous intéresse ici est la prise en compte de l'utilisateur final dans cette démarche, notamment au travers d'outils Web 2.0. Alors que le Web 2.0 facilite la production de contenus documentaires, nous avons souhaité approfondir la manière dont il permet la création, l'évolution et le partage de données, toujours au sens *Web of Data*, via ces outils Web 2.0. Plus particulièrement nous nous sommes ici intéressés :

- à l'utilisation de wikis pour le peuplement d'ontologies, en étudiant de quelle manière ces outils permettent un peuplement ouvert, collaboratif et évolutif d'ontologies de domaine ;
- aux relations entre les systèmes d'indexation libre (et spontanée) à base de tags et des processus d'indexation sémantique plus classiques où les termes d'indexation sont liés à des ressources termino-ontologiques.

Nos travaux dans ce domaine nous permettent ainsi d'envisager en quoi les outils et les processus du Web 2.0 peuvent faciliter l'émergence de données représentées selon les principes du Web Sémantique.

L'exploitation de graphes d'annotations sémantiques pour l'interopérabilité, la mise en commun et la recherche d'information

Enfin, une troisième thématique que l'on peut extraire de nos travaux et qui vient en corollaire des deux précédentes est l'exploitation de graphes d'annotations sémantiques pour

⁸<http://w3c.org/2001/sw>

proposer de nouveaux services à valeur ajoutée aux utilisateurs finals. Une des problématiques du Web Sémantique est en effet le problème classique de *la poule et l'œuf* : il est nécessaire de disposer de données pour en montrer toute la puissance mais il est également nécessaire de disposer d'outils les exploitant pour inciter à leur production. Afin de mettre ce cercle vertueux en place, différentes questions se posent, principalement vis à vis des outils permettant l'exploitation de ces annotations :

- de quelle manière utiliser un nombre croissant d'annotations distribuées dans un objectif de signalement pertinent d'information ?
- comment masquer à l'utilisateur la complexité des graphes d'annotations et des algorithmes de parcours et de requêtes associés ?
- comment mettre en avant les résultats obtenus pour que l'utilisateur final prenne conscience de la valeur des données produites et accentue cette démarche de production ?

Ainsi, si l'on devait résumer nos motivations et la problématique scientifique de cette thèse en une phrase synthétique, nous pourrions reformuler de la manière suivante : *Comment combiner Web Sémantique et Web 2.0 afin de tirer profit d'interactions sociales issues d'outils du Web 2.0 pour la représentation et l'exploitation de connaissances formalisées selon les principes du Web Sémantique*? Notons également, comme le titre de ce mémoire l'indique, que nos motivations autour de cette convergence entre Web Sémantique et Web 2.0 sont liées à l'essor récent de la notion d'Entreprise 2.0, qui met en avant l'utilisation des technologies et principes du Web 2.0 au sein de la sphère professionnelle.

Principaux résultats

Réflexions sur la complémentarité entre Web 2.0 et Web Sémantique

De manière générale, nous avons détaillé à travers nos travaux en quoi cette complémentarité entre Web 2.0 et Web Sémantique nous paraissait nécessaire pour conduire à un Web où les interactions sociales sont omniprésentes dans un objectif de production de données interprétables et interopérables. Ainsi, nous avons montré en quoi le Web Sémantique et ses formalismes de représentation des connaissances (au sens RDF(S)/OWL) ne s'opposaient pas – bien au contraire – à l'utilisation d'outils et de principes Web 2.0 [Passant et Laublet, 2008c]. Nos réflexions se sont portées entre autres sur l'utilisation couplée d'ontologies et de bases de connaissances en support de systèmes à base de tags et de folksonomies [Passant *et al.*, 2006] [Passant, 2007c], ou encore sur l'utilisation de wikis sémantiques pour permettre un peuplement d'ontologies collaboratif, évolutif et ouvert [Passant et Laublet, 2008e]. Dans ces deux cas, il nous semble important de signaler que nous avons pris en compte le rôle actif de l'utilisateur, proposant ainsi une vision du Web Sémantique pensée pour l'utilisateur final aussi bien en termes de production que d'utilisation d'annotations sémantiques.

Ces réflexions sur la complémentarité entre Web 2.0 et Web Sémantique ont également donné lieu à l'organisation de différents ateliers [Giboin *et al.*, 2008] [Breslin *et al.*, 2008], la participation à plusieurs tutoriels sur le sujet dans des conférences comme WWW⁹ ou ISWC¹⁰ et la corédaction d'un livre sur le sujet [Breslin *et al.*, 2009].

⁹World Wide Web Conference – <http://www.iw3c2.org/>

¹⁰International Semantic Web Conference – <http://iswc.semanticweb.org/>

Modèles de représentation

Afin de mettre en pratique ces réflexions, nous nous sommes attachés à la définition de différentes ontologies permettant de modéliser à la fois les activités, les interactions et les contenus créés par des communautés Web 2.0 à l'aide de technologies du Web Sémantique. Bien que vouées à des utilisations distinctes, ces différentes ontologies s'articulent de manière complémentaire au sein d'une architecture de médiation sémantique pour l'Entreprise 2.0.

En termes de modélisation des métadonnées socio-structurelles, nous avons ainsi contribué activement à SIOC – *Semantically-Interlinked Online Communities* [Breslin *et al.*, 2005] –, de ses débuts à sa Soumission Membre au W3C en Juin 2007 [Berrueta *et al.*, 2007], en tant que coauteur de la spécification et éditeur de deux documents associés. Concernant nos travaux autour de la complémentarité entre ontologies et tags, nous avons défini le modèle MOAT – *Meaning Of A Tag* [Passant et Laublet, 2008b] – permettant de résoudre les problèmes classiques des systèmes à base de tags via l'utilisation de bases de connaissances formelles venant en support des folksonomies. Enfin, de manière plus proche des besoins de cette convention CIFRE, nous avons également développé plusieurs vocabulaires permettant la représentation des connaissances métier, en se basant notamment sur des modèles publics et abondamment utilisés sur le Web Sémantique et en proposant certaines bonnes pratiques dans ce contexte.

Ainsi, nos différentes réflexions en termes de modèles de représentation ont été bénéfiques aussi bien dans le contexte d'entreprise de cette thèse que de manière plus large sur le Web.

Réalisations logicielles

En plus des modèles évoqués précédemment, nos travaux ont également conduit à la réalisation de différentes implémentations logicielles. Si celles-ci sont liées aussi bien au contexte d'entreprise de notre thèse qu'à des développements plus larges sur le Web, elles ont toutes en commun l'objectif de mettre en avant ce lien fort entre Web Sémantique et Web 2.0.

D'une part, nous avons mis en place un ensemble d'outils pour l'Entreprise 2.0 agrémentés de modules permettant la production automatisée d'annotations sémantiques, notamment à partir de blogs, ainsi qu'un serveur de wikis sémantiques permettant la représentation de données métier formalisées selon les principes du Web Sémantique [Passant et Laublet, 2008d]. En termes d'utilisation de ces annotations, nous avons développé différents services de visualisation de données RDF ainsi qu'un moteur de recherche sémantique pour l'entreprise venant exploiter ontologies et annotations sémantiques pour la recherche de documents annotés. Cette architecture logicielle, proposée sous la forme d'un médiateur sémantique pour l'Entreprise 2.0 [Passant, 2008a], combine ainsi outils et principes du Web 2.0 pour la production et visualisation d'annotations et technologies du Web Sémantique pour la représentation de celles-ci.

D'autre part, nous avons développé différentes applications Web dans cet objectif de convergence entre Web 2.0 et Web Sémantique, certains développements ayant été mutualisés avec les outils mis en place en entreprise, comme par exemple différents *plug-in* pour

la production d'annotations sémantiques depuis le système Drupal en utilisant les vocabulaires SIOC et MOAT. Nous avons également proposé une API permettant de généraliser la production automatique d'annotations sémantiques socio-structurelles avec SIOC [Bojārs *et al.*, 2006], ainsi que des applications comme LODr [Passant, 2007a], permettant d'appliquer les principes de MOAT à des contenus Web 2.0 issus de services comme Flickr ou Delicious, ou SMOB, service de microblogging ouvert et décentralisé reposant entièrement sur les standards et technologies du Web Sémantique [Passant *et al.*, 2008]. En termes de visualisation de données, nous pouvons également citer FOAFMap [Passant, 2006], un des premiers services de *mash-up* sémantique, proposant la géolocalisation de réseaux sociaux modélisés en RDF.

ORGANISATION DU MÉMOIRE

Plan du mémoire

Ce manuscrit est découpé en cinq chapitres auxquels viennent s'ajouter cette introduction et une conclusion. Si le plan général ne suit pas une approche traditionnelle qui consiste à introduire l'état de l'art puis nos travaux et leur évaluation, chacun des chapitres reviendra sur ces différents aspects en fonction du domaine abordé. Ce mémoire, qui peut se considérer à la fois comme un ensemble de propositions autour de la convergence entre Entreprise 2.0 (et plus généralement Web 2.0) et Web Sémantique et comme l'étude d'un cas pratique autour de cette convergence, s'organise ainsi de la manière suivante.

Chapitre 1: Vers une convergence entre Web Sémantique et Web 2.0, page 11

Ce premier chapitre introduira les notions de Web Sémantique et de Web 2.0, essentielles pour la bonne compréhension de ce mémoire. Dans la première partie, nous présenterons un bref historique du Web et introduirons ensuite les fondements du Web Sémantique. Nous expliciterons RDF et la notion d'URIs pour la représentation de données, l'utilisation de RDFS et OWL pour la définition d'ontologies et l'utilisation de SPARQL pour l'interrogation de données. Nous reviendrons également sur le projet *Linking Open Data* et la vision d'un *Web of Data*, notamment par rapport au Web tel que nous le connaissons aujourd'hui. La seconde partie détaillera la notion de Web 2.0 et les principaux changements introduits par celui-ci. Nous présenterons tout d'abord les principes généraux de cette vision participative du Web, puis introduirons différents composants qui seront au cœur de nos travaux parmi lesquels blogs, wikis et systèmes d'annotation à base de tags. Enfin, nous présenterons un aperçu général de la convergence possible entre ces deux domaines. Nous conclurons ainsi ce chapitre en introduisant certains des travaux qui seront détaillés par la suite dans ce mémoire, comme la notion de modèles communs pour les outils Web 2.0 ou l'utilisation de wikis sémantiques pour le peuplement d'ontologies.

Chapitre 2: SemSLATES : Une approche sémantique pour l'Entreprise 2.0, page 49

Nous introduirons le chapitre suivant en présentant la notion d'Entreprise 2.0 et le système d'information initial que nous avons mis en place au sein d'EDF. Nous identifierons ensuite ses limites, qui motivent nos travaux relatifs à la méthodologie *SemSLATES* que nous

avons définie et qui sera détaillée dans ce chapitre. Nous présenterons ainsi l'apport d'une architecture de médiation sémantique dans ce contexte d'Entreprise 2.0, architecture venant se greffer au dessus de l'existant sans pour autant remettre en cause celui-ci. Nous verrons en quoi l'ajout de différents composants logiciels sur des outils déjà présents permet de bénéficier d'une sémantique commune qui ouvre la voix à une interopérabilité accrue entre applications. Nous comparerons également notre proposition à certains travaux similaires, et tâcherons de montrer en quoi notre approche nous semble novatrice et pertinente par rapport à l'état de l'art. Ce chapitre nous permettra également d'introduire les trois chapitres suivants, qui détailleront les différents aspects nécessaires pour mener à bien cette approche, à savoir (1) des modèles communs de représentation, (2) des outils d'annotations sémantiques et de peuplement d'ontologies et (3) des services exploitant ces ontologies et bases de connaissances.

Chapitre 3: Rôle et définition d'un ensemble d'ontologies pour l'Entreprise 2.0, page 83

Ce troisième chapitre présentera en détail différentes ontologies que nous avons mises en place dans ce contexte d'Entreprise 2.0, en distinguant les modèles axés sur la représentation de données métier et ceux mis en place pour la représentation des structures documentaires et des interactions sociales sur le Web 2.0. La première partie présentera principalement nos travaux autour de SIOC, modèle pour la représentation des métadonnées socio-structurelles pour les outils et communautés Web 2.0. Nous présenterons d'autres modèles poursuivant un but similaire et détaillerons l'alignement de SIOC avec des vocabulaires existants. Nous aborderons également le rôle de SIOC vis-à-vis des problématiques de portabilité des données sociales. La seconde partie présentera ensuite les différentes ontologies de domaine utilisées dans notre architecture de médiation. Alors que ces modèles sont par nature dépendants du contexte applicatif, il nous semble utile de revenir dessus notamment pour expliciter en quoi l'utilisation et l'extension de vocabulaires existants nous semble une bonne pratique dans un contexte d'entreprise. Nous présenterons également certaines problématiques de modélisation d'ontologies auxquelles nous avons été confrontées, et comment nous y avons fait face. Nous détaillerons ensuite nos travaux en matière de représentation des tags et plus particulièrement la définition de MOAT, modèle permettant de prendre en compte et de modéliser la signification des tags via des concepts du Web Sémantique, offrant ainsi la possibilité d'établir un lien souple entre folksonomies et ontologies. Cette partie sera également l'occasion de comparer ce modèle aux autres ontologies permettant la représentation des tags et des folksonomies mais aussi de faire le parallèle avec les approches permettant l'enrichissement sémantique de folksonomies de manière automatique ou semi-automatique.

Chapitre 4: Annotations sémantiques et peuplement collaboratif d'ontologies, page 137

Après avoir présenté les différents modèles utilisés dans de tels écosystèmes sémantique, nous détaillerons dans ce quatrième chapitre les moyens mis en place pour permettre leur peuplement et ainsi produire les annotations sémantiques s'y rattachant. Nous détaillerons ainsi les différentes extensions que nous avons mises en place pour les outils existants, en présentant également certains de nos efforts plus généraux pour simplifier l'annotation sémantique et le peuplement d'ontologies depuis des services Web 2.0. Ce chapitre,

plus technique que le précédent, nous permettra tout d'abord de présenter les processus de production automatique d'annotations sémantiques modélisées avec SIOC depuis des outils existants. Nous nous attarderons ensuite sur notre prototype de wiki sémantique, Ufo-Wiki, notamment sur la manière dont il permet de coupler la création d'annotations socio-structurelles et le peuplement d'ontologies métier. Enfin, nous expliciterons les processus participatifs associés à MOAT, permettant l'indexation sémantique de contenus à partir de systèmes à base de tags et détaillerons différentes implémentations logicielles associées, utilisées aussi bien dans ce contexte de médiation pour l'Entreprise 2.0 que sur le Web.

Chapitre 5: Intégration et utilisation d'annotations sémantiques distribuées, page 185

Après avoir présenté la définition de différentes ontologies pour l'Entreprise 2.0 et la production des annotations sémantiques associées, nous détaillerons leur utilisation. Nous reviendrons tout d'abord sur le caractère distribué de ces annotations et le besoin de disposer d'une architecture nous permettant facilement d'effectuer des requêtes sur celles-ci via un entrepôt de données centralisé. Ceci nous permettra de présenter les différents protocoles de communication mis en place, à la fois en termes d'agrégation de données et d'exploitation de celles-ci. Nous présenterons ensuite différents services venant enrichir les outils existants par l'intermédiaire des annotations produites. Nous détaillerons principalement (1) l'enrichissement des wikis sémantiques via un système de macros, (2) l'utilisation d'interfaces de visualisation avancées et la mise en place de *mash-ups* sémantiques et (3) la réalisation d'un moteur de recherche sémantique venant s'intégrer à cette architecture. Plus particulièrement, nous insisterons dans ce chapitre sur la manière dont ces outils permettent de masquer la complexité des requêtes et des modèles et langages utilisés à l'utilisateur final.

Chapitre 5.4.3: Conclusion générale, page 223

Enfin, nous conclurons ce mémoire en revenant sur les différents travaux présentés et la manière dont ils répondent aux problématiques initiales, tout en essayant de porter un regard critique sur ceux-ci. Nous envisagerons également certains travaux futurs qu'il nous semble important de garder à l'esprit dans cette perspective globale de convergence entre Web Sémantique et Web 2.0.

Guide de lecture

Afin de guider le lecteur dans le parcours de ce mémoire, nous proposons le guide de lecture suivant (Figure 0.1, page 9). Pour un aperçu global de nos travaux, on pourra se limiter aux deux premiers chapitres qui donneront une vision générale et synthétique de nos recherches. Le premier chapitre présente ainsi les différentes notions manipulées alors que le second donne une aperçu global de nos problématiques de recherche et des solutions apportées. Les trois chapitres suivants détaillent en profondeur nos travaux et peuvent par ailleurs se considérer comme un tout permettant d'approfondir les thèmes abordés dans le second chapitre.

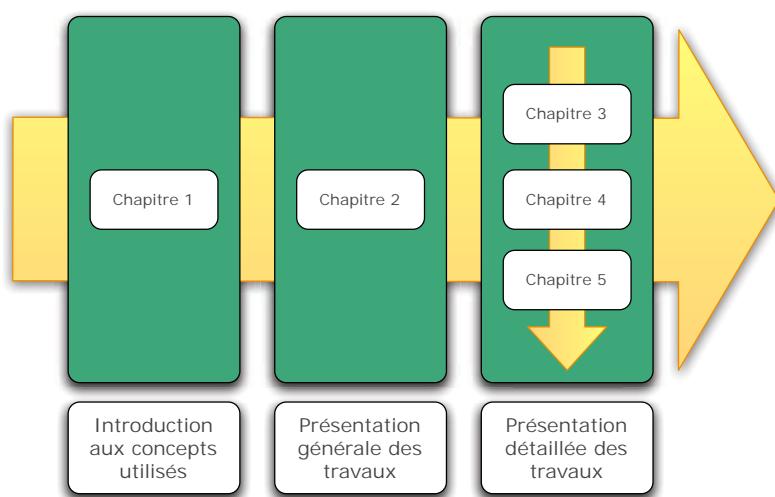


Figure 0.1: Organisation des chapitres

Chapitre 1

Vers une convergence entre Web Sémantique et Web 2.0

INTRODUCTION

Ces dernières années ont vu la montée en puissance de deux visions du Web, que l'on pourrait à première vue considérer comme disjointes. La première, le Web Sémantique, propose une *extension* de celui-ci définissant des formalismes de représentations unifiées pour les données dans une optique d'échange et de compréhension de celles-ci par les agents logiciels [Berners-Lee *et al.*, 2001]. L'autre, communément appelée Web 2.0¹, est beaucoup plus pragmatique et met l'accent sur la place centrale de l'utilisateur au sein de la démarche de production d'information [O'Reilly, 2005]. Elle met en avant les échanges, l'ouverture et la collaboration entre internautes par l'intermédiaire d'outils et services simples d'utilisation.

Dans ce chapitre, nous introduirons tout d'abord les principes du Web Sémantique et des formalismes de représentation associés, tout en revenant plus particulièrement sur certains aspects qui nous paraissent fondamentaux pour la bonne compréhension de ce mémoire. Nous introduirons donc la notion d'URI et présenterons le langage RDF (Section 1.1.2, page 16), qui permet de représenter les données sur le Web Sémantique selon la vision du W3C, avant d'aborder les notions de vocabulaires et d'ontologies ainsi que les langages associés, à savoir RDFS et OWL (Section 1.1.2, page 21). Si ces différents points sont relatifs à la production de données, ou annotation sémantique, il nous semble également intéressant de présenter les mécanismes relatifs à leur interrogation. Nous expliciterons ainsi l'utilisation du langage et protocole SPARQL (Section 1.1.3, page 25), celui-ci jouant un rôle important dans l'avènement du Web Sémantique et plus concrètement dans les outils que nous allons présenter par la suite. Enfin, nous aborderons l'initiative *Linking Open Data*, qui vise à traduire en RDF et interconnecter un grand nombre de données présentes sur le Web, dans une vision plus pragmatique du Web Sémantique et de ce que l'on appelle maintenant plus communément *Web of Data* (Section 1.1.4, page 27).

Dans la seconde partie du chapitre, nous présenterons ce qui caractérise le Web 2.0 et expliciterons en quoi cette vision n'introduit selon nous pas de révolution technologique majeure (particulièrement en termes de représentation des connaissances) mais en contre-

¹Nous ne discuterons pas l'utilisation de ce terme. Gardons simplement à l'esprit que, malgré les appellations, il n'y a qu'un seul Web.

partie modifie de manière profonde la façon dont les contenus sont publiés et échangés en ligne (Section 1.2, page 30). Cette rupture concernant la production d'informations en ligne, qui est donc plus sociale que technologique s'accompagne d'un certain nombre d'outils que nous présenterons ici. En particulier, nous détaillerons deux outils phares de cette mouvance, à savoir les blogs (Section 1.2.2, page 33) et les wikis (Section 1.2.2, page 35), ainsi que la notion de réseaux sociaux (Section 1.2.3, page 41), les principes de syndication de contenu (Section 1.2.2, page 36) et la notion de *tagging* (Section 1.2.3, page 38), méthode collaborative, incrémentale et ouverte de catégorisation. Ces différents points étant au cœur des travaux qui seront présentés par la suite dans ce mémoire, il nous semble important de bien détailler leur fonctionnement et d'entrevoir certaines de leurs limites que nous présenterons par la suite (Section 2, page 49).

Enfin, nous indiquerons dans la troisième partie de ce chapitre pourquoi il nous semble utile, voire nécessaire, de faire cohabiter ces deux visions pour parvenir à terme à un Web où l'utilisateur est au centre de la production de données, mais où celles-ci sont représentées de manière unifiée afin d'automatiser, ou tout du moins de simplifier, certaines tâches (Section 1.3, page 42). Nous reviendrons ici sur les préjugés supposés entre ces deux visions avant d'étudier cette convergence, qui conduira à des espaces informationnels combinant principes Web 2.0 et technologies du Web Sémantique. Ainsi, nous présenterons d'une part quels peuvent être les avantages du Web 2.0 pour le Web Sémantique, essentiellement en termes d'interfaces d'édition et d'annotations sémantiques et d'autre part les avantages du Web Sémantique pour le Web 2.0, cette fois-ci en termes de structuration de données et de formats d'échange. Ces deux aspects nous permettront ainsi de voir de quelle manière cette convergence conduit à un cercle vertueux entre Web Sémantique et Web 2.0. Cette dernière partie du chapitre permettra également d'entrevoir plus en détail les travaux qui seront développés dans la suite de ce mémoire, à savoir l'utilisation des technologies du Web Sémantique pour modéliser et structurer les données issues de services Web 2.0, de manière à enrichir leurs fonctionnalités.

1.1 FORMALISMES ET STRUCTURES DE DONNÉES AVEC LE WEB SÉMANTIQUE

1.1.1 Vers un Web interprétable par les machines

En 1989, Tim Berners-Lee imagine pour le CERN² une architecture informatique distribuée permettant d'interconnecter les différents éléments du système d'information interne [Berners-Lee, 1989]. Il représente alors celui-ci comme un graphe où les noeuds, tout comme les arcs, sont typés et peuvent ainsi représenter (pour les noeuds) des outils, des documents, des projets ou des personnes ou bien encore (pour les arcs) des relations de production, d'inclusion ou d'appartenance. Afin de faciliter la navigation dans un tel système, sa proposition se base sur l'utilisation de l'hypertexte, tel que défini par Ted Nelson dès les années 60 au sein du projet Xanadu³ [Nelson, 1965]. C'est cette proposition d'architecture décentralisée qui donnera par la suite naissance au *World Wide Web* tel que nous le connaissons aujourd'hui.

²Organisation européenne pour la recherche nucléaire – <http://cern.ch>

³<http://www.xanadu.com/>

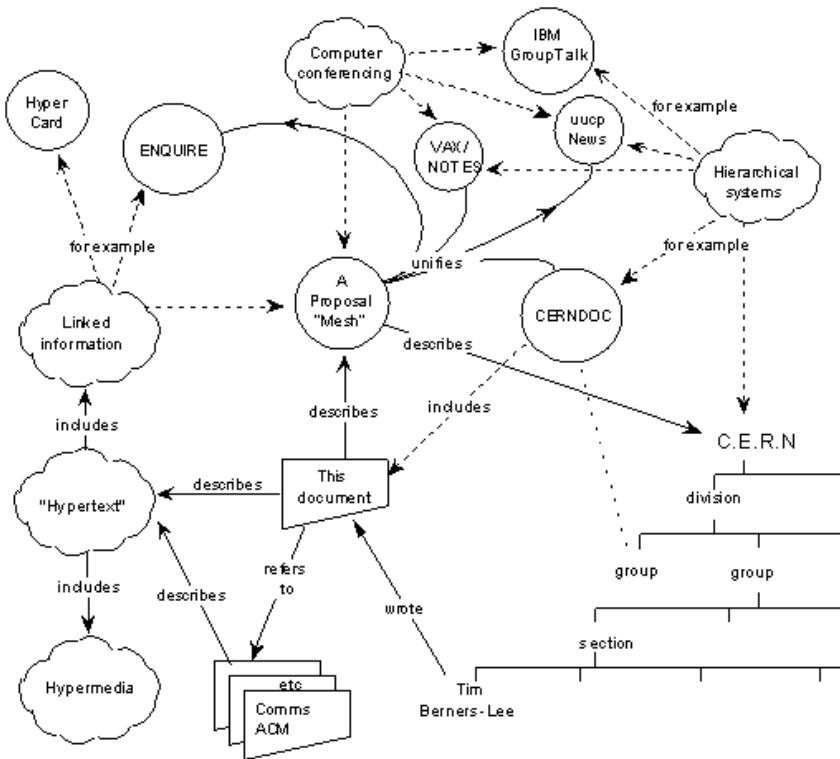


Figure 1.1: Proposition d'architecture distribuée qui conduira au *World Wide Web* [Berners-Lee, 1989]

Si l'on observe le schéma correspondant à cette vision d'origine du Web (Figure 1.1, page 13) et que l'on prend en compte l'état actuel de celui-ci, on ne peut s'empêcher de constater que là où la proposition initiale fait état de ressources et de liens fortement typés, le Web tel que nous le connaissons aujourd'hui ne considère que des documents, qu'ils soient textuels ou multimédia et des liens hypertextes non typés pour établir des relations entre ceux-ci. Ainsi, si un utilisateur est en mesure d'identifier le concept induit par un document (une personne ou un projet donné ...) ainsi que la nature du lien défini entre deux concepts (à partir des liens entre documents), cette identification n'est pas réalisable de manière simple par un agent logiciel. En effet, celui-ci ne considère que des documents plein-texte (encodés dans un langage dont il ne sait pas interpréter la sémantique) connectés entre eux par des hyperliens unidirectionnels non typés. De plus, les métadonnées associées à ces documents (auteur, date de création ...) sont elles aussi difficilement interprétables. Enfin, même pour un utilisateur, ces interprétations peuvent-être biaisées puisqu'elles font appel à l'expérience, la culture, et l'affect mental de celui-ci, qui peut différer selon les personnes pour un même document.

Ainsi se pose le problème d'un Web interprétable non seulement par les humains mais surtout par les machines. C'est en ce sens que se situe l'initiative du Web Sémantique qui vise à résoudre cette problématique d'interprétation des données par les agents logiciels :

"The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" [Berners-Lee et al., 2001]. C'est donc bien d'*extension* et non pas de refonte dont il est question pour définir ce Web compréhensible par les machines⁴. On parle également de *Web de Données* (*Web of Data*) afin d'évoquer la façon dont celui-ci permet de modéliser sur le Web des représentations interprétables de *données* et non plus uniquement de *documents* au sujet de ces données. Nous reviendrons un peu plus tard sur cet aspect (Section 1.1.4, page 27).

Cette évolution du Web repose sur la présence d'annotations sémantiques, permettant de modéliser de manière formelle (1) les métadonnées (date de création, auteur, etc.) associées aux documents présents sur le Web et (2) les données présentes au sein de ces documents. Ces annotations sémantiques, qui permettent ainsi d'envisager l'interprétation des contenus en ligne, sont envisageables à partir du moment où l'on dispose :

- d'une part d'un *modèle commun* pour représenter les ressources sur le Web. C'est le rôle joué par l'utilisation des URIs – *Uniform Resource Identifier* [Berners-Lee et al., 2005] – couplées à RDF – *Ressource Description Framework* [Klyne et Carroll, 2004] – (Section 1.1.2, page 16) ;
- d'autre part de vocabulaires permettant de définir de manière formelle, mais surtout interprétable et interopérable, la sémantique de ces données. Les *ontologies*, au sens informatique du terme [Gruber, 1995], jouent ici un rôle important. Nous verrons plus loin comment modéliser des ontologies sur le Web Sémantique avec des langages RDFS – RDF Schema [Brickley et Guha, 2004] – et OWL – *Web Ontology Language* [Patel-Schneider et al., 2004] – (Section 1.1.2, page 21).

Nous verrons par la suite que ces annotations peuvent être produites selon différents objectifs, de l'indexation de documents à la modélisation du contenu de ceux-ci, les deux approches pouvant également être associées (Section 2.3.1, page 69).

Si cette initiative est aujourd'hui essentiellement guidée par les travaux du W3C, via différents groupes de travail et efforts de standardisation menés depuis 2001⁵, il est important de signaler d'autres travaux précurseurs, notamment SHOE⁶ [Heflin et Hendler, 2000]. Ce projet intègre en effet différents composants permettant de rendre le contenu de pages Web compréhensible et exploitable par des agents logiciels :

- un *langage* – SHOE : *Simple HTML Ontology Extensions*⁷ [Luke et Heflin, 2000] – défini sous forme d'extension de HTML et permettant d'inclure directement des données interprétables au sein de pages Web. Celui-ci permet d'une part de modéliser les données mais aussi de définir leur sémantique via la description d'*ontologies* (Section 1.1.2, page 21) au sein des pages ;
- un *agent* – Exposé – permettant de retrouver sur le Web les différentes pages annotées pour les stocker ensuite dans un système dédié - PARKA [Rager et al., 1997], sur lequel

⁴La machine n'interprétant qu'une succession de 0 et 1, il est délicat de parler réellement de compréhension par les machines. On devrait plutôt parler de contraintes d'interprétation, comme le souligne [Bachimont, 2000] en évoquant la notion d'engagement ontologique. On utilisera cependant ce terme compréhension par abus de langage au sein de ce mémoire.

⁵<http://www.w3.org/2001/sw>

⁶<http://www.cs.umd.edu/projects/plus/SHOE/>

⁷<http://www.cs.umd.edu/projects/plus/SHOE/spec.html>

il est possible d'effectuer différentes *requêtes* via un langage spécifique – PIQ.

On retrouve bien dans la vision actuelle du Web Sémantique des similarités avec cette approche combinant (1) des langages de description de données et de modélisation d'ontologies comme RDF(S)/OWL et (2) des langages de requêtes comme SPARQL (Section 1.1.3, page 25) et l'utilisation d'entrepôts de données RDF. À ceux-ci viennent s'ajouter des notions de logique formelle, de preuve et de confiance utilisées à terme par différentes applications et reprenant certains principes de l'Intelligence Artificielle [Russell et Norvig, 2003] (Figure 1.2, page 15).

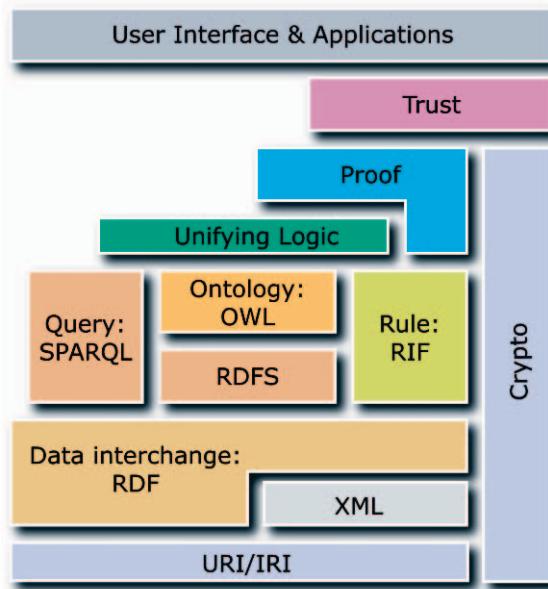


Figure 1.2: Pile du Web Sémantique, Février 2008⁸

Pour terminer cette introduction au Web Sémantique et avant de détailler les différents formalismes de représentation utilisés dans ce contexte, nous signalerons les travaux de visionnaires comme Vannevar Bush et le Memex [Bush, 1945], Ted Nelson et Xanadu, ou encore Douglas Engelbart et ses propositions de systèmes informatiques pour augmenter l'efficience intellectuelle [Engelbart, 1962] ou ses travaux sur l'*Open Hyperdocument System*⁹ [Engelbart, 1990]. Ceux-ci imaginaient il y a plusieurs dizaines d'années déjà des méthodes pour unifier et connecter des représentations du monde réel via des relations typées, couplées à des processus de navigation dans ces représentations. C'est également ce que [Berners-Lee, 1989] proposait dans sa vision d'origine d'un système d'informations interconnectées : "The system we need is like a diagram of circles and arrows, where circles and arrows can stand for anything". Nous pensons que les travaux du Web Sémantique permettront à terme de réaliser ces visions d'un système où l'information est universellement accessible,

⁸<http://www.w3.org/2001/sw/>

⁹<http://www.csli.sri.com/projects/ohs/>

interconnectée mais surtout définie avec une sémantique formelle et interprétable par des agents logiciels autonomes, de manière à proposer de nouveaux services innovants notamment en termes de navigation et de recherche d'information. C'est également de cette manière que les *social machines* définies par [Berners-Lee et Fischetti, 1999] pourront également voir le jour, dans un modèle unifié d'interactions entre humains et machines.

1.1.2 Représentation des connaissances avec RDF(S) et OWL

Avant-propos

Nous présenterons ici uniquement des formalismes proposés ou standardisés via les activités du W3C, formalismes que nous utilisons par ailleurs au sein des différents travaux présentés dans cette thèse. Pour d'autres modes de représentation des connaissances, en particulier les Topic Maps [Biezunski *et al.*, 2002] [Auillans *et al.*, 2002] et leur utilisation sur le Web Sémantique, le lecteur pourra se référer à la thèse [Amardeilh, 2007].

Représentation des ressources : les URIs et RDF

RDF – *Ressource Description Framework* [Klyne et Carroll, 2004] – est un élément fondamental du Web Sémantique puisqu'il permet de représenter des ressources sur le Web de manière uniforme pour les agents logiciels là où ceux-ci ne voient dans un document texte qu'une succession de caractères inexploitables. Pour ce faire, chaque ressource est identifiée de manière universelle par une URI, qui peut être assignée aussi bien à (1) une donnée présente sur le Web (un document, un compte utilisateur sur un service donné ...), (2) un objet du monde réel (un pays, une personne ...) auquel on souhaite associer un identifiant dans ce contexte de représentation en ligne, ou encore (3) une relation (l'appartenance, la filiation ...). Par exemple :

- <http://example.org/blog/112> identifie un billet de blog sur un site donné ;
- <http://sws.geonames.org/3017382/> identifie la France en tant que zone géographique ;
- <http://apassant.net/alex> identifie l'auteur de ce mémoire (et non sa page personnelle) ;
- <http://www.w3.org/2000/01/rdf-schema#label> identifie la relation qui lie une ressource à son label.

Afin de décrire ces ressources, RDF se base sur la notion de triplets, permettant de définir des assertions au sujet de celles-ci. Chaque triplet se compose de :

- un *sujet*, *i.e.* la ressource à laquelle on assigne une propriété, identifiée par une URI ;
- un *prédicat*, *i.e.* la propriété assignée à la ressource, également identifiée par une URI ;
- un *objet*, *i.e.* la valeur de la propriété. Celle-ci peut être de type primitif (chaîne de caractère, entier ...) ou être à nouveau une ressource. Elle peut ainsi être à son tour sujet d'un autre triplet conduisant à la formation d'un graphe, les nœuds tout comme les arcs étant représentés par des URIs. Tim Berners-Lee considère ainsi le Web Sémantique comme un *Giant Global Graph* par analogie avec le *World Wide Web*¹⁰, dans le sens où il connecte des ressources typées via des propriétés identifiées, là où le Web connecte simplement des documents via des liens hypertextes (Section 1.1.4, page 27).

¹⁰<http://dig.csail.mit.edu/breadcrumbs/node/215>

Différentes sérialisations permettent de représenter des assertions modélisées en RDF. C'est le cas de N3 [Berners-Lee, 2006b], Turtle [Beckett et Berners-Lee, 2008] (sous-dialecte du précédent), RDF/XML¹¹ [Beckett, 2004], ou encore des représentations graphiques¹². Ainsi, les deux exemples de code et la figure qui suivent (Figure 1.3, page 18) définissent les mêmes informations qui se traduisent par "*EDF est une organisation située en France*", information constituée dans cet exemple de deux triplets¹³, la sérialisation RDF/XML étant elle sous forme condensée. Nous remarquerons aussi dans cet exemple l'utilisation de préfixes et d'espaces de noms ainsi que la présence du raccourci N3 a utilisé pour rdf:type.

```

@prefix foaf: <http://xmlns.com/foaf/0.1> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix geonames: <http://geonames.org/ontology#> .
@prefix athena: <http://athena.edf.fr/data/> .

athena:EDF a foaf:Organization ;
  geonames:locatedIn <http://sws.geonames.org/3017382/> ;
  rdfs:label "Électricité de France" .

```

Listing 1.1: Représentation Turtle de triplets RDF

```

<rdf:RDF
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:geonames="http://geonames.org/ontology#"
  <foaf:Organization rdf:about="http://athena.edf.fr/data/EDF"
    >
    <geonames:locatedIn rdf:resource="http://sws.geonames.org
      /3017382/">
    <rdfs:label>Électricité de France</rdfs:label>
  </foaf:Organization>
</rdf:RDF>

```

Listing 1.2: Représentation RDF/XML de triplets RDF

Si les annotations sémantiques sont en général représentées sous la forme de documents RDF indépendants des éventuels documents (X)HTML associés, on peut remarquer que l'ajout de métadonnées directement au sein de pages Web (comme le proposait SHOE)

¹¹Une erreur courante est ainsi de présenter RDF comme une application (au sens schéma ou DTD) de XML, alors que RDF/XML est uniquement une des sérialisations possibles de celui-ci. Par ailleurs, la syntaxe de RDF/XML diffère en quelques points de la syntaxe XML classique [Beckett, 2004].

¹²Dans ce cas, l'interprétation est impossible à moins d'utiliser un format graphique interprétable comme SVG (*Scalable Vector Graphics*) [SVG Working Group, 2003].

¹³Nous aurions très bien pu écrire cette affirmation à l'aide d'un ou au contraire de quatre triplets en fonction des modèles utilisés (Section 3.2, page 104).

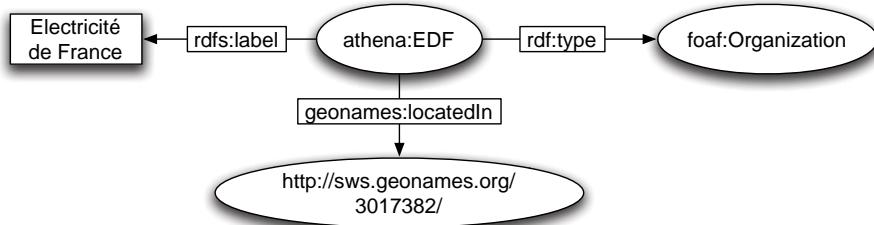


Figure 1.3: Représentation graphique de triplets RDF

est aujourd’hui au cœur de différents travaux. En effet, représenter les annotations au sein de documents annexes introduit généralement un problème de duplicité d’informations. Dans l’exemple précédent, on peut supposer que le fait de définir la chaîne de caractère "Electricité de France" comme valeur pour `rdfs:label` est redondant avec une information déjà présente au sein de la page Web associée, certes en (X)HTML mais avec cette même chaîne de caractères (par exemple dans une balise `<h1>`). Des travaux comme eRDF¹⁴ ou RDFa [Adida et Birbeck, 2008] permettent ainsi l’inclusion directe d’annotations RDF au sein de documents (X)HTML, le second se basant sur l’introduction de nouveaux attributs XHTML pour y parvenir, comme le montre l’exemple ci-dessous (Listing 1.3, page 18).

```
<html xmlns="http://www.w3.org/1999/xhtml"
      xmlns:foaf="http://xmlns.com/foaf/0.1/"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
      xmlns:geonames="http://geonames.org/ontology#>
<body about="http://athena.der.edf.fr/data/EDF" typeof="foaf:Organization">
  <h1 property="rdfs:label">Electricité de France</h1>
  <p>
    EDF est située en <a rel="geonames:locatedIn" href="http://
      sws.geonames.org/3017382/">France</a>.
  </p>
</body>
</html>
```

Listing 1.3: Exemple d’assertions modélisées avec RDFa

Dans cette même optique d’annotations intégrées au sein même des pages, nous pouvons également citer également les microformats¹⁵, effort communautaire qui offre aussi la possibilité de définir certaines données structurées (événements, contacts ...) au sein de pages Web via de simples attributs de balises. Ceux-ci ne sont malheureusement pas aussi puissants que RDF(S)/OWL en termes d’expressivité (subsomption, inférence ...), mais sont

¹⁴<http://research.talis.com/2005/erdf/wiki/Main/RdfInHtml>

¹⁵<http://microformats.org>

néanmoins utilisés plus fréquemment sur le Web. De plus, ceux-ci ne bénéficient pas de la même ouverture que les ontologies, puisqu'un microformat ne peut évoluer qu'après consensus de la communauté. Ces différentes limites leurs valent parfois le nom de *lower-case semantic web*, en opposition au Web Sémantique et ses modèles plus formels. Néanmoins, l'utilisation de GRDDL – *Gleaning Resource Descriptions from Dialects of Languages* – [Connolly, 2007] permet de faire le pont entre ces différentes visions. GRDDL offre en effet la possibilité de traduire différents dialectes XML en RDF et permet ainsi de transformer un document XHTML contenant des microformats ou des annotations RDFa en données RDF brutes qui peuvent être utilisées comme n'importe quelles données RDF natives.

Pour en revenir aux assertions RDF elles-mêmes, il est également possible de considérer un ou plusieurs triplets RDF comme source(s) de nouveaux triplets, par exemple pour définir la date à laquelle une assertion a été établie. Si une première approche pour modéliser ce processus se base sur l'utilisation des principes de réification RDF¹⁶, celle-ci introduit différents problèmes (notamment une explosion du nombre de triplets [Carroll et Stickler, 2004]) que [Carroll *et al.*, 2005] permettent de résoudre avec l'utilisation des graphes nommés (*named graphs*). La notion de graphes nommés étend celle de graphe RDF (*i.e.* un ensemble de triplets¹⁷) en permettant d'assigner à chacun une URI propre. Cette URI permet de considérer chaque graphe comme une ressource à part entière et donc de l'utiliser comme sujet d'une nouvelle relation. Il est ainsi possible de modéliser l'auteur d'un ensemble de triplets (Figure 1.4, page 20) ou encore de certifier les informations via un système de signature de graphes [Carroll, 2003] dans une optique de confiance des sources d'informations comme définie par la pile du Web Sémantique (Figure 1.2, page 15). Malgré ces avantages et en raison de la structure par triplets de RDF, l'utilisation des graphes nommés au sein de documents RDF est complexe et nécessite une évolution des syntaxes actuelles. Les extensions TRIX¹⁸ [Carroll et Stickler, 2004] ou TRIG [Bizer et Cyganiak, 2007] permettent de modéliser ces graphes nommés respectivement en RDF/XML et Turtle. [Bottollier *et al.*, 2007] ont proposé une nouvelle manière de procéder via l'utilisation d'une propriété spécifique (<http://www.inria.fr/acacia/corese#graph>) pour indiquer la source d'un ensemble de triplets au sein de documents RDF/XML. En pratique cependant, une manière simple de procéder à l'identification de ces sources et de regrouper les triplets dans un document accessible en ligne est de considérer l'URL du dit document comme l'URI du graphe source. Ces méthodes sont en outre toutes compatibles avec l'utilisation de la clause GRAPH au sein de requêtes SPARQL (Section 1.1.3, page 25).

Il est également important lorsqu'on modélise une ressource sur le Web Sémantique, de faire la distinction entre son URI (*i.e.* son identifiant) et l'URL du ou des documents la décrivant, qu'il s'agisse d'un document RDF regroupant un certain nombre d'assertions à son sujet ou d'une description (X)HTML. On considère ainsi à ce sujet [Lewis, 2007] :

- les ressources informationnelles (un document, un billet de blog ...) pour lesquelles l'URL du document peut correspondre à l'URI de son identifiant. Il est en effet cohérent de considérer que le document identifié par cette URI correspond au document

¹⁶<http://www.w3.org/TR/rdf-mt/#ReifAndCont>

¹⁷<http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/#dfn-rdf-graph>

¹⁸<http://sw.nokia.com/trix/TriX.html>

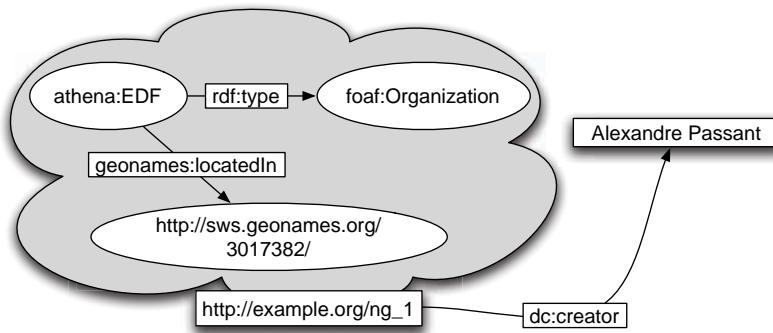


Figure 1.4: Graphes nommés et identification de l'auteur d'un ensemble de triplets

situé à cette même adresse ;

- les ressources non-informationnelles, *i.e.* des données monde réel (une personne, un pays ...) que l'on souhaite représenter sur le Web et où la distinction est nécessaire. On ne peut en effet pas considérer que <http://fr.wikipedia.org/wiki/France> correspond à l'identifiant de la France, puisqu'on a d'un côté un document Web et de l'autre un pays¹⁹.

Ainsi :

- <http://sws.geonames.org/3017382> correspond à une URI identifiant la France (et non pas un document à son sujet) ;
- <http://sws.geonames.org/3017382/about.rdf> correspond au document RDF associé comportant un certain nombre de triplets à son sujet ;
- <http://www.geonames.org/3017382/public-of-france.html> correspond à sa description (X)HTML associée.

Cette distinction est particulièrement importante au moment de la définition d'assertions. Lorsqu'on va modéliser des informations au sujet du pays (par exemple sa population), on va utiliser l'URI identifiant la ressource (*e.g.* <http://sws.geonames.org/3017382>) en tant que sujet des différents triplets mais si l'on souhaite définir une assertion au niveau du document (*e.g.* son auteur) on utilisera l'URL d'un document la décrivant (*e.g.* <http://www.geonames.org/3017382/public-of-france.html>). Afin de faire le lien entre ces niveaux de représentation, une bonne pratique veut que chaque URI associée à une ressource soit déréférençable²⁰ et renvoie vers un ensemble d'informations à son sujet en (X)HTML ou RDF selon l'agent logiciel utilisé pour déréférencer cette URI. Pour plus de détails sur la définition d'URIs pour le Web Sémantique, on pourra consulter [Ayers et Völkel, 2008] et différentes discussions à ce sujet au sein du W3C²¹.

¹⁹Ceci introduirait de plus des problèmes de consistance. Par exemple FOAF définit les classes Agent et Document comme disjointes, ce qui implique qu'une même URI ne peut pas représenter à la fois une personne (ressource non-informationnelle) et sa page personnelle (ressource informationnelle).

²⁰"Agents may use a URI to access the referenced resource; this is called dereferencing the URI." [W3C Technical Architecture Group, 2004]

²¹<http://www.w3.org/2001/tag/issues.html#httpRange-14>

Enfin, signalons que rien n'oblige l'ensemble des triplets concernant une ressource donnée à être stocké au sein du même graphe ou document. Au contraire, puisque cette approche se base sur les principes du Web et donc d'une architecture distribuée il est tout à fait possible de définir ces informations dans plusieurs documents, l'identification des sources permettant par la suite de tracer l'origine de chaque assertion. Nous détaillerons cette pratique de gestion distribuée des connaissances dans le chapitre consacré à l'annotation sémantique et au peuplement d'ontologies (Section 4.2.2, page 154).

Vocabulaires et ontologies : RDFS et OWL

Si les URIs et RDF offrent un cadre commun pour représenter les informations dans le contexte du Web Sémantique, ils ne permettent pas de définir la sémantique des données manipulées. Il faut en effet fournir aux agents capables de lire ces informations un moyen d'interpréter que, par exemple, <http://xmlns.com/foaf/0.1/Organization> représente le concept d'organisation qui peut avoir différentes propriétés et être également lié à d'autres concepts également interprétables. Pour ce faire, il est nécessaire de disposer de vocabulaires ou d'*ontologies* pour modéliser cette sémantique.

Si le terme *ontologie* trouve son origine en philosophie où l'*ontologie* représente la "*spéculation sur l'être en tant qu'être*", l'informatique la définit comme étant "*la spécification explicite d'une conceptualisation*" [Gruber, 1995]. Nous verrons plus loin (Section 3, page 83) qu'il existe plusieurs types d'*ontologies*, partageant cependant en commun différents éléments :

- des *concepts*, ou classes, définissant des ensembles d'objets, abstraits ou concrets, que l'on souhaite modéliser pour un domaine donné. En fonction de celui-ci et des choix de modélisation (puisque la définition d'une *ontologie* implique un certain point de vue) le niveau d'abstraction peut-être très large (ex : la notion de temps) ou au contraire très précis (ex : un élément de robinetterie nucléaire). Il faut donc prendre cette notion de concept de manière très large fortement dépendant du domaine ;
- des *propriétés* attribuées à ces concepts parmi lesquelles on distingue généralement (1) les relations qui peuvent exister entre concepts ou instances de ces concepts et (2) les attributs primitifs qu'il est possible d'associer aux différents concepts ou à leurs instances (chaîne de caractère, entier ...). Un exemple particulier de relation fréquemment utilisée dans les *ontologies* est la relation de subsomption qui permet d'établir des hiérarchies de concepts ;
- des *axiomes*, qui permettent de modéliser des assertions logiques et qui sont utilisés dans la définition de concepts ou de propriétés afin d'affiner celles-ci. Associés à des raisonneurs, ils permettent d'établir de nouveaux faits à partir des connaissances de base ou de vérifier la consistance d'un ensemble d'assertions.

On peut associer à chaque concept différentes déclinaisons linguistiques (ou termes) et il est important de bien distinguer le terme du concept comme le rappelle [Bachimont, 2000]. [Kassel et Perrette, 1999] va également plus loin dans cette distinction en considérant pour chaque concept les termes associés, la notion (*i.e.* l'intention du concept) et l'objet (*i.e.* son extension). Nous reviendrons sur cette distinction en présentant nos propositions permettant de lier tags et *ontologies* de domaine (Section 3.3.3, page 128).

On a généralement coutume de distinguer l'*ontologie* (*i.e.* le modèle) des individus ou

instances (*i.e.* les réalisations des différents concepts présents dans le modèle) et de considérer que ceux-ci ne font pas partie de l'ontologie mais appartiennent à la base de connaissance associée, l'ontologie étant alors un modèle conceptuel venant en support de cette base de connaissances et des faits qu'elle contient [Guarino et Giaretta,]. Pour reprendre l'exemple précédent, les notions d'organisation et de zone géographique feront ainsi partie d'une ontologie donnée et EDF, la France et le fait qu'EDF soit une organisation basée en France seront eux des éléments de la base de connaissance associée. Cette distinction entre instances et base de connaissance est par ailleurs similaire à ce que proposent les logiques de description [Baader *et al.*, 2003] en distinguant les ABox et TBox. Pour plus de détails sur ces principes de modélisation, ainsi que les réseaux sémantiques [Quillian, 1968], les graphes conceptuels [Sowa, 1984] et autres formalismes de représentation des connaissances ayant précédé les ontologies, on pourra se référer aux thèses [Troncy, 2004] et [Isaac, 2005] ou à l'ouvrage *Ingénierie des connaissances* [Charlet *et al.*, 2000].

Pour prendre un exemple concret, on peut imaginer une ontologie qui définit :

- des concepts : *Agent*, *Entreprise* et *Personne* ;
- des propriétés :
 - *isA*, relation de subsomption telle que *isA(Entreprise, Agent)* et *isA(Personne, Agent)* ;
 - *instanceOf*, relation d'instanciation telle que *instanceOf(AlexandrePassant, Personne)*²² ;
 - *aPourEmploye*, relation telle que *aPourEmploye(Entreprise, Personne)* ;
 - *aPourNom*, attribut assigné aux concepts *Agent*, *Entreprise* et *Personne*²³
 - *aPourNSS*, attribut assigné au concept *Personne* ;
- des axiomes :
 - $x, aPourNSS(x) = 1$ indiquant que les réalisations des concepts *Agent* n'ont qu'un seul numéro de sécurité sociale ;
 - $x, aPourNom(x) >= 1$ indiquant que les réalisations des concepts *Agent*, *Entreprise* et *Personne* ont au moins un nom ;
 - $(x, y), aPourEmployé(x, y) >= 1$ indiquant que toute réalisation du concept *Entreprise* a au moins un employé (défini en tant que *Personne*) ;

et y associer deux individus au sein de la base de connaissances qui suit (Listing 1.4, page 22).

```

instanceOf(AlexandrePassant, Personne)
instanceOf(EDF, Entreprise)
aPourNom(AlexandrePassant) = "Alexandre Passant"
aPourNSS(AlexandrePassant) = "1800669XXXXXXX"
aPourNom(EDF) = "Électricité de France"
aPourEmployé(EDF, AlexandrePassant)

```

Listing 1.4: Exemple de base de connaissances associée à une ontologie

²²On peut en fait considérer que les relations *isA* et *instanceOf* ne font pas partie de l'ontologie elle-même, mais d'un métamodèle permettant la définition d'ontologies, comme nous allons le voir avec RDFS et OWL.

²³Notons ici qu'en fonction des langages utilisés pour définir l'ontologie, il peut suffire de définir cet attribut comme propriété de *Agent* pour que les concepts *Entreprise* et *Personne* en héritent en raison de règles d'inférence associées à l'utilisation de la relation de subsomption *isA*.

RDFS – RDF Schema [Brickley et Guha, 2004] – est une première étape pour modéliser des ontologies sur le Web Sémantique. Ce langage introduit les notions de classe (`rdfs:Class`) et de propriété (`rdfs:Property`) associées à des relations de subsomption permettant de définir des hiérarchies de classes et de propriétés, respectivement `rdfs:subClassOf` et `rdfs:subPropertyOf`. RDFS permet également pour chaque propriété de définir son domaine (`rdfs:domain`) et son codomaine (`rdfs:range`), soit respectivement `Entreprise` et `Personne` pour la relation `aPourEmployé` de l'exemple précédent.

Une ontologie RDFS s'écrit sous forme de triplets RDF qui vont ainsi définir des identifiants pour ses différentes classes et propriétés, ceux-ci étant uniques puisque basés sur des URIs. Le code qui suit (Listing 1.5, page 23) représente une ontologie modélisant une partie des classes et propriétés que nous avons présentées dans l'exemple précédent. Il introduit également la possibilité dans une ontologie d'utiliser et d'étendre des classes et propriétés définies dans d'autres modèles. Dans notre cas, la classe `Entreprise` étend la classe `Organisation` définie dans l'ontologie FOAF [Brickley et Miller, 2004a].

```

@prefix : <#> .
@prefix foaf: <http://xmlns.com/foaf/0.1> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

:Entreprise a rdfs:Class ;
  rdfs:subClassOf foaf:Organization ;
  rdfs:label "Entreprise" .
:aPourEmploye a rdfs:Property ;
  rdfs:domain :Entreprise
  rdfs:range foaf:Agent

```

Listing 1.5: Exemple d'ontologie représentée en RDFS et sérialisée en Turtle

Pour finir, RDFS laisse entrevoir via la sémantique associée à RDF/RDFS [Hayes, 2004] des premières règles d'inférence simples permettant à une base de connaissance de s'enrichir de nouvelles assertions à partir du moment où certains faits sont présents dans celle-ci. Ces règles incluent notamment la transitivité des propriétés `subClassOf` (règle `rdfs7`) et `subPropertyOf` (règle `rdfs9`) comme le montre le tableau qui suit (Tableau 1.1, page 23).

Règle	Si	Alors
<code>rdfs7</code>	aaa rdfs :subPropertyOf bbb . uuu aaa yyy .	uuu bbb yyy.
<code>rdfs9</code>	uuu rdfs :subClassOf xxx . vvv rdf :type uuu .	vvv rdf :type xxx.

Tableau 1.1: Règles d'inférence RDFS

L'expressivité de RDFS est malgré tous assez restreinte. Ce langage ne permet pas entre autres de définir la notion de symétrie d'une propriété qui pourrait s'appliquer à une rela-

tion *aPourVoisin*²⁴. Ainsi, pour aller plus loin dans la définition d'ontologies pour le Web Sémantique, le W3C a mis en place dès 2001 un groupe de travail autour d'OWL – *Web Ontology Language* [Bechhofer et al., 2004] –, langage de définition d'ontologies sur le Web dans la continuité de DAML+OIL [Horrocks, 2002], issu lui-même des projets et langages OIL [Fensel et al., 2000] en Europe et DAML-Ont [McGuinness et al., 2003] aux Etats-Unis. OWL, passé au statut de recommandation du W3C en 2004, reprend ainsi les notions de classes et de propriétés définies en RDFS en les précisant respectivement par `owl:Class` (sous-classe de `rdfs:Class`) et `owl:dataTypeProperty` et `owl:objectProperty` (sous-classe de `rdfs:Property`) distinguant ainsi les attributs (types primitifs) des relations (liens vers d'autres classes). Surtout, OWL ajoute de nouveaux constructeurs et axiomes permettant d'accroître l'expressivité des ontologies, avec une sémantique plus poussée que celle de RDFS [Patel-Schneider et al., 2004]. OWL se compose en réalité de trois sous langages, à l'expressivité croissante²⁵ :

- OWL-Lite qui étend RDFS et ajoute de nouveaux constructeurs comme la symétrie des propriétés et des contraintes de cardinalité (uniquement 0 ou 1) ;
- OWL-DL dont le nom est hérité des logiques de description et qui ajoute des constructeurs supplémentaires (et regroupe en fait l'ensemble des constructeurs disponibles en OWL), comme les notions de combinaisons booléennes de classes (union ou intersection), des axiomes de classes (disjonction) et étend les contraintes de cardinalité d'OWL-Lite ;
- OWL-Full qui n'ajoute pas de constructeur par rapport à OWL-DL mais qui les interprète différemment offrant ainsi une expressivité plus forte (toute classe est vue à la fois comme une classe, un individu et un ensemble d'individus) mais sans garantie de calculabilité, OWL-Full n'étant pas décidable.

Les différents axiomes définis dans une ontologie OWL peuvent être pris en compte dans un processus de raisonnement avec des systèmes comme Pellet²⁶ [Sirin et al., 2007] ou Racer²⁷ [Haarslev et Möller, 2001]. Ceux-ci peuvent être utilisés par exemple (1) pour la classification automatique d'instances en fonction de leurs propriétés et des axiomes définis dans l'ontologie ou (2) pour la création de nouvelles relations entre instances en fonction de l'état initial d'une base de connaissance. Par exemple, un axiome définissant la symétrie d'une propriété *aPourVoisin* conduira à la règle suivante :

$$(x, y), \text{aPourVoisin}(x, y) \quad \text{aPourVoisin}(y, x) \quad (1.1)$$

En outre, il est important de garder à l'esprit que ces langages (RDFS et OWL) se situent dans l'hypothèse d'un monde ouvert et donc que l'absence de déclaration d'un fait ne permet pas de considérer celui-ci comme faux. Ainsi, si dans un ensemble d'assertions aucune d'entre elles n'indique qu'EDF est situé en France, un système basé sur ces langages

²⁴Nous considérons ici la notion de voisinage au sens large, i.e. ne distinguons pas *aPourVoisin* et *aPourVoisine*.

²⁵Alors que OWL 2 est en cours de standardisation, notons que tout au long de ce mémoire, nous étudierons uniquement sa version 1 et utiliserons l'appellation OWL (et non pas OWL 1) par simplicité.

²⁶<http://pellet.owldl.com/>

²⁷<http://www.racer-systems.com/>

ne déduira pas qu'EDF n'est pas une entreprise française, mais simplement qu'il n'est pas capable de répondre à cette question. De ce fait, l'absence de définition d'une propriété dans une ontologie n'interdit pas son utilisation et de même le fait de déclarer une ressource instance d'une classe n'interdit pas de l'utiliser comme instance d'une autre classe. Si l'on souhaite *a contrario* qu'une propriété ne puisse pas être utilisée pour une classe donnée ou qu'être instance d'une classe ne permette pas d'être instance d'une seconde, il est nécessaire de spécifier certains axiomes (cardinalité et disjonction dans notre cas) dans l'ontologie, le passage de RDFS à OWL étant alors requis. De plus, notons que ces langages sont descriptifs et non prescriptifs²⁸ [Davis, 2005]. Ainsi, le fait de définir `Entreprise` comme domaine de `aPourEmploye` n'implique pas que les instances associées à cette propriété soient explicitement typées `Entreprise` mais qu'elles le deviennent par inférence lorsque la propriété leur est assignée. Les possibilités de raisonnement présentées en amont peuvent alors se révéler utiles pour vérifier la consistance du modèle, dans cette vision du monde ouvert qui déroute parfois, notamment lorsqu'on est habitué à des représentations plus classiques de typage.

1.1.3 Interrogation de données avec SPARQL

Alors que RDFS et OWL permettent de définir des ontologies sur le Web Sémantique et RDF de modéliser des assertions en se basant sur celles-ci, il est nécessaire pour en tirer parti de disposer d'un langage de requête adapté. SPARQL – *SPARQL Protocol and RDF Query Language* [Prud'hommeaux et Seaborne, 2008] – propose ainsi à la fois un langage et un protocole pour interroger des données modélisées en RDF. Ces travaux s'inscrivent dans la continuité de RDQL [Seaborne, 2004] et l'on peut voir SPARQL comme le SQL du Web Sémantique : "Tenter d'utiliser le Web sémantique sans SPARQL revient à exploiter une base de données relationnelle sans SQL"²⁹. SPARQL utilise le principe d'identification de chemins dans un graphe [West, 2000] pour récupérer les résultats d'une requête donnée. Ainsi, une requête SPARQL se compose d'un opérateur (définissant le type de requête), d'un patron (la partie nécessaire pour l'identification des graphes correspondants) et de modificateurs (par exemple, `ORDER BY`). Une requête peut interroger un ou plusieurs documents RDF, soit par l'utilisation d'un attribut `FROM` en début de requête, soit par l'intermédiaire d'APIs – *Application Programming Interface* – qui permettent de considérer simultanément plusieurs sources, soit via l'utilisation d'entrepôts de données RDF associés à des points d'accès (ou *endpoints*) SPARQL (Section 5.1, page 186). SPARQL dispose des quatre opérateurs suivants³⁰ :

- `SELECT` qui comme son nom l'indique va sélectionner différents éléments selon un patron de requête particulier. Une requête destinée à récupérer la localisation d'EDF pourrait être :

```
SELECT ?pays
WHERE { athena:EDF geonames:locatedIn ?pays }
```

Listing 1.6: Exemple de requête SPARQL SELECT

²⁸<http://lists.w3.org/Archives/Public/public-xg-geo/2007Jan/0002.html>

²⁹<http://www.w3.org/2007/12/sparql-pressrelease>

³⁰Nous avons ici volontairement supprimés les définitions de préfixes pour des raisons de lisibilité.

- **CONSTRUCT** qui permet de transformer un graphe RDF en un autre graphe. On peut ainsi voir cet opérateur comme le XSLT du Web Sémantique (Section 4.1.2, page 140). Par exemple, pour passer de notre modèle à un autre vocabulaire, on peut utiliser :

```
CONSTRUCT { ?entreprise mon_ontologie:situeDans ?pays }
WHERE { ?entreprise geonames:locatedIn ?pays }
```

Listing 1.7: Exemple de requête SPARQL CONSTRUCT

- **ASK** qui permet de répondre à une requête, en identifiant si oui ou non le patron recherché est présent dans le graphe interrogé. Ainsi, "*EDF est-il situé en France ?*" peut s'écrire :

```
ASK { athena:EDF geonames:locatedIn geonames:3017382/ }
```

Listing 1.8: Exemple de requête SPARQL ASK

- **DESCRIBE** qui renvoie sous forme d'un graphe RDF l'ensemble des triplets ayant pour sujet la ressource passée en argument. Par exemple, pour connaître l'ensemble des assertions relatives à EDF, on écrira :

```
DESCRIBE athena:EDF
```

Listing 1.9: Exemple de requête SPARQL DESCRIBE

SPARQL offre également la possibilité d'utiliser les graphes nommés via le patron GRAPH, par exemple pour restreindre les graphes où l'identification de patrons doit être appliquée. Nous reviendrons en détail sur cette utilisation dans la partie consacrée aux wikis sémantiques (Section 4.2.1, page 148) et plus particulièrement sur l'utilisation que nous en faisons au sein de l'outil que nous avons mis en place (Section 4.2.2, page 154).

SPARQL souffre cependant de différentes limites, notamment par rapport à un langage comme SQL. Par exemple, il ne propose pas pour le moment de fonctions d'agrégat, ni de possibilité d'ajouter des données dans un graphe, SPARQL étant uniquement dédié à des requêtes en lecture seule. Diverses extensions veillent cependant à résoudre ces limites et ajouter de nouvelles fonctionnalités. Citons pas exemple la recherche par chemins et plus uniquement par triplets (SPARQLer [Kochut et Janik, 2007]) ou l'approximation de requêtes (iSPSPARL [Kiefer et al., 2007]), des fonctionnalités proches étant implémentées dans le moteur SPARQL Corese [Corby et al., 2004]. Pour en revenir aux fonctions d'agrégat, si elles ne sont pas définies par la sémantique de SPARQL, elles sont malgré tout implémentées dans des moteurs comme ARC³¹ ou Virtuoso³². Nous détaillerons plus tard les efforts concernant l'ajout et les modifications de données RDF avec SPARUL – SPARQL Update

³¹<http://arc.semsol.org>

³²<http://virtuoso.openlinksw.com/>

[Seaborne *et al.*, 2008] – (Section 5.1.3, page 192). Notons également que certaines de ces propositions sont à l'ordre du jour du nouveau groupe de travail au W3C autour de SPARQL³³ dont nous sommes aujourd'hui membre.

Enfin, [Pérez *et al.*, 2006] ont montré que certains types de requêtes faisaient partie de la catégorie des problèmes *NP complets* [Garey et Johnson, 1979] étant donné le principe de parcours de graphes qu'utilise SPARQL. Cependant, il est intéressant de constater que les requêtes peuvent, de façon plus générale, être optimisées en fonction de l'ordre des patrons de requêtes, de manière à réduire successivement le graphe où la requête s'applique [Stocker *et al.*, 2008]. Nous pouvons imaginer qu'à l'avenir, ces stratégies d'optimisations seront implémentées dans la plupart des moteurs SPARQL, à la manière de ce qui se fait pour la réécriture automatique de requêtes dans les systèmes SQL [Kraft *et al.*, 2003].

1.1.4 Web Sémantique et *Web of Data*

Malgré les efforts de standardisation de ces différents langages qui posent les bases de la représentation et de l'interrogation de données sur le Web Sémantique, il faut reconnaître que jusqu'à récemment, les données RDF disponibles sur le Web étaient peu nombreuses. Si FOAF, notamment au travers d'exports natifs depuis certains sites comme LiveJournal³⁴, a permis d'entrevoir une démocratisation de ces données, le domaine est longtemps resté limité. En contrepartie, de nombreuses données libre d'accès (utilisant par exemple des licences Creative Commons³⁵) sont aujourd'hui disponibles sur le Web. C'est devant ce double constat qu'est née l'initiative *Linking Open Data*, supportée par le groupe *Semantic Web Education and Outreach* du W3C³⁶, avec l'objectif d'exposer en RDF un grand nombre de données déjà présentes sur le Web (mais dans des formats hétérogènes ou sous forme de simples documents HTML) et d'interconnecter celles-ci.

Pour parvenir à cette vision plus pragmatique du Web Sémantique (au sens où ce sont les données et les bases de connaissances qui sont mises en avant, et non pas les ontologies et les possibilités qu'elles offrent), le projet repose sur quatre principes de base définis par [Berners-Lee, 2006a] :

- utiliser des URIs pour nommer les choses ;
- utiliser des URIs HTTP afin que l'on puisse déréférencer ces choses ;
- lorsque quelqu'un déréfère une URI, lui fournir des informations utiles à son sujet ;
- inclure des liens vers d'autres URIs, afin que l'on puisse découvrir plus d'informations ;

L'initiative, débutée en Juin 2007, a permis de produire un nombre impressionnant de données liées (Figure 1.5, page 28), estimées aujourd'hui à plusieurs milliards d'assertions et issues de différentes sources de données aussi diverses que DBpedia³⁷ (export RDF de Wikipedia) [Auer *et al.*, 2007], les programmes de la BBC [Scott *et al.*, 2008] ou encore les profils utilisateurs de Flickr³⁸ [Passant, 2008b]. Différentes stratégies sont utilisées pour produire

³³<http://www.w3.org/2009/01/sparql-charter.html>

³⁴<http://livejournal.com>

³⁵<http://creativecommons.org>

³⁶<http://www.w3.org/2001/sw/swoe/>

³⁷<http://dbpedia.org>

³⁸<http://apassant.net/blog/2007/12/18/rdf-export-flickr-profiles-foaf-and-sioc>

ces liens entre données, de la contribution manuelle utilisateur [Hausenblas *et al.*, 2008] à l'utilisation d'heuristiques plus poussées [Raimond *et al.*, 2008], notamment pour gérer les problèmes d'ambiguïté qui se posent.

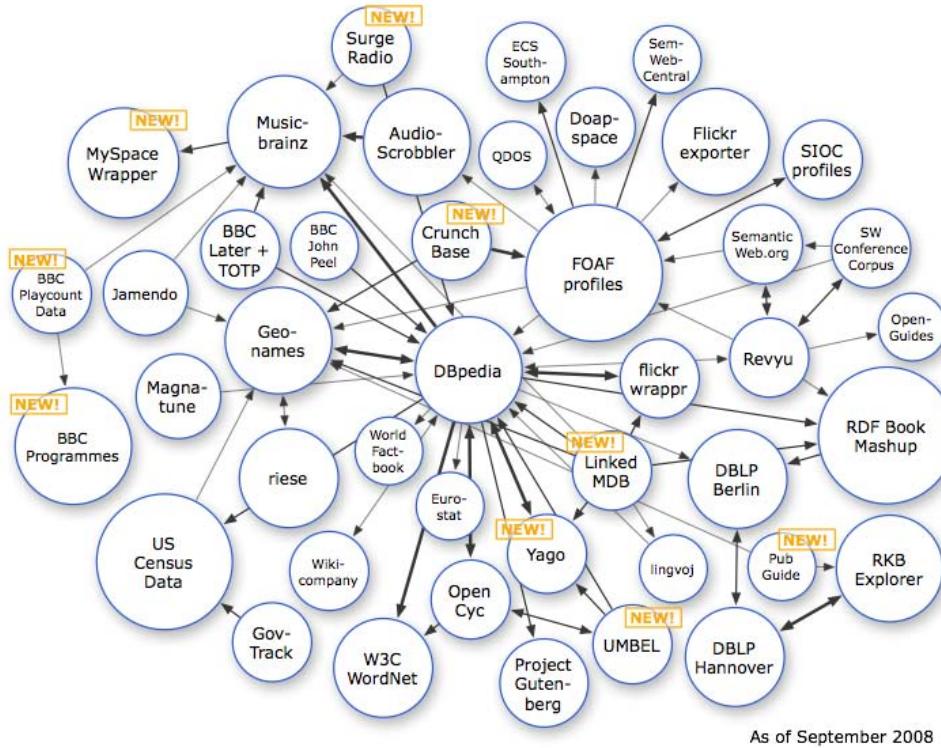


Figure 1.5: Nuage de données du projet Linking Open Data³⁹

Ces données étant issues de sources existantes, pour la plupart des documents (X)HTML, nous pouvons nous poser la question de la relation qui existe entre (1) le document Web tel que nous le connaissons aujourd'hui, avec sa notion hyperliens, et (2) les données associées et les relations qu'elles partagent. Selon nous, le document peut être vu comme un simple support à ces données via les annotations sémantiques associées, que celles-ci soient modélisées dans un document RDF annexe associé à la page (X)HTML ou incluses directement avec RDFA. Par ailleurs, le document peut être un support à la fois pour la production (annotations extraites depuis celui-ci) et la visualisation de données (interface de navigation dans un graphe de connaissances). Nous aborderons ces deux points de vue dans la suite de ce mémoire, tout d'abord concernant la production (Section 4, page 137), puis la visualisation (Section 5, page 185) de données RDF. Comme le montre la figure qui suit, cette correspondance se fait de plus assez naturellement en termes de représentation (Figure 1.6, page 29). Une prochaine étape est selon nous l'exploitation de toutes ces données et plus uniquement des documents comme le font la majorité des moteurs de recherche tradition-

³⁹<http://richard.cyganiak.de/2007/10/1od/>

nels ou les navigateurs Web. Nous reviendrons sur cette exploitation de données RDF dans le dernier chapitre de cette thèse (Section 5, page 185).

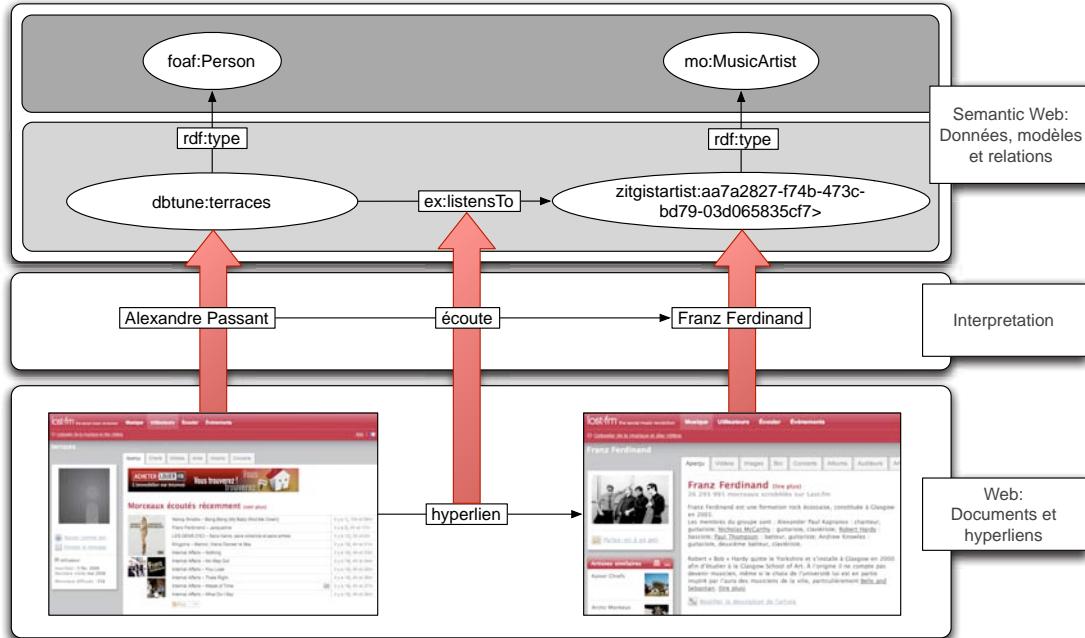


Figure 1.6: Le document en tant que support de données pour le Web Sémantique

On peut cependant reprocher certaines limites à cette initiative *Linking Open Data*, notamment parfois un manque de formalisme dans les représentations extraites. Par exemple, les premières versions de DBpedia ne reposaient sur aucune ontologie, la version 3.2 résolvant ce problème⁴⁰ mais avec une ontologie qui ne suis pas toujours ce que nous considérons être des bonnes pratiques de modélisation (Section 3.2.4, page 109). Également, l'utilisation abondante de certaines propriétés à la sémantique forte, comme `owl:sameAs` qui introduit la notion d'identité sur le Web Sémantique, se fait parfois au détriment de la qualité des annotations produites et inférées. Cette propriété est en effet fréquemment utilisée comme une simple relation entre deux ressources, sans prendre en compte les probables problèmes de consistance qui peuvent arriver lors de la fusion de celles-ci, dues à la sémantique même de `owl:sameAs`.

Malgré tout, nous ne pouvons que nous réjouir de la vivacité du projet *Linking Open Data* et de sa communauté (notamment via l'organisation de *workshops* dédiés [Bizer et al., 2008]) et en conséquence de cet amas de données RDF aujourd'hui disponibles en ligne. Celles-ci ont permis de faire un grand pas en avant dans l'acceptation du Web Sémantique en tant que graphe global de connaissances, notamment au niveau du grand public⁴¹ et dans la sphère

⁴⁰<http://lists.w3.org/Archives/Public/public-lod/2008Nov/0025.html>

⁴¹Même si cela reste pour le moment destiné à une audience technophile.

entrepreneuriale avec des entreprises comme Zemanta⁴² ou Freebase⁴³ qui rejoignent ces efforts en termes de production et d'interconnexion de données. Nous pensons également que celles-ci vont permettre d'aborder de nouveaux domaines de recherche en grandeur nature, comme la notion de confiance des sources de données ou l'inférence à large échelle.

1.2 DU CONSOMMATEUR AU PRODUCTEUR AVEC LE WEB 2.0

1.2.1 Une vision participative du Web

En contrepartie de cette évolution du Web vers un modèle où les données formalisées permettent de faciliter les échanges d'information, ces dernières années ont vu apparaître une autre vision du Web, plus sociale – et économique – que technique, communément appelée Web 2.0. [O'Reilly, 2005] définit le Web 2.0 comme "*a set of principles and practices that ties together a veritable solar system of sites that demonstrate some or all of those principles, at a varying distance from that core*". Si cette définition reste assez floue on trouve cependant parmi cet ensemble de principes fondamentaux (Figure 1.7, page 31) deux notions qui nous paraissent particulièrement importantes, à savoir celles de *Web en tant que plate-forme* et celle d'*architecture participative (architecture of participation)*.

Cette première notion reconsideère l'utilisation du Web et de ses principes pour y fournir des services et applications à forte valeur ajoutée plutôt que des contenus essentiellement statiques. Le rôle du Web peut même être dédié à celui de simple plate-forme d'échange et de transit de l'information comme dans le cas de RSS (Section 1.2.2, page 36). Par extension, on regroupe également sous ce terme la migration de services traditionnels (client mail, suite bureautique ...) vers des applications en ligne.

Dans ce contexte, la notion d'architecture participative met en avant la production de contenus à forte valeur ajoutée par effet de bord des usages réguliers et des intérêts personnels que chacun poursuit en utilisant ces applications. Ceci se fait par ailleurs de manière autonome en raison de la manière même dont ces applications ont été conçues. Nous verrons par exemple en détaillant l'utilisation des wikis (Section 1.2.2, page 35) de quelle manière des modifications établies individuellement permettent d'enrichir globalement un document ou un site de manière collaborative mais surtout continue et transparente. [O'Reilly, 2005] fait ainsi l'analogie avec Dan Bricklin qui présente comment les processus de développement *open-source* (le développement de fonctionnalités par un utilisateur pour un besoin précis impliquant une évolution générale de l'application dont tous peuvent bénéficier) et les architectures *peer-to-peer* (chaque consommateur devenant à son tour fournisseur de données) parviennent à ce même objectif⁴⁴. On peut également comparer ces principes à l'architecture même du Web, l'ajout d'hyperliens entre documents permettant d'accroître la structure du graphe global qu'il représente, renforçant ainsi les possibilités générales de navigation.

Plus généralement, on peut considérer le Web 2.0 comme une vision du Web mettant à disposition des utilisateurs un ensemble de services et de technologies visant à faciliter la production et le partage d'informations de manière intuitive et collaborative. Ainsi, le

⁴²<http://zemanta.com>

⁴³<http://freebase.com>

⁴⁴<http://www.bricklin.com/cornucopia.htm>

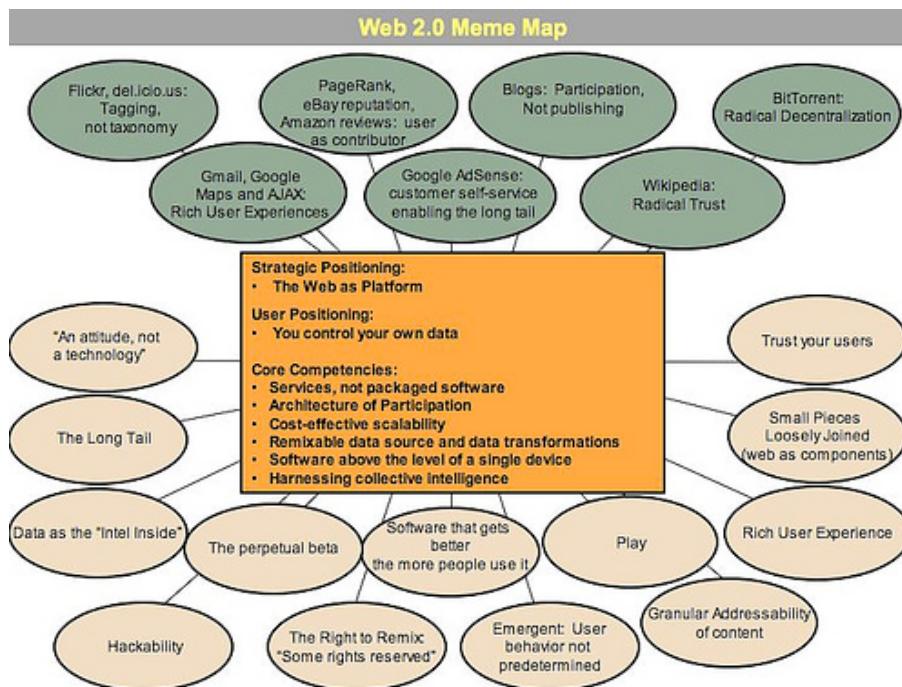


Figure 1.7: L'écosystème Web 2.0 [O'Reilly, 2005]

Web devient un média participatif *many-to-many*, plus qu'un simple espace de stockage à titre essentiellement consultatif, l'utilisateur final ayant de ce fait un rôle central dans cette démarche⁴⁵. Pour y parvenir, les services Web 2.0 partagent pour la plupart un ensemble de principes communs :

- l'utilisateur est au centre du service, en termes de publication et de réaction. On peut même aller jusqu'à dire qu'il fait l'outil, la valeur de ce dernier dépendant de son contenu. Nous nous situons ici dans un schéma inverse de celui des portails Web de la fin des années 90 abondés par une autorité ou une équipe de rédaction établie à priori. On peut ainsi considérer que de nombreux services Web 2.0 sont des conteneurs vierges de tout contenu, ceux-ci étant soumis à l'adoption de l'outil par les utilisateurs ;
- le passage du statut de consommateur à celui de producteur doit se faire simplement. Le lecteur doit être en mesure de réagir à l'information qu'il consulte, *a minima* à un niveau inférieur à celui du producteur originel de l'information consultée (commentaires sur les blogs), au mieux au même niveau que celui-ci (édition de contenu sur un wiki, services de partage de contenu, etc.). Pour accentuer cette simplicité, les interfaces se doivent également d'être intuitives et sans prérequis technique ;
- la composante sociale se doit d'être présente non seulement en termes de publication mais aussi en termes d'échanges entre membres de la plate-forme. De tels services

⁴⁵C'est ainsi que le Time a consacré les internautes *personnalité de l'année 2006*. – <http://www.time.com/time/magazine/article/0,9171,1569514,00.html>

doivent être en mesure de stimuler les synergies entre internautes, voire de participer à l'élaboration de réseaux sociaux, virtuels ou réels, certains outils y étant entièrement dédiés.

Malgré ces caractéristiques communes, les services offerts sont relativement divers. Ainsi, les blogs (Section 1.2.2, page 33) mettent en avant l'individu, en offrant un système de publication personnelle en ligne. Les wikis (Section 1.2.2, page 35) ont quant à eux pour objectif de participer à l'élaboration collective et consensuelle de contenu. De nombreux services de partage de données ont également fait leur apparition, favorisant généralement la définition de réseaux sociaux (Section 1.2.3, page 41). Afin d'offrir un cadre légal à ces différentes données publiées et partagées, l'initiative Creative Commons permet aux utilisateurs de choisir, via différentes licences, de quelle manière ils autorisent la réutilisation des données ainsi mises à disposition sur le Web. Un autre aspect important dans l'utilisation de ces outils est la notion de *mash-up*, ou application composite, permettant de combiner les données provenant de divers services ou de les visualiser avec de nouvelles interfaces. De nombreux services Web 2.0 proposent en effet aux développeurs des APIs permettant de réutiliser les données produites en leur sein. Il est ainsi possible de combiner les données provenant des différentes applications, mais également de les combiner avec d'autres interfaces de visualisation, par exemple dans un but de géolocalisation comme le propose l'API Google Maps⁴⁶. Malheureusement, ces APIs sont le plus souvent propriétaires au sens où chaque service dispose d'une API distincte, contraignant les développeurs à apprendre les spécificités de chacune. Il en est de même pour les formats de réponse de ces APIs, dépendants des services interrogés. Pour une liste et une actualité plus complète des pratiques et services qui fleurissent chaque jour sur le Web, on pourra consulter un site comme Techcrunch⁴⁷.

Indépendamment de cette diversité, il est intéressant de constater que cette génération d'outils a introduit de nouvelles pratiques sociales en termes de partage d'information. Là où l'internaute avait auparavant tendance à restreindre la diffusion d'information à un public prédéfini, ces outils introduisent une notion de publication ouverte, où la cible n'est pas contrôlée par l'utilisateur. Si l'on peut comprendre une certaine réticence à partager sa vie privée ou son savoir de cette manière, notamment dans un contexte d'entreprise (Section 2.1.4, page 59), nombreux sont les utilisateurs publient volontairement de cette manière. On parle ainsi de *social software* ou de *social media* pour évoquer certains de ces outils et les paradigmes sociaux associés.

Bien que l'utilisation de l'indice 2.0 laisse entrevoir, à la manière des versions logicielles, une évolution du Web par rapport à sa vision originelle (Section 1.1.1, page 12), il s'agit principalement d'évolutions sociologiques et économiques comme le souligne l'ouvrage *Wikinomics*⁴⁸ [Tapscott et Williams, 2007]. Malgré tout, en raison de sa forte interaction avec les utilisateurs, cette évolution a introduit de nouvelles pratiques en matière de développement logiciel, notamment un certain nombre de *design patterns* spécifiques au Web 2.0 [Nickull *et al.*, 2008]. Parmi ceux-ci, [O'Reilly, 2005] incite les concepteurs de services à dépasser les processus traditionnels de développement et de livraison de nouvelles versions

⁴⁶<http://code.google.com/intl/fr-FR/apis/maps/>

⁴⁷<http://techcrunch.com>

⁴⁸<http://www.wikinomics.com/book/>

par paliers pour proposer aux utilisateurs de tester en flux continu leurs nouvelles idées, et bénéficier d'un retour sur expérience immédiat, avec cette notion de *bête perpétuelle*. Pour compléter ce point, le lecteur pourra se référer aux études sociologiques de Danah Boyd sur la manière dont les adolescents s'approprient ou font évoluer des services à forte audience comme MySpace⁴⁹ par leur pouvoir d'acceptation ou de refus de nouvelles fonctionnalités [Boyd, 2008]. D'un point de vue plus technique, on peut faire l'analogie entre ces pratiques et les principes du développement agile [Cohen *et al.*, 2004], mis en avant par l'avènement de frameworks logiciels comme Ruby On Rails⁵⁰. Ceux-ci mettent aussi l'accent sur des interactions fréquentes entre clients et maître d'ouvrage à la manière de ce que peut proposer l'*extreme programming* [Beck, 1999].

1.2.2 Blogs, wikis, réseaux sociaux et syndication de contenu

Blogs et publication personnelle d'information

Un blog, diminutif de weblog, est un site présentant sur sa page d'accueil un ensemble de billets (*posts* dans le vocabulaire anglophone) consistant en des notes ou articles plus ou moins longs et ordonnés de manière antéchronologique, l'usage d'hyperliens (internes et externes) y étant abondant. Un blog est en général personnel et donc maintenu par un unique auteur – ou blogueur –, mais peut aussi être partagé entre plusieurs rédacteurs, chacun ayant alors pour habitude de signer distinctement ses billets. En effet, le blog, contrairement au wiki que nous évoquerons dans la section suivante, met fortement l'accent sur la notion d'identité de l'auteur en tant que producteur de contenu. La notion de collaboration n'est alors pas liée à la rédaction de billets, mais à la possibilité que les lecteurs ont de réagir aux propos consultés par l'intermédiaire de commentaires associés aux billets. Cet aspect participatif permet ainsi à chacun de former et de fidéliser une communauté de lecteurs évolutive et réactive autour de soi et de ses écrits ou opinions, notion de réseau social que nous détaillerons par la suite (Section 1.2.3, page 41).

À nouveau, ce n'est pas l'aspect technologique des blogs qui fait leur force, mais leur simplicité de mise en œuvre et d'utilisation couplée à la composante collaborative évoquée ci-dessus. De nombreux services proposent la création d'un blog en quelques minutes (Blogger⁵¹, Wordpress.com⁵² ...) et les outils pour installer son propre système sont également nombreux. La publication se fait sans connaissance technique via une interface Web ou dans certains cas directement depuis son poste de travail ou un terminal mobile, contribuant à l'ubiquité de la présence en ligne d'un individu. Ainsi, les blogs ont remis au goût du jour le concept de page personnelle, la nature spontanée et régulière des billets et leur présentation antéchronologique offrant cependant une dynamique tout autre.

La nature des blogs aujourd'hui disponibles sur le Web est assez diverse, puisqu'on y trouve aussi bien des journaux intimes d'adolescents, des blogs d'exports, que des blogs d'opinion. Certains d'entre eux, notamment les blogs d'opinion ou les blogs politiques, qui mettent en avant le concept de journalisme citoyen, ou *grassroots journalism*, peuvent

⁴⁹<http://myspace.com>

⁵⁰<http://rubyonrails.com>

⁵¹<http://blogger.com>

⁵²<http://wordpress.com>

même concurrencer en termes d'audience les grands quotidiens comme le montrent notamment des études de Technorati sur le sujet⁵³. Il est également intéressant de constater, toujours dans cette perspective de rapport à l'actualité, le parallèle en termes de publication et de temporalité de l'information qui existe entre blogs et médias traditionnels [Cointet *et al.*, 2007]. Pour une étude sociologique plus complète sur ce phénomène de journalisme citoyen on pourra consulter l'ouvrage *We the Media*⁵⁴ [Gillmor, 2004]. On peut enfin également noter que si les contenus sont variables, tout comme les fréquences de mise à jour, le nombre de blogs est en constante augmentation. Ainsi, Technorati⁵⁵, service référençant les blogs sur le Web et proposant un moteur de recherche associé, en recensait plus de 70 millions début 2007 (Figure 1.8, page 34) et plus de 130 millions en 2008⁵⁶.

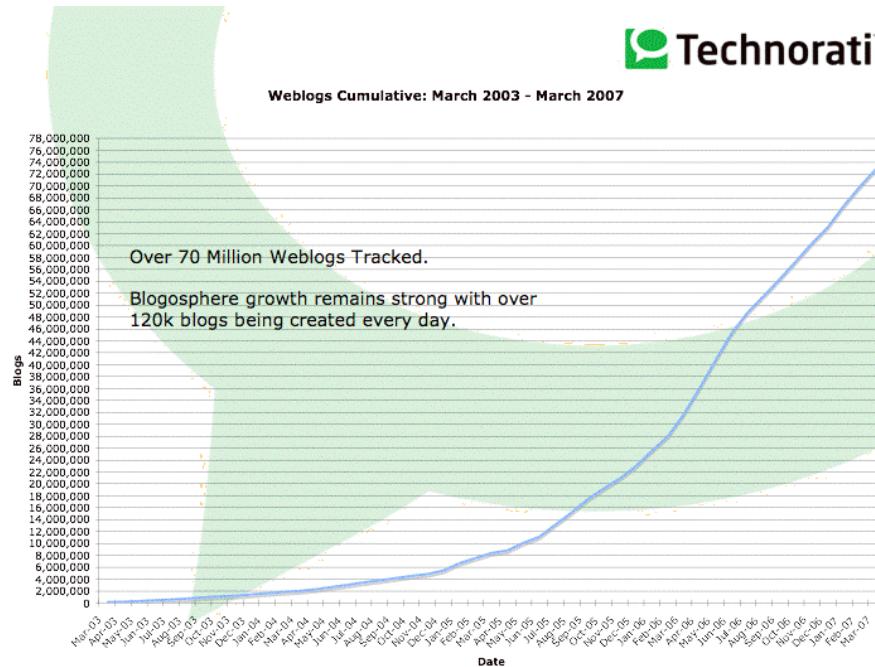


Figure 1.8: Etat de la blogosphère, Avril 2007⁵⁷

Une des forces des blogs, comme nous l'avons évoquée, est la possibilité d'expression spontanée qu'ils offrent et en conséquence les discussions qu'ils engendrent. À cet égard, il nous semble important de signaler l'explosion récente du phénomène de microblogging, popularisé par Twitter⁵⁸. À mi-chemin entre le blog et la messagerie instantanée, ce mode de communication se traduit par la publication de courts messages (généralement moins de

⁵³<http://technorati.com/weblog/2006/02/83.html>

⁵⁴<http://wethemedia.oreilly.com/>

⁵⁵<http://technorati.com>

⁵⁶<http://technorati.com/blogging/state-of-the-blogosphere/>

⁵⁷<http://www.sifry.com/alerts/archives/000493.html>

⁵⁸<http://twitter.com>

140 caractères) non-titrés et sans restriction de contenu. Si ces messages sont généralement proches de la notification de statut personnel, ils peuvent aussi servir au signalement léger d'informations (en postant par exemple un simple lien vers une ressource en ligne jugée intéressante) et permettent de manière plus générale une communication agile entre les personnes les postant et ceux y répondant ou simplement les suivant [Java *et al.*, 2007]. Puisque nous évoquions auparavant la notion de journalisme citoyen, notons également le rôle joué par Twitter à ce sujet, du fait des différentes possibilités qu'il offre pour la publication de message (via Web, e-mail, SMS, etc.) ainsi que vis-à-vis des modes de réactions associés et de la propagation de ces messages⁵⁹.

Wikis et consensus informationnel

Un wiki [Leuf et Cunningham, 2001] est un site Web dynamique et évolutif, au sens où il permet à chaque lecteur de modifier les pages consultées et d'en ajouter de nouvelles mais aussi d'en supprimer. Ainsi, la dynamique d'un wiki s'observe non seulement vis-à-vis du contenu de ses pages mais aussi via l'architecture générale de celui-ci, évoluant selon les actions utilisateurs. Un wiki n'est généralement pas axé sur des informations contextualisées temporellement et produites par un auteur unique identifié (cas du blog), mais sur la construction collaborative et incrémentale de contenu consensuel. Les usages des wikis sont divers, de l'encyclopédie généraliste – l'exemple le plus parlant étant Wikipedia⁶⁰ – à la documentation de projets *open-source* (par exemple Trac⁶¹), l'ouverture du site s'inscrivant ici dans la continuité du libre accès du code. Même si ces outils ont été popularisés récemment, le premier prototype de wiki date de 1994⁶² le nom trouvant son origine dans le terme hawaïen *wiki wiki*, signifiant vite. Parmi les caractéristiques essentielles des wikis, nous retiendrons :

- des processus simples pour la participation. Par défaut, chaque lecteur doit être en mesure d'éditer le contenu d'un wiki quelque soit le niveau de modification souhaité (ajout, création ou suppression de pages) via le même outil que celui qui permet la visualisation du site. Pour ce faire, une syntaxe particulière est généralement utilisée et des processus de normalisation tels que Creole⁶³ ont été proposés à ce sujet, sans grand succès cependant ;
- en conséquence de cette édition ouverte, chaque page doit bénéficier d'un historique des modifications. Celui-ci permet de revenir simplement à une version précédente (en cas de modifications jugées non souhaitées pour la communauté, ou de vandalisme) ou simplement de consulter les modifications apportées entre deux versions. Certains wikis permettent également de s'abonner au flux des modifications d'une page (Section 1.2.2, page 36) ;
- le rôle important joué par les hyperliens. Un wiki doit permettre d'établir facilement des liens entre pages du même wiki. Pour ce faire, on utilise généralement la syntaxe

⁵⁹http://www.journalisme.sciences-po.fr/index.php?option=com_content&task=view&id=303&Itemid=112

⁶⁰<http://wikipedia.org>

⁶¹<http://trac.edgewall.org/>

⁶²<http://c2.com/cgi/wiki>

⁶³<http://www.wikicreole.org/wiki/Creole1.0>

MotWiki qui permet d'établir automatiquement un lien vers une page portant ce nom ou d'en créer une si celle-ci n'existe pas. Cette pratique renforce la dynamique des wikis et évite la présence de pages orphelines, *i.e.* sans lien entrant. La notion de rétrolien est également très présente, chaque page listant l'ensemble des pages ayant un lien entrant vers celle-ci. Cette pratique étend ainsi la notion de source et de direction des hyperliens pour offrir une navigation à double sens entre les pages.

Si le principe d'ouverture des wikis en fait dans l'idéal un outil adéquat pour la constitution collaborative de documents ou de sites, il soulève de nombreuses questions et introduit également des problèmes de *spam* ou de vandalisme. Ainsi, si certains systèmes introduisent des restrictions d'accès pour la modification des pages, d'autres s'organisent comme des espaces autogérés où les utilisateurs rectifient eux-mêmes les pages modifiées dans un sens n'allant pas avec celui défini, explicitement ou non, par la communauté. Nous reviendrons plus loin dans ce manuscrit sur des exemples d'utilisation des wikis dans un contexte d'entreprise et sur les problèmes rencontrés pour faire accepter l'outil dans un tel milieu (Section 2.2, page 62).

Syndication de contenu et personnalisation de l'accès à l'information avec RSS

Devant cette abondance de contenus en ligne et leur régulière évolution, il est nécessaire de fournir un moyen d'obtenir le signalement d'informations pertinentes selon les centres d'intérêt de chacun. La syndication de contenu a pour objectif de répondre à ce problème, en offrant aux sites un moyen de délivrer automatiquement un flux constamment actualisé de leurs dernières mises à jour, auquel les lecteurs peuvent s'abonner. Dans le but de formaliser ce processus et d'offrir un format standard de données, plusieurs modèles ont vu le jour, comme NewsML⁶⁴ dès 2000 pour les échanges entre fournisseurs d'informations et agrégateurs de données. Aujourd'hui, ces flux majoritairement modélisés en RSS ou Atom et généralement serialisés en XML sont disponibles sur la plupart des plates-formes de blogs et de wikis et sur une majorité d'applications Web 2.0. L'utilisateur peut souscrire à ces flux via un agrégateur, logiciel client ou service en ligne offrant une vision humainement lisible de ces informations brutes et tirant partie des différentes métadonnées contenues dans ces flux pour ordonner les éléments par date, source ou encore par auteur. Ces applications permettent également de récupérer à intervalles réguliers les dernières mises à jour des dits flux, certains flux RSS spécifiant directement leur fréquence de rafraîchissement via leurs métadonnées⁶⁵. RSS se décline en différentes versions, certaines provenant de groupements privés (0.9, proposée par Netscape), certaines fermées (2.0, figée) et d'autres provenant de groupes de travail ouverts (1.0⁶⁶, basée sur RDF en conséquence évolutive puisque permettant l'intégration de vocabulaires externes). C'est notamment devant cette confusion que s'est créé le groupe de travail autour d'Atom, aujourd'hui standard de l'IETF [Nottingham et Sayre, 2005] et également associé à un protocole de publication⁶⁷.

⁶⁴<http://www.newsml.org>

⁶⁵<http://web.resource.org/rss/1.0/modules/syndication/>

⁶⁶<http://web.resource.org/rss/1.0/>

⁶⁷Une version RDF d'Atom est également disponible avec Atom-OWL. – <http://bblfish.net/work/atom-owl/2006-06-06/AtomOwl.html>

Quelque soit sa version, un flux RSS se compose d'un conteneur (`channel`), contenant un certain nombre d'éléments (`item`) généralement limités aux 10 ou 20 dernières mises à jour de ce conteneur. Un exemple classique de flux RSS pour un blog va par exemple lister les 20 derniers billets postés sur celui-ci. Atom, propose quand à lui cette même organisation, mais via l'utilisation des éléments `feed` et `entry`. À chaque élément est associé un certain nombre de métadonnées comme la date, l'auteur ou encore l'URL de l'élément, certaines obligatoires, d'autres optionnelles, ce point variant selon les formats. Le code qui suit présente un flux RSS 2.0 pour un blog comportant deux entrées (Listing 1.10, page 37).

```
<rss version="2.0">
  <channel>
    <title>Mon site exemple</title>
    <description>Flux RSS exemple</description>
    <link>http://www.example.org</link>
    <item>
      <title>Actualité 2</title>
      <description>Contenu d'un billet</description>
      <pubDate>Wed, 27 Jul 2007 04:30:00 -0700</pubDate>
      <link>http://www.example.org/actu2</link>
    </item>
    <item>
      <title>Actualité 2</title>
      <description>Contenu d'un billet</description>
      <pubDate>Mon, 25 Jul 2007 00:30:30 -0700</pubDate>
      <link>http://www.example.org/actu1</link>
    </item>
  </channel>
</rss>
```

Listing 1.10: Exemple de flux RSS 2.0

Alors que les outils du Web 2.0 ont remis en cause certains principes de publication sur le Web, en mettant l'utilisateur final au centre de cette démarche de *production*, la syndication de contenu met celui-ci au centre de la *consommation* d'information. Celle-ci est devenue également personnalisable et paramétrable, par opposition à la diffusion d'informations par e-mail où le lecteur ne peut contrôler la fréquence où l'information lui est délivrée. On retrouve de plus via la syndication de contenu cette notion de Web en tant que plate-forme évoquée précédemment (Section 1.2, page 30), le Web pouvant même être utilisé comme simple interface de diffusion de données. Ceci est d'autant plus flagrant lorsque les contenus sont postés et consommés depuis des applications hors-ligne, le Web n'étant alors plus qu'un *hub* numérique de transit de l'information (Figure 1.9, page 38). Enfin, comme nous le montrerons tout au long de ce mémoire via nos travaux, nous retiendrons de cette section que l'utilisation de modèles communs pour la représentation de données du Web 2.0, dont RSS est un premier exemple, offre de nombreuses perspectives en termes d'échange, d'ouverture et d'interrogation de l'information dans un contexte distribué comme celui du Web.

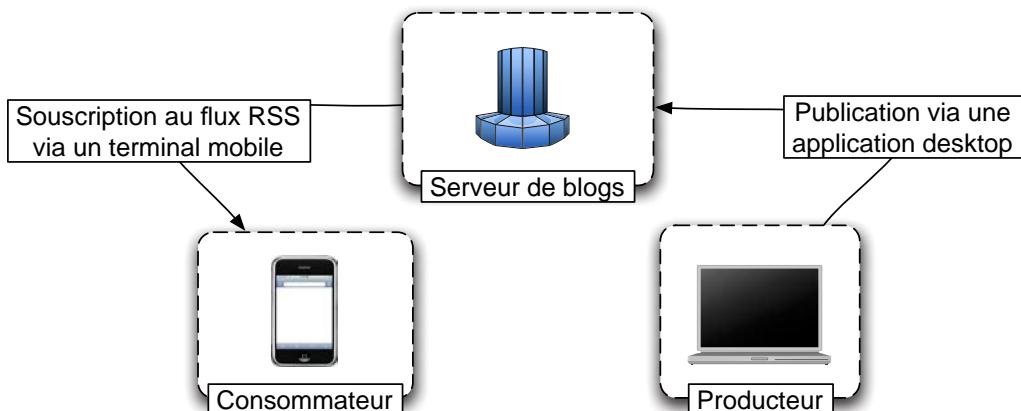


Figure 1.9: Le Web en tant que plate-forme, l'exemple de RSS

1.2.3 Métadonnées sociales : tags et folksonomies

Enfin, face à cette abondance d'informations, facilitée par les outils et services présentés en amont, se pose le problème d'un accès pertinent à celle-ci. Jusqu'à présent, cette tâche était essentiellement rendue possible via des systèmes classiques d'indexation de pages Web. Le Web 2.0 a introduit une autre pratique, basée sur la catégorisation des contenus par les utilisateurs eux-mêmes via l'association aux ressources en ligne de mots-clés libres (aussi bien en type, nombre ou langue), ou tags. Il est important de noter que :

- d'une part cette pratique ne se limite pas aux données textuelles mais qu'il est possible de taguer des ressources numériques aussi diverses que des photos (Flickr) ou des vidéos (YouTube) comme nous le verrons par la suite (Section 1.2.3, page 41)
- d'autre part, certains sites proposent d'étiqueter non seulement les contenus des utilisateurs, mais aussi ceux, déjà tagués, d'autres utilisateurs (Delicious).

Cette pratique s'est également répandue sur la blogosphère, de nombreux billets de blog étant annotés de cette manière, un service comme Technorati permettant ensuite de visualiser ceux-ci et de restreindre la recherche d'information à un tag précis.

De par son rattachement à un contenu existant, un tag peut essentiellement être vu comme une métadonnée supplémentaire associée à une ressource. Cependant, alors qu'un outil de blog associe automatiquement à un billet la date de création de celui-ci et le nom de son auteur, qu'une photo possède ses métadonnées EXIF pour identifier ses caractéristiques, les métadonnées ici générées sont de l'ordre de métadonnées contrôlées et personnalisées par l'utilisateur lui-même, ou métadonnées sociales. Si l'on se réfère à l'usage de métadonnées dans les bibliothèques numériques, on peut en identifier trois types [Taylor, 1999] :

- les métadonnées *descriptives*, caractérisant le contenu de la ressource et utilisées essentiellement dans une optique de recherche d'information ;
- les métadonnées *structurelles*, établissant des liens entre ressources et établies généralement de manière automatique depuis ces mêmes ressources ;
- les métadonnées *administratives*, qui définissent par exemple les droits d'accès ou les

restrictions de copyright de la ressource.

Il est intéressant de constater que si la majorité des tags peuvent facilement être perçus comme des métadonnées descriptives (car essentiellement relatifs au contenu de la ressource, y décrivant ses sujets principaux), certains sont utilisés par les utilisateurs comme des métadonnées administratives ou même structurelles. Ainsi, on observe sur Delicious l'utilisation des tags `creativecommons` ou `gpl` relatifs aux licences du contenu annoté, ou encore `w3c` ou `slashdot` pour indiquer que la ressource est issue du site en question. Des études ont également montré que les tags pouvaient se révéler de diverse nature. Ainsi, [Golder et Huberman, 2006] ont identifié sept usages différents des tags comme l'annotation relative au contenu du document annoté (cas le plus classique), la référence personnelle (`a_lire`), ou l'opinion au sujet d'une ressource (`drole`). [Marlow *et al.*, 2006] ont également montré que les tags pouvaient dans certains cas avoir un aspect social permettant à l'utilisateur de se mettre en avant (ex : `vu_en_concert`). Enfin, [Berendt et Hanser, 2007] ont montré que les tags pouvaient dans certains cas, plus que des métadonnées, être considérés comme du contenu additionnel relatif à la ressource annotée. Quoi qu'il en soit, la pratique des tags est donc assez diverse et dépend fortement du contexte d'utilisation et du vécu numérique des utilisateurs, ce que nous confirmerons plus tard en présentant une utilisation de ces mêmes principes dans un contexte d'entreprise (Section 2.2.3, page 63).

À l'utilisation de ces tags est liée la pratique d'étiquetage ou de *tagging*, association par un utilisateur d'un tag à une ressource donnée (billet de blog, photo ...). Cette relation qui forme ainsi une relation tripartite [Mika, 2005] peut se représenter par :

$$\text{Tagging}(\text{Utilisateur}, \text{Ressource}, \text{Tag}) \quad (1.2)$$

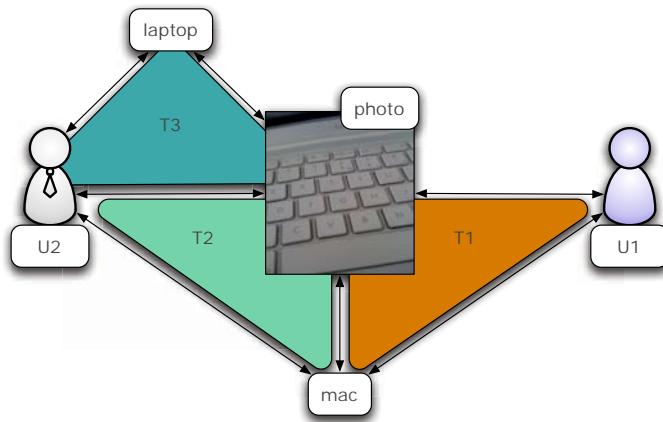
telle que :

- *Utilisateur* correspond à l'utilisateur qui effectue l'action ;
- *Resource* correspond à la ressource annotée (billet de blog, page Web ...);
- *Tag* correspond au tag utilisé ;
- *Tagging* correspond à l'action liant ces trois éléments.

Certains ont proposé de contextualiser cette relation temporellement [Newman *et al.*, 2005] ou en fonction de la source (*i.e* le site) où l'action a été effectuée [Gruber, 2007]. Nous verrons plus tard comment nous proposons d'étendre ce modèle en prenant en compte la signification d'un tag dans un contexte particulier de *tagging* (Section 3.3.2, page 127).

Étant donné que plusieurs tags peuvent être associés par un même utilisateur à une même ressource, et qu'un même tag peut être associé à une même ressource par différents utilisateurs, les actions de *tagging* ne sont en général pas isolées (Figure 1.10, page 40). On utilise donc l'appellation de *social tagging* ou de métadonnée sociales comme nous l'avons évoqué auparavant pour définir ce phénomène. Ainsi, la figure qui suit représente trois actions de *tagging* (T_1, T_2, T_3) associés à une même ressource (*photo*) via deux utilisateurs (U_1, U_2) et deux tags distincts (*mac, laptop*) de la manière suivante :

- $T_1(U_1, \text{mac}, \text{photo})$
- $T_2(U_2, \text{mac}, \text{photo})$
- $T_3(U_3, \text{laptop}, \text{photo})$


 Figure 1.10: Actions de *tagging* combinées autour d'une même photo

Cet ensemble d'actions de *tagging* au sein d'un espace donné (site Web, plate-forme de blogs ...) forme ce qu'on appelle une folksonomie [Vander Wal, 2007], terme hérité du croisement entre *folks* (les gens) et *taxonomy* (taxonomie) et dont la pratique a fait l'objet de nombreuses publications ces dernières années [Mathes, 2004] [Halpin *et al.*, 2007]. Une folksonomie est ainsi issue d'un ensemble d'actions de *tagging* et peut se formaliser comme suit.

$$\text{Folksonomie}(\text{User}^*, \text{Resource}^*, \text{Tag}^*, \text{Tagging}) \quad (1.3)$$

telle que :

- *User** correspond à un ensemble (fini) d'utilisateurs ;
- *Resource** correspond à un ensemble (fini) de ressources annotées ;
- *Tag** correspond à un ensemble (fini) de tags ;
- *Tagging* correspond à la relation qui permet de lier les éléments de ces différents ensembles, telle que définie précédemment (Équation 1.2, page 39).

Si la simplicité de l'approche fait la force des systèmes à base de tags, ceux-ci souffrent de nombreux défauts en termes de recherche d'information, causés aussi bien par les problèmes d'ambiguïté ou de synonymie des mots-clés que par leur nature totalement plate et l'absence de liens entre tags. Nous détaillerons ces différents problèmes par la suite (Section 2.2.3, page 63). *A contrario*, une de leurs forces se situe dans leur utilisation en termes de navigation et dans les possibilités qu'ils offrent pour la découverte de nouvelles informations. L'organisation des liens entre ressources, tags et utilisateurs forme en effet un graphe dans lequel il est possible de naviguer renforçant ainsi la sérendipité, qu'il s'agisse de découvrir de nouveaux documents ou de nouveaux utilisateurs. La popularité des tags dans une folksonomie est d'autre part rendue visible par l'utilisation de nuages de tags ou *tagclouds*, offrant également un autre mode de navigation pour les systèmes à base de tags (Figure 1.11, page 41). Ces interfaces permettent également d'avoir un aperçu du champ lexical associé à une folksonomie, et peuvent être restreintes aux tags d'un utilisateur donné.



Figure 1.11: Exemple de nuage de tags (Delicious)

Partage de contenus, réseaux sociaux en object-centered sociality

En complément des outils et pratiques présentés jusqu'à maintenant, il nous semble important d'évoquer des services de partage de contenus et les notions de réseaux sociaux qui en découlent. Si les outils comme les blogs et les wikis permettent principalement l'édition et le partage de documents textuels, de nombreux services visent à permettre la mise en ligne (toujours via des interfaces simples d'accès) de divers types d'informations : *bookmarks* (Delicious), photos (Flickr), vidéos (YouTube), transparents (SlideShare⁶⁸), etc. Dans un objectif de découverte des contenus publiés, ces services se basent généralement sur l'utilisation de tags tels que présentés auparavant (Section 1.2.3, page 38).

Plus que le simple partage de contenus, une des forces de ces services est la notion de réseaux sociaux associés aux différents contenus publiés. Par exemple, Flickr permet de poster des photos mais surtout de créer des groupes autour de différentes thématiques, sur lesquels des discussions peuvent être menées. Il en est de même sur SlideShare, où les transparents peuvent être rassemblés au sein de groupes particuliers, offrant ainsi une visibilité plus grande aux contenus publiés. Dans un registre différent, last.fm⁶⁹ permet d'établir des communautés d'intérêt autour d'artistes ou de styles musicaux. Si d'autres services sont consacrés à la simple élaboration de réseaux sociaux, qu'ils soient destinés à un public professionnel comme LinkedIn⁷⁰ ou de manière plus large comme Facebook⁷¹, les services sont en général centrés autour de notions (photos, vidéos, etc.) ou de thématiques particulières (style musical, technologie, etc.). Ainsi, comme souligné par [Breslin et Decker, 2007], on parle souvent d'*object-centered sociality*⁷², les utilisateurs échangeant et se retrouvant autour de ces objets particuliers pour y former des réseaux sociaux.

⁶⁸<http://slideshare.net>

⁶⁹<http://last.fm>

⁷⁰<http://linkedin.com>

⁷¹<http://facebook.com>

⁷²http://www.zengestrom.com/blog/2005/04/why_some_social.html

Cette notion de réseaux sociaux est également présente sur les blogs. [Cardon *et al.*, 2007] ont ainsi présenté différentes manières dont des communautés pouvaient se former autour d'un blog donné, en fonction des liens que les membres d'une communauté partagent dans la vie réelle, ou bien selon des centres d'intérêts communs. Dans ce contexte, des services comme MyBlogLog⁷³ permettent en outre de formaliser ces communautés et d'interconnecter certaines en fonction de leurs membres.

1.3 COMPLÉMENTARITÉ ENTRE LES DEUX DOMAINES

1.3.1 Synthèse des deux visions

À la lecture des sections précédentes, nous pouvons donc identifier :

- d'un côté une vision du Web axée sur la représentation formelle des données et des moyens d'échanger celles-ci (RDF, ontologies, SPARQL ...) avec le Web Sémantique (Section 1.1, page 12) ;
- de l'autre une vision centrée sur la collaboration entre internautes via des outils à l'ergonomie attractive (système de tags, *mash-ups* ...) (Section 1.2, page 30).

Le tableau suivant synthétise, de manière volontairement exagérée, différents aspects de ces deux visions (Tableau 1.3.1, page 42).

	Web Sémantique	Web 2.0
Destination	Agents logiciels	Humains
Aprioris	Complexité	Pragmatisme
Background	Académique	Développeurs Web
Languages de représentation	RDF(S)/OWL	(X)HTML, Microformats
Modes de publication	Centralisée	Collaboration
Indexation	Annotations et Ontologies	Tags et Folksonomies
Interrogation	SPARQL	APIs propriétaires

Tableau 1.2: Caractéristiques comparées du Web Sémantique et du Web 2.0

Si cette synthèse cloisonne fortement et volontairement ces deux visions, il faut reconnaître que certains aprioris ici présents ont souvent été évoqués pour mettre en opposition celles-ci. On peut par exemple se référer aux discussions via blogs interposés entre Clay Shirky⁷⁴ et James Hendler⁷⁵ au sujet des folksonomies et des ontologies où à l'opposition entre Web Sémantique *top-down* et *bottom-up*⁷⁶, qui est selon nous un non-sens⁷⁷. Ces discussions font écho à une incompréhension générale qui a longtemps causé du tort au Web

⁷³<http://mybloglog.com>

⁷⁴http://www.shirky.com/writings/ontology_overrated.html

⁷⁵<http://www.mindswap.org/blog/2007/11/21/shirkyng-my-responsibility/>

⁷⁶http://www.readwriteweb.com/archives/the_top-down_semantic_web.php

⁷⁷On peut certes considérer qu'il existe des ontologies *top-down* ou *bottom-up*, notamment via la notion de sémantique émergente à partir des tags (Section 3.3.1, page 122), mais l'appellation Web Sémantique *bottom-up* nous semble inappropriée à partir du moment où l'on parle d'un mode de représentation de données.

Sémantique, à savoir la vision d'une unique ontologie centralisée et référente pour décrire le monde, chose dont il n'a jamais été fait état, du moins dans [Berners-Lee *et al.*, 2001]

Malgré ces distinctions, nous pensons comme d'autres [Gandon, 2006] [Gruber, 2008] [Ankolekar *et al.*, 2008] que ces deux visions ne sont pas contradictoires et que, bien au contraire, elles peuvent - et doivent - chacune bénéficier des apports et travaux de l'autre communauté. Ceci doit permettre de converger vers une unique vision du Web, optimisé à la fois pour les humains et les machines, au niveau des modes de publication pour le premier et de la modélisation des données pour le second. C'est cette convergence qui, selon nous, permettra d'aboutir à un *Web de Données* issues d'interactions sociales tout en étant réutilisable de manière autonome via des agents logiciels au sein d'écosystèmes informationnels sémantiques et sociaux (*Social Semantic Information Spaces*) (Figure 1.14, page 46).

Nous allons ainsi dans les sections suivantes présenter de manière assez générale comment nous envisageons cette convergence et comment se situent certains travaux au sein de cette mouvance de *Social Semantic Web* ou *Semantic Web 2.0* qui progresse depuis quelques années [Breslin et Decker, 2006]. Ces réflexions seront au centre des travaux présentés dans les chapitres suivants, où nous détaillerons les modèles de représentation et les outils que nous avons mis en place pour y parvenir, notamment au sein d'un médiateur sémantique collaboratif pour l'Entreprise 2.0 (Section 2.3, page 69). Les idées qui suivent sont ici présentées essentiellement dans une perspective de réflexion qui permettra au lecteur de mieux appréhender la suite de ce mémoire. Celles-ci seront en outre reprises en détail dans les chapitres suivants.

1.3.2 Apports du Web 2.0 pour le Web Sémantique

Si l'on se base sur la vision du Web 2.0 en tant que système centré sur l'utilisateur (Section 1.2, page 30), il nous semble important pour le Web Sémantique de réutiliser certains paradigmes de celui-ci afin de monter en puissance :

- l'utilisation d'outils simples pour la production à grande échelle de données formalisées selon les principes du Web Sémantique, publiées de manière personnelle (blogs) ou collaborative (wikis). Ainsi, et nous le verrons par la suite, les blogs et les wikis peuvent s'avérer des interfaces efficaces pour la production d'annotations sémantiques, sans pour autant confronter l'utilisateur aux modèles sous-jacents (Section 4, page 137) ;
- la masse importante d'utilisateurs passés du statut de consommateur à celui d'acteur. Si le Web 2.0 est en effet un *read-write Web*, qui plus est collaboratif, les outils du Web Sémantique peuvent ainsi bénéficier d'une masse importante d'utilisateurs producteurs de données formalisées, pour autant que les outils soient simples et adaptés comme indiqué dans le point précédent ;
- la collaboration entre utilisateurs à des fins de création collective et consensuelle d'informations et de connaissances, en corollaire des éléments précédents. Ainsi, les folksonomies mais surtout les wikis peuvent être utilisés pour peupler ou maintenir des ontologies de manière collaborative, comme nous le verrons par la suite (Section 4.2.1, page 148) ;
- l'utilisation d'interfaces simples, ergonomiques et intuitives, pour la visualisation et

la navigation de graphes complexes d'annotations sémantiques. Si ces structures de données sont relativement complexes, l'utilisation d'outils comme des interfaces à facettes permettront de masquer cette complexité aux utilisateurs (Section 5.2.3, page 203) ;

- plus généralement la mise en place de *mash-ups* sémantiques attractifs, proposant des nouveaux moyens de parcourir et visualiser ces informations modélisées en RDF et issues de sources de données réparties sur le Web (Section 5.2.3, page 206).

On considère donc ici le Web 2.0 comme un support à la création, visualisation et manipulation de données formalisées selon les principes du Web Sémantique (Figure 1.12, page 44). En effet, la plupart des outils actuels du Web Sémantique nécessitent un temps d'apprentissage qui n'est pas négligeable, à ajouter aux connaissances nécessaires pour aborder la modélisation de données selon ses principes et l'utilisation d'ontologies. Si la réussite du Web en termes de publication de contenu est passée en partie par l'utilisation d'éditeurs (X)HTML simples, d'interfaces WYSIWYG et autres outils avec un temps d'apprentissage très faible (blogs, wikis ...), nous pensons que les technologies du Web Sémantique ont tout à gagner en proposant également des interfaces intuitives et collaboratives pour la visualisation et la production d'annotations sémantiques.

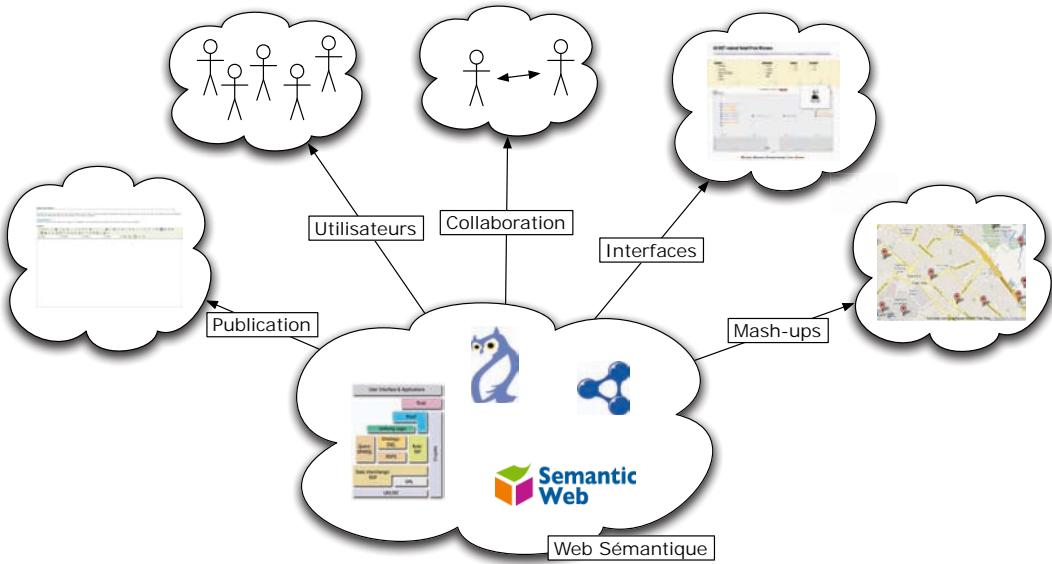


Figure 1.12: Web 2.0 pour le Web Sémantique

1.3.3 Apports du Web Sémantique pour le Web 2.0

Réiproquement, si les outils du Web 2.0 proposent des méthodes qui nous semblent efficaces en termes d'usages et d'interfaces, nous pensons que l'utilisation des technologies

du Web Sémantique ne peut être qu'un plus en termes de structuration et d'échange de données sur le Web 2.0.

En se basant sur les outils et langages du Web Sémantique, les outils du Web 2.0 peuvent ainsi profiter de (Figure 1.13, page 45) :

- l'utilisation de modèles communs pour représenter leurs métadonnées, logiquement basés sur RDF, en lieu et place d'APIs hétérogènes. SIOC [Breslin *et al.*, 2005] répond à cette problématique en proposant un vocabulaire pour définir les métadonnées du Web 2.0 (notion de billet, d'utilisateur, de commentaire ...) via les technologies du Web Sémantique. Il permet ainsi de ne plus considérer blogs, wikis et autres services en ligne comme des îlots indépendants et isolés sur le Web, mais comme des services interconnectés où l'échange d'informations peut se faire de manière transparente (Section 3.1, page 84) ;
- l'utilisation d'ontologies métier pour permettre la structuration des connaissances produites via ces outils. L'utilisation d'ontologies du domaine doit permettre de capitaliser des connaissances (issues par exemple de blogs ou de wikis) de manière formelle à des fins de réutilisation entre services. En ce sens, les wikis sémantiques nous semblent un bon exemple d'utilisation de technologies du Web Sémantique pour augmenter le potentiel d'outils existants et déjà bien acceptés sur le Web 2.0 (Section 4.2.1, page 148) ;
- l'utilisation de protocoles de requêtes et d'échange standardisés. L'utilisation de RDF pour la production de données et de SPARQL pour leur interrogation permet ainsi de simplifier l'interopérabilité entre applications. On favorise en ce sens la découverte de contenus répartis sur différents services Web 2.0 ainsi que la création de *mash-ups* sémantiques à moindre coût.

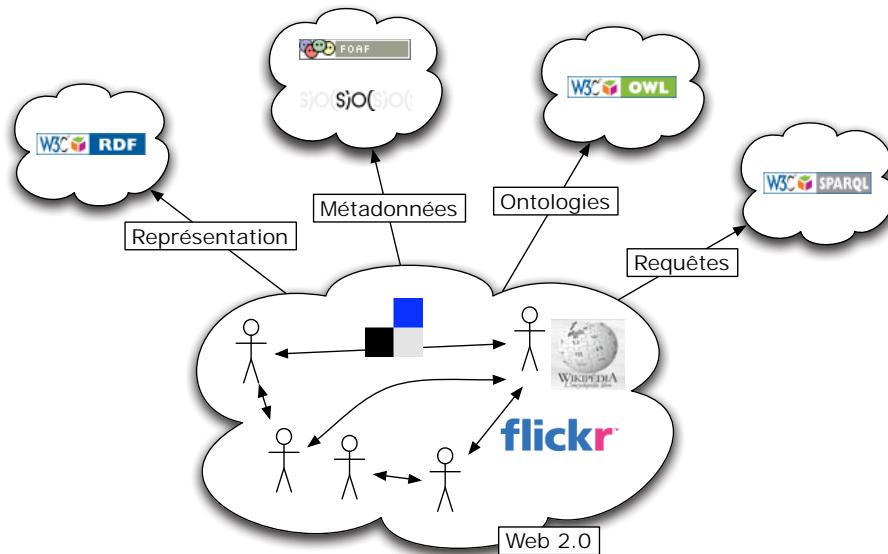


Figure 1.13: Web Sémantique pour le Web 2.0

Ainsi, les outils du Web 2.0 peuvent bénéficier des technologies du Web Sémantique pour assurer la structuration et l'homogénéité des données produites : en s'affranchissant des formats internes et d'APIs propriétaires, on facilite les échanges entre et depuis des systèmes originellement hétérogènes. En conséquence, les outils du Web 2.0 peuvent également participer à cet essor du *Web of Data*, en produisant non plus de simples documents, mais un ensemble de données interopérables.

CONCLUSION

Ce chapitre nous a permis de présenter différents concepts essentiels pour la compréhension de nos travaux. Nous avons tout d'abord détaillé les principes et langages du Web Sémantique, en termes de représentation des connaissances et d'interrogation, puis présenté l'initiative *Linking Open Data* (Section 1.1, page 12). Dans la seconde partie, nous avons introduit les changements et paradigmes introduits par le Web 2.0, en présentant plus précisément certains outils et pratiques, à savoir blog, wikis, syndication de contenu et principes de *tagging* (Section 1.2, page 30). La dernière partie de ce chapitre nous a par la suite permis d'introduire certaines pistes relatives à la convergence entre ces deux visions, convergence qui sera au cœur des travaux que nous allons présenter dans la suite de ce mémoire.

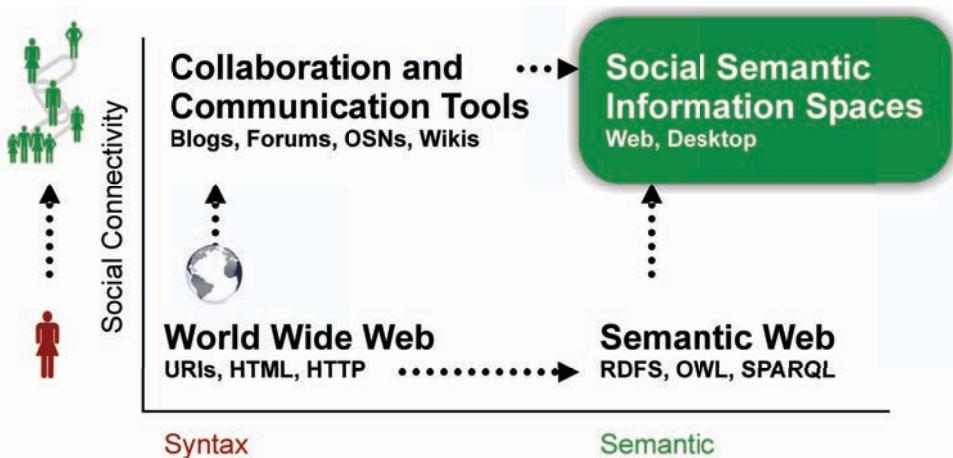


Figure 1.14: Convergence entre Web Sémantique et Web 2.0 [Breslin et Decker, 2006]

Cette convergence, qu'on l'appelle Web n.0, *Social Semantic Web* ou *Metaweb*⁷⁸, permettra d'aboutir à :

- des contenus Web issus d'interactions sociales entre internautes et interopérables grâce à l'utilisation combinée de RDF et d'ontologies pour définir la structure et la sémantique de ces contenus ;
- un *Web de Données*, et non plus seulement un Web de documents, puisque l'on considère alors les systèmes Web 2.0 comme fournisseurs de données interopérables, définies selon les principes évoqués au point précédent ;

⁷⁸http://novaspivack.typepad.com/nova_spivacks_weblog/2003/12/the_birth_of_th.html

- des outils en ligne simples d'utilisation pour créer et mettre à jour ces différentes données, comme les blogs et les wikis agrémentés de capacités de représentation sémantique de l'information ;
- des interfaces de navigation, d'interrogation, de visualisation et des *mash-up* intuitifs et simples d'accès capables d'absorber ces données complexes et réparties pour proposer des services pertinents à l'utilisateur final.

C'est à travers cette complémentarité que pourront se former des espaces informationnels à la frontière de ces deux domaines, utilisant au maximum le potentiel des différents courants actuels du Web (Figure 1.14, page 46). Nous allons ainsi dans la suite de cette thèse identifier différents moyens, aussi bien en termes de modèles de représentation (Section 3, page 83) que d'applications (Section 4, page 137) (Section 5, page 185), de parvenir à cette convergence. Enfin, pour conclure ce chapitre, on citera [Berners-Lee, 2005b] : "*I think we could have both Semantic Web technology supporting online communities, but at the same time also online communities can also support Semantic Web data by being the sources of people voluntarily connecting things together*", pour mettre à nouveau l'accent sur ce qui est non pas un apport à sens unique, mais une véritable complémentarité entre ces deux visions.

Chapitre 2

SemSLATES : Une approche sémantique pour l'Entreprise 2.0

INTRODUCTION

Après avoir introduit dans le chapitre précédent les caractéristiques du Web Sémantique et du Web 2.0 ainsi que différentes pistes relatives à une complémentarité entre ces deux approches, nous allons présenter ici l'utilisation des technologies du Web 2.0 dans un contexte d'entreprise.

Nous présenterons tout d'abord le contexte du projet Athéna au sein d'EDF R&D (Section 2.1, page 50). Ce projet vise à introduire au sein de l'entreprise différents outils Web 2.0 pour faciliter les échanges et la constitution collaborative d'information au sein de celle-ci, dans la mouvance de l'Entreprise 2.0 [Mcafee, 2006] (Section 2.1.1, page 50). Nous reviendrons sur les différents besoins du projet et la manière dont les différents outils mis en place (Section 2.1.2, page 53) offrent une complémentarité permettant de répondre au paradigme *SLATES* introduit par cette notion d'Entreprise 2.0 (Section 2.1.3, page 57). Cette première partie sera également l'occasion de revenir sur les statistiques d'utilisation de ces outils au sein d'EDF R&D (Section 2.1.4, page 59).

Nous montrerons ensuite en quoi cette plate-forme et plus généralement les systèmes d'information d'Entreprise 2.0 (tout comme les outils Web 2.0 pris indépendamment) sont limités sur certains points et nous proposerons ainsi la méthodologie *SemSLATES* permettant d'enrichir ce paradigme via une architecture de médiation de données basée sur les technologies du Web Sémantique (Section 2.3.1, page 69). Notre entendons ici par architecture de médiation la mise en place d'un ensemble d'outils et de modèles permettant d'intégrer les données des différents services de cette plate-forme grâce à une sémantique commune, afin notamment de proposer un ensemble de services additionnels. Nous présenterons alors la vision générale et l'architecture de ce système de représentation et d'intégration sémantique pour l'Entreprise 2.0. Nous présenterons ensuite les différents composants de cette architecture, aussi bien en termes de modèle que de production de données formalisées, qui nous permettront de détailler les limites des outils Web 2.0 que notre méthode vise à résoudre aussi bien en termes de capitalisation des connaissances que d'indexation documentaire.

Outre une description approfondie de l'écosystème sémantique d'entreprise que nous proposons, ce chapitre donnera une vision d'ensemble de nos travaux vis-à-vis (1) de la définition de modèles de représentation de données pour l'Entreprise 2.0 et (2) de la mise

en place d'outils associés pour produire et utiliser des connaissances représentées selon ces modèles. Ces points seront en outre approfondis dans les chapitres suivants de ce mémoire.

2.1 WEB COLLABORATIF EN ENTREPRISE : LE PROJET ATHÉNA

2.1.1 Origine et objectifs du projet

Afin de collecter, analyser et diffuser l'information provenant de différentes sources externes à destination de ses ingénieurs, chercheurs et dirigeants, EDF R&D dispose du groupe ID-Net et plus particulièrement de la Cellule Appui-Veille (CAV), situés au sein du Secrétariat Général¹. La mission essentielle de cette cellule est d'offrir une structure d'Intelligence Economique [Wilensky, 1967] à la R&D, sur des thèmes aussi bien techniques que sociaux ou économiques. [Martre, 1994] définit l'Intelligence Economique comme "*l'ensemble des actions coordonnées de recherche, de traitement et de distribution, en vue de son exploitation, de l'information utile aux acteurs économiques*". Dans le contexte d'EDF ces actions permettent ainsi au personnel de la R&D d'être au fait des dernières innovations, partenariats et technologies utilisées – ou potentiellement utilisables – concernant leur activité. Elles permettent de plus aux dirigeants d'avoir une vision globale de différents domaines permettant d'élaborer ou d'affiner la stratégique du groupe.

Jusqu'à récemment, une partie de ce processus reposait sur des méthodes classiques de veille, capitalisation et diffusion de l'information en entreprise. Parmi les process et outils mis en place, citons l'utilisation d'outils de collecte et de suivi de sites Web comme WebSite-Watcher², la capitalisation de connaissances via des bases Lotus Notes³ ou encore la diffusion d'informations par la voie classique du courrier électronique. Début 2005, commandité par la direction de la R&D, le projet Athéna a vu le jour, avec des objectifs doubles :

- d'une part, optimiser et mutualiser la collecte, la capitalisation et la diffusion de l'information via des solutions innovantes ;
- d'autre part, mettre en place des processus collaboratifs à différents niveaux de cette chaîne informationnelle, notamment en termes d'échanges et de coconstruction de connaissances.

De part son domaine d'activité et son historique, EDF est une entreprise où la culture du secret et des réseaux sociaux informels prédomine, au détriment d'une circulation globale de l'information entre individus. Ceci s'explique en partie par la nature des sujets abordés par les experts de l'entreprise, qu'ils soient sensibles pour des raisons de sécurité (nucléaire) ou de stratégie et d'innovation (énergies renouvelables). Plus généralement, une autre composante de cette absence d'échange intra-entreprise s'explique, comme dans beaucoup d'organisations, par la nature même du savoir, souvent équivalent au pouvoir. Les connaissances sont ainsi la propriété de celui qui les possède, disséminées au compte-goutte de façon plus ou moins formelle et généralement uniquement à un cercle privé de relations. En conséquence, cette rétention d'information se fait au détriment de l'entreprise, de ses compétences et éventuellement de sa stratégie à adopter vis-à-vis de domaines émergents.

¹Celui-ci gère l'administration des trois sites français de la R&D.

²<http://aignes.com>

³<http://www.ibm.com/software/fr/lotus>

Ainsi, en cherchant à repousser les frontières d'une information cloisonnée tout en y introduisant une composante participative, le projet vise à faire entrer l'Intelligence Collective [Bonabeau et Theraulaz, 1994] au sein de l'entreprise. Un des objectifs visé par le projet est donc d'entraîner une synergie permettant de faire émerger des connaissances supérieures à celles que pourraient produire isolément chacun des individus, selon la maxime "*We are smarter than me*"⁴, [Libert *et al.*, 2007]. La réussite de ce projet ne repose donc pas uniquement sur la technique – avec la mise en place de nouveaux outils (Section 2.1.2, page 53) – mais également sur des aspects sociologiques et organisationnels, à savoir l'adoption de ces outils et des pratiques associées par les utilisateurs. De manière plus globale, le projet Athéna se situe dans la mouvance de l'Entreprise 2.0 [McAfee, 2006], vision où les outils du Web 2.0 et les méthodes collaboratives associées – de plus en plus communes dans la sphère personnelle – pénètrent les murs de l'entreprise : "*Enterprise 2.0 is the use of emergent social software platforms within companies, or between companies and their partners or customers*"⁵. Cette vision de l'entreprise où le côté social joue un rôle majeur dans l'élaboration de connaissances rejoint également la notion d'*écologie de l'information* proposée par [Davenport et Prusak, 1997], où l'humain est au centre du système d'information.

Tout comme le Web 2.0, la notion d'Entreprise 2.0 est relativement porteuse, que cela soit pour la communication interne ou externe des entreprises. Même si nous nous sommes intéressés à cette mouvance principalement en termes d'informations internes (Section 2.1.2, page 53), notons la place importante de ces solutions pour favoriser la communication entre certaines entreprises et leurs clients ou le grand public. Une récente étude montre ainsi que près de 13% des entreprises du top 500 de Fortune ont un blog public maintenu par les employés⁶. Les blogs ne sont d'ailleurs pas les seuls outils utilisés puisque l'on retrouve certaines entreprises sur Twitter ou SecondLife⁷, univers virtuel en ligne. Le premier peut être utilisé pour informer ses clients de la mise en place de nouveaux services ou pour simplement communiquer directement avec eux, comme le fait par exemple le service Web 2.0 SlideShare (Figure 2.1, page 52), alors que le second est utilisé dans certains cas pour procéder à des entretiens de recrutement en ligne⁸.

D'un point de vue de l'impact économique de l'Entreprise 2.0, le marché est également porteur et devrait en outre, selon différentes études, évoluer dans les années qui viennent. Forrester Research prédit ainsi un marché global pour les solutions d'Enterprise 2.0 de 4.6 milliards de dollars en 2013 <http://www.forrester.com/Research/Document/Excerpt/0,7211,43850,00.html> alors que Gartner identifie que les plates-formes de *social computing*⁹ seront adoptées par les entreprises dans les dix prochaines années¹⁰. Autre signe de cet essor, de nombreuses solutions logicielles clé-en-main sont aujourd'hui disponibles, comme

⁴<http://www.wearesmarter.org/>

⁵<http://andrewmcafee.org/blog/?p=76>

⁶<http://www.asia.socialtext.net/bizblogs/index.cgi>

⁷<http://secondlife.com>

⁸http://online.wsj.com/public/article/SB118229876637841321-NkCuEAak8wFXmvmpVWkALxqNS3M_20070719.html

⁹Comprendre réseaux sociaux.

¹⁰<http://gartner.com/it/page.jsp?id=739613>



Figure 2.1: Utilisation de Twitter par le service Web 2.0 Slideshare pour communiquer avec ses utilisateurs

IBM Lotus Connections¹¹ ou Jive Clearspace¹². Certaines entreprises se spécialisent également dans ce domaine aussi bien d'un point de vue technique que pour l'accompagnement à l'utilisation de tels outils, comme SocialText¹³ ou HeadShift.

Pour en revenir à la notion même d'Entreprise 2.0, [Mcafee, 2006] évoque en définissant ce terme la manière dont des outils comme les blogs et les wikis permettent de transformer les intranets en structures dynamiques et évolutives grâce à la participation des utilisateurs. Il caractérise également les différents principes introduits par ces outils par l'acronyme *SLATES* :

- *Search* – Recherche d'information ;
- *Links* – Liens entre contenus ;
- *Authoring* – Publication aisée ;
- *Tags* – Annotations des contenus par tags ;
- *Extensions* – Extension de la navigation ;
- *Signals* – Signalement d'information.

Par exemple, les blogs et les wikis peuvent être utilisés pour la publication d'information (*Authoring*) et la définition de liens entre document (*Links*) de manière intuitive sans aucun prérequis technique. Les systèmes à base de tags peuvent quant à eux être utilisés pour annoter les contenus publiés (*Tags*) et favoriser la découverte de nouvelles informations (*Extensions*). De plus, les principes de syndication RSS mais aussi des outils comme le microblogging peuvent être utilisés pour favoriser le signalement de nouvelles informations (*Signals*). Ce dernier mode de communication et de partage de l'information nous semble de plus

¹¹<http://www-01.ibm.com/software/lotus/products/connections/>

¹²<http://www.jivesoftware.com/products/clearspace>

¹³<http://www.socialtext.com/>

particulièrement adapté à cette notion de signalement puisqu'il offre une méthode de communication agile et spontanée au sein de l'entreprise. En complément, la plupart des outils bénéficient de capacité de recherche d'information, qu'il s'agisse de recherche plein-texte ou de recherche par tags (*Search*). Ainsi, si l'on peut difficilement contredire le fait que ces outils permettent aux utilisateurs de simplement lier, éditer ou taguer des contenus, nous sommes plus réservés quand à leur capacité à offrir une recherche d'information efficace, des extensions de celle-ci et un signalement d'informations pertinent, comme nous le montrerons plus tard (Section 2.2, page 62).

2.1.2 Répondre efficacement aux différents besoins

Comme nous l'avons évoqué, un des objectifs d'Athéna est la mise en place de nouveaux outils pour faciliter la constitution et l'échange d'informations au sein de la R&D, notamment dans un contexte de veille informationnelle. Différents services ont ainsi été successivement mis en place, labellisés de manière unifiée sous le nom de plate-forme Hermès.

Flux RSS et mutualisation des sources d'information

La première phase du projet a consisté en la mise en place d'un système de collecte et d'abonnement à des flux RSS issus du Web. L'objectif visé est ainsi d'optimiser la collecte, la diffusion et la mutualisation d'informations externes au sein de l'entreprise. C'est d'ailleurs ce que [Mcafee, 2006] identifie comme les canaux de communication permettant de répondre à la problématique de signalement (le second *S* de *SLATES*). Cette pratique d'utilisation de flux RSS externes au sein de l'entreprise est en outre aujourd'hui de plus en plus courante. Un sondage Ipsos datant de décembre 2007 montre ainsi que 21% des décideurs informatiques ont recours aux abonnements à des flux RSS¹⁴. Plus récemment, une étude d'AIIM¹⁵ indique que cette technologie est déjà acquise par 51% des entreprises sondées et que 21% ont prévu de l'intégrer dans leur stratégie [Frappaolo et Keldsen, 2008].

La sélection des flux à collecter se fait de manière continue par la CAV selon les demandes des clients de la plate-forme, *i.e.* les entités de la R&D qui souhaitent suivre l'actualité d'un thème donné. Ces flux sont classés selon différentes thématiques (énergie solaire, télécommunications ...), les utilisateurs pouvant ensuite s'y abonner. Cette interface d'abonnement permet également d'avoir accès aux dernières nouvelles des flux souscrits (Figure 2.2, page 54), ceux-ci étant rafraîchis plusieurs fois par jour.

Les flux sélectionnés peuvent en outre provenir de sites de nature relativement diverse : grands quotidiens, sites d'actualité, mais aussi forums ou blogs d'experts. Cette perspective permet de bien comprendre à quel point la diffusion des connaissances sur le Web, accentuée par l'utilisation d'outils Web 2.0, peut être bénéfique pour une entreprise en termes d'acquisition de nouveaux savoirs. Il est en effet possible de tirer parti des connaissances d'un expert sans que celui-ci n'ait de relation directe avec l'entreprise et ce à moindre coût et sans démarche proactive, au contraire de ce que proposent les *Ideagoras* [Tapscott et Williams, 2007].

¹⁴<http://www.ipsos.fr/CanalIpsos/poll/8359.asp>

¹⁵<http://aiim.org>

The screenshot shows a web-based RSS reader interface. At the top, there are four tabs: "Mes flux", "Mes nouvelles", "Tous les flux", and "Les nouveaux flux". Below the tabs, there are three main sections, each with a newspaper icon and a title:

- Le Monde -- Actualités -- Une -- lemonde.fr -- (FR)**
 - Le parquet va demander l'audition du président du Sénat [3 heures 23 min]
 - L'équipage du "Ponant" a été libéré sans incident [3 heures 38 min]
 - Cameroun : un opposant à Paul Biya jeté en prison pour avoir contesté la révision constitutionnelle [3 heures 50 min]
 - Le marché publicitaire français est touché par la frilosité des annonceurs [4 heures 24 min]
 - 57 600 euros pour le nu de Carla Bruni [5 heures 19 min]

En (sa)voir plus...
- Open Source -- lemondeinformatique.fr -- (FR)**
 - Linux continue de croître sur sa base install@e [1 jour 14 heures]
 - Sun certifie Ubuntu pour ses nouveaux serveurs [1 semaine 14 heures]
 - OpenXML : l'April dénonce « l'influence politique » exercée sur l'ISO et l'Afnor [1 semaine 3 jours]
 - OpenXML devient un standard ISO [1 semaine 3 jours]
 - Annuels Red Hat : le support dope les revenus [1 semaine 4 jours]

En (sa)voir plus...
- EDF dans la presse française -- Google Actualités France -- (FR)**
 - Décraytage: les yeux dans le Bleu ciel d'EDF - Energie2007 [3 heures 17 min]
 - La Bourse salue l'hypothèse d'une offre d'EDF sur British Energy ... - La Tribune.fr [3 heures 22 min]
 - EDF : Les cours forment une tête-et-épaules. - Boursorama [4 heures 7 min]
 - EDF prêt à offrir 11 milliards de livres pour British Energy - Romaniedie.com [5 heures 3 min]
 - Le Cac 40 se reprend, Bouygues, France Télécom et EDF en soutien - Les Échos [5 heures 30 min]

Figure 2.2: Interface personnelle de visualisation de flux RSS au sein d'Hermès

Par rapport à la méthodologie utilisée avant cette pratique d'agrégation, trois progrès importants sont à retenir :

- là où les processus précédents faisaient intervenir différents outils pour agréger les données source, définissant chacun leur propre format, nous disposons via l'utilisation de RSS d'un modèle standard pour la représentation des informations collectées. Ceci se traduit par la possibilité d'utiliser des outils génériques pour la lecture et le stockage des informations agrégées (en l'occurrence des APIs dédiées aux flux RSS) ;
- ce nouveau processus permet également de mutualiser les sources d'information à destination des utilisateurs. Cette mutualisation est une première étape dans la démarche d'Intelligence Collective visée par le projet Athéna. Les flux ne sont en effet plus seulement réservés aux commanditaires de la veille thématique, mais disponibles pour tous les utilisateurs de la plate-forme via l'interface d'abonnement ;
- contrairement à la pratique précédente où les informations étaient envoyées par e-mail à intervalles réguliers, celles-ci sont ici fournies à l'utilisateur à sa demande, *i.e.* à chaque consultation de son interface de lecture, évitant la surcharge d'informations non sollicitées.

Réactions et informations spontanées grâce aux weblogs

Si cette première étape permet de simplifier et de mutualiser l'acquisition et la diffusion d'informations externes au sein de l'entreprise, elle ne prend pas en compte une autre des problématiques initiales. En effet, un autre besoin est de fournir une certaine valeur ajoutée à ces informations brutes et d'échanger autour de celles-ci ou au sujet de nouvelles informations. Nous avons ainsi mis en place une plate-forme proposant un blog à chaque utilisateur le souhaitant. Un premier objectif est la valorisation des éléments de flux RSS, en permet-

tant de créer simplement un billet à partir d'une nouvelle, à la manière d'un outil comme ReBlog¹⁶. Bien entendu, le système ne se limite pas à la création de contenus à partir d'éléments existants, mais offre la possibilité de créer des billets originaux et de commenter les billets existants, intégrant ainsi une composante participative au service. Ce processus répond ainsi au *A* de *SLATES* en permettant à tous de passer du statut de consommateur à celui de rédacteur via la publication de nouvelles informations ou en accentuant le signalement d'informations existantes (second *S* de *SLATES*).

L'intérêt de cette démarche est double :

- premièrement, en matière de mise en valeur de l'information. Une nouvelle issue d'un flux se retrouve rapidement noyée au sein d'une masse importante d'informations. De plus elle n'est pas immédiatement accessible pour les utilisateurs qui n'ont pas souscrit au flux d'origine. La plate-forme de blogs dressant une liste antéchronologique et visible par tous des derniers billets créés, les éléments y bénéficient d'une meilleure visibilité (certes parfois courte, mais qui permet cependant à tous de les remarquer) ;
- en second, en matière de valeur ajoutée et d'analyse pertinente de l'information. Dans le cas où le billet est issu d'informations agrégées, si rien n'empêche l'utilisateur de republier l'information telle quelle, l'objectif est d'y ajouter une analyse personnelle ou *a minima* de la situer dans le contexte EDF. Les aspects les plus pertinents d'une actualité donnée peuvent ainsi être mis en avant par le rédacteur du billet.

Chaque blog disposant à son tour d'un flux RSS, il est possible de s'y abonner pour limiter sa veille personnelle aux informations d'une thématique donnée, chaque utilisateur ayant pour habitude de créer des billets autour d'un sujet spécifique (nucléaire, énergies solaires ...). Là aussi, plusieurs avantages sont à signaler par rapport à l'échange d'information par e-mail. Tout d'abord, en raison de la nature ouverte de la publication (*a contrario* d'un e-mail adressé à une communauté restreinte et établie à priori par le rédacteur), l'information circule de manière plus large. En conséquence, il est possible à un plus grand nombre de personnes d'y réagir, favorisant ainsi les échanges spontanés et l'acquisition de nouveaux savoirs. D'autre part, en plus d'être ouvertes et mutualisées, les informations deviennent pérennes via un système d'archives des billets contrairement (1) aux archives d'e-mails qui disparaissent généralement lorsque leur propriétaire quitte l'entreprise et (2) aux éléments de flux RSS dont la survie dans notre agrégateur n'est pas toujours assurée¹⁷.

Capitalisation d'information via les wikis

Revenons maintenant sur un autre aspect déterminant pour le projet, celui de la capitalisation des connaissances. Bien qu'un pas ait déjà été franchi dans ce domaine avec l'utilisation des blogs, il faut garder à l'esprit qu'un billet de blog représente généralement une connaissance établie à un instant *t*. Un billet de blog insiste en général sur une actualité contextualisée temporellement, comme par exemple la fusion de deux entreprises ou le lancement d'un projet. De ce fait, les informations de ce type ne peuvent pas – du moins sous cette forme de billet brut – être considérées comme des connaissances encyclopédiques (les secteurs d'activité d'une entreprise, la liste de ses dirigeants ...). De plus, en raison de la pré-

¹⁶<http://reblog.org>

¹⁷Pour des raisons légales, certains éléments de flux sont supprimés passé un certain délai.

sentation antéchronologique des blogs, ces billets sont voués à être rapidement remplacés par d'autres en termes d'affichage. Le besoin initial de capitalisation n'est donc pas complètement satisfait et il est nécessaire de fournir une solution permettant de produire efficacement des documents de référence sur divers domaines. Qui plus est, cette solution doit aussi permettre de faire évoluer ces documents, par opposition aux fonds documentaires généralement figés¹⁸.

Devant ce besoin, nous avons naturellement opté pour la mise en place d'un serveur de wikis afin de capitaliser et de construire, non pas des informations volatiles mais des connaissances pérennes et consensuelles. Chaque utilisateur a ainsi la possibilité de créer son propre wiki dédié à un projet ou une thématique donnée mais peut aussi agir sur les différents wikis mis en place par les autres utilisateurs de la plate-forme.

C'est essentiellement via l'utilisation de ces outils que l'on parviendra à visualiser l'émergence d'une Intelligence Collective visée par le projet : l'agrégation d'un ensemble de processus individuels (ajout d'une nouvelle page, modification de contenu existant ...) devant conduire à terme à l'apport de connaissances ayant une valeur ajoutée plus forte que celle des connaissances individuelles (Figure 2.3, page 56). La plate-forme mise en place conserve en outre les caractéristiques essentielles des wikis évoquées précédemment : utilisation importante des hyperliens (*L* de *SLATES*), rétroliens (que l'on peut voir d'une certaine manière comme une extension de la navigation, *E* de *SLATES*), historique des versions, création aisée de nouvelles pages, etc. Nous avons de plus, tout comme pour les blogs, intégré un éditeur WYSIWYG afin de faciliter la courbe d'apprentissage de l'outil, nouveau pour la quasi-totalité des utilisateurs.

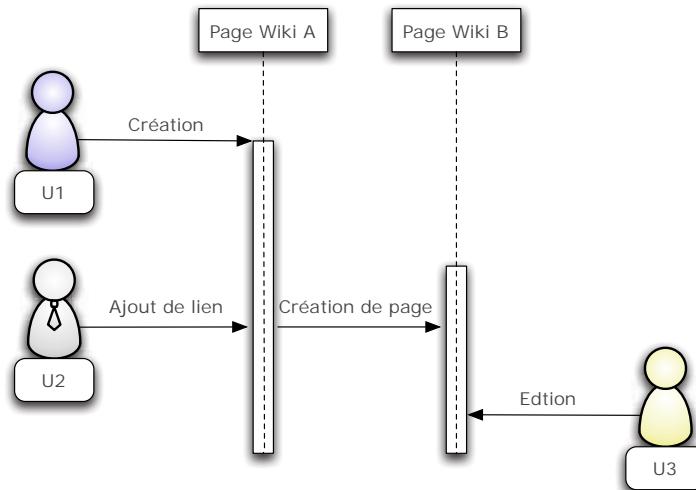


Figure 2.3: Coconstruction de connaissances avec les wikis

¹⁸Ou, quand ils ne le sont pas, nécessitent un processus complexe pour mettre un document à jour.

Indexation documentaire et recherche d'information

La masse d'information que les outils précédents délivrent impose la mise en place de services permettant un accès optimal à celles-ci. Deux processus ont été mis en place dans cet objectif :

- une indexation automatique des contenus reposant sur un moteur plein-texte. Les données provenant de flux RSS, tout comme les billets de blog et les pages wikis internes sont donc indexées à intervalle régulier ;
- une annotation manuelle des contenus produits par les utilisateurs via un système de tags. Les billets de blogs et pages wikis peuvent bénéficier de ce processus permettant à l'utilisateur d'indexer librement ses contenus selon les termes qui lui semblent les plus pertinents. Afin de faciliter le processus et favoriser l'utilisation de tags déjà présents, un système d'autocomplétion a été mis en place. De plus, un système de recherche et de navigation est associé à ce processus. Nous reviendrons plus tard dans ce chapitre sur une analyse de l'utilisation des tags dans ce contexte et sur les problèmes qu'ils soulèvent (Section 2.2.3, page 63).

2.1.3 Complémentarité générale des outils

L'ensemble des outils mis en place permet ainsi de répondre aux objectifs de *SLATES* de la manière suivante (Tableau 2.1.3, page 57) :

Règle	Pratique et outils associés
<i>Search</i>	Moteur de recherche plein-texte (blogs, wikis et flux RSS) et recherche par tags (blogs et wikis)
<i>Link</i>	Utilisation d'hyperliens entre documents internes (notamment via les wikis) ou entre documents internes et informations externes (via la republication RSS)
<i>Authoring</i>	Publication dynamique, personnelle (blogs) ou collaborative (wikis) et facilitée par des interfaces intuitives (éditeur WYSIWYG notamment)
<i>Tags</i>	Utilisation de tags avec système d'autocomplétion pour en faciliter l'ajout (blogs et wikis)
<i>Extension</i>	Liens, rétro-liens et références vers des informations externes (wikis et blogs) ainsi que navigation associée au système de tags
<i>Signals</i>	Agrégation mutualisée d'informations externes (flux RSS) et republication pour leur mise en avant (blogs)

Tableau 2.1: *SLATES* et la plate-forme Hermès

On peut modéliser un scénario optimal d'utilisation de ces différents outils de la manière suivante (Figure 2.4, page 58) :

- un premier utilisateur consulte une nouvelle provenant d'un flux RSS issu du Web auquel il est abonné et signale l'information sur son blog ;
- un second lit ce billet, puis capitalise l'information sur un wiki dédié à la thématique associée ;

- un troisième consulte cette page puis contribue au wiki en créant une nouvelle page à partir de celle-ci ;
- un quatrième intervenant va lire puis commenter le billet d'origine ;
- le second utilisateur va enfin consulter puis éditer la page wiki nouvellement créée.

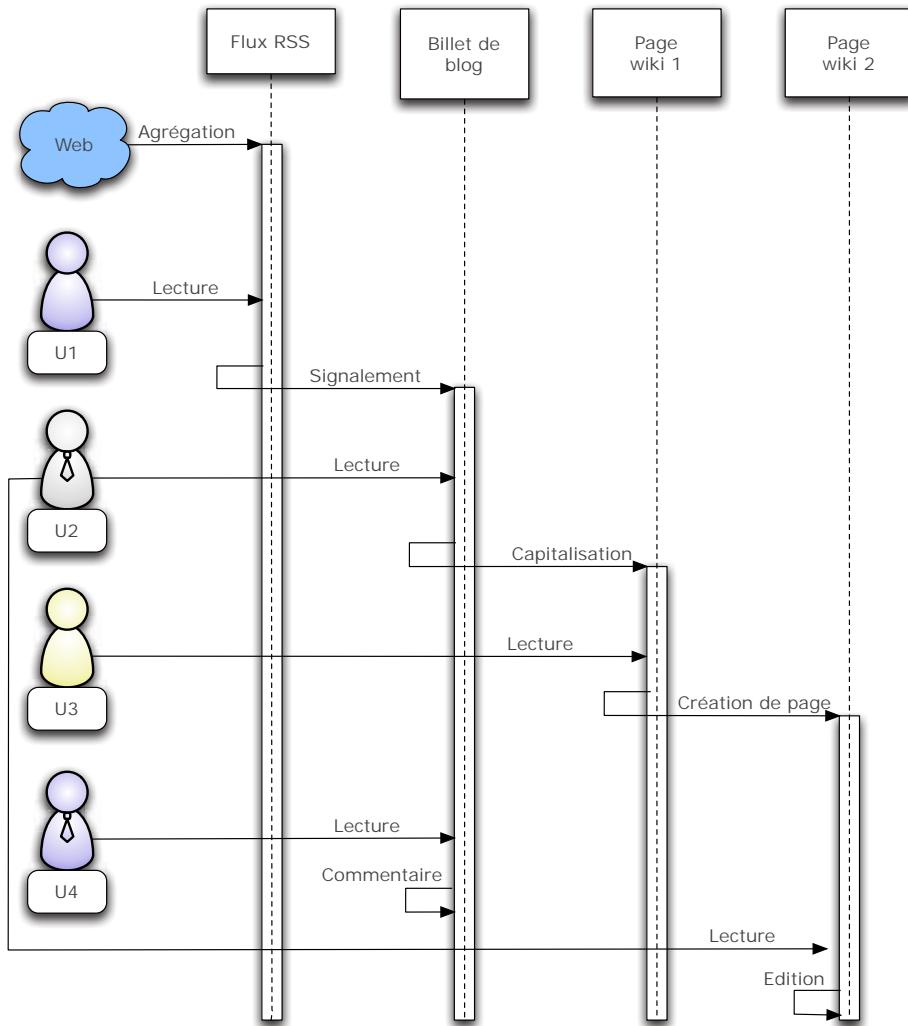


Figure 2.4: Scénario idéal d'utilisation des différents éléments de publication de la plate-forme

Ce scénario met en avant les différents outils et processus introduits par la plate-forme Hermès dans cet objectif d'Intelligence Collective. On y retrouve en particulier les notions de partage d'information et de constitution collaborative de connaissances évolutives.

2.1.4 Retour sur expérience

Avant de revenir sur les limites de cette approche (Section 2.2, page 62), faisons un point sur quelques statistiques qui nous permettent d'évaluer la plate-forme en termes d'acceptation par les utilisateurs. Environ trois ans après son lancement initial et une année après que la plate-forme ait été officiellement labellisée comme élément de l'Intranet de la R&D, les chiffres sont plutôt concluants, puisque plus de 3000 utilisateurs ont fait la démarche de s'y inscrire (Tableau 2.2, page 59). Cependant, environ 6% seulement ont franchi l'étape consistant à passer du statut de consommateur à celui de producteur. Sur ce pourcentage, on notera comme le montre le tableau suivant¹⁹ que la majorité sont des utilisateurs de blogs, même si certains utilisent uniquement les wikis. Notons également qu'environ la moitié des contributeurs ont déjà posté un commentaire sur la plate-forme.

Nombre d'utilisateurs	3068
Nombre global de contributeurs	203
Nombre de contributeurs dans les blogs	167
Nombre de contributeurs dans les wikis	88
Nombre de contributeurs dans les commentaires	109

Tableau 2.2: Utilisateurs et contributeurs au sein d'Hermès

Concernant les flux RSS, la plate-forme dispose de plus de 1500 flux à disposition des abonnés, répartis en près de 300 thèmes (Tableau 2.3, page 59).

Nombre de flux	1528
Nombre de thèmes	295
Nombre moyen d'abonnés par flux	4.46
Nombre maximum d'abonnés à un flux	118

Tableau 2.3: Statistiques des flux RSS au sein d'Hermès

Au niveau des outils eux-mêmes, on constate donc que le blog est l'outil le plus utilisé, avec près de 16000 billets. Seulement 600 d'entre eux ont cependant été commentés, ce qui montre malgré le nombre de billets et d'utilisateurs actifs une certaine timidité dans les réactions (Tableau 2.4, page 60). La composante sociale en termes de conversations et d'échanges sur des sujets d'actualité ou des réflexions personnelles n'est donc pas aussi avancée que souhaitée et les discussions spontanées ne naissent visiblement pas aussi facilement que nous l'aurions espéré. La figure qui suit (Figure 2.5, page 60) illustre également cette différence entre billets et commentaires. On peut notamment la comparer à une étude menée chez DrKW à ce sujet où les commentaires sont dans ce cas beaucoup plus nombreux que les billets [Mcafee, 2006]. Nuançons cependant ce rapport assez faible par le fait que de nombreux billets sont, comme nous l'avons présenté, rédigés dans un processus de signalement des nouvelles issues de flux RSS et n'appellent pas nécessairement à discussion.

¹⁹Statistiques de décembre 2008, tout comme l'ensemble des statistiques qui suivent.

Nombre de billets	21614
Nombre de billets commentés	700
Nombre de commentaires	1195
Nombre de wikis	83
Nombre de pages wikis	4378

Tableau 2.4: Statistiques des contributions utilisateur au sein d'Hermès

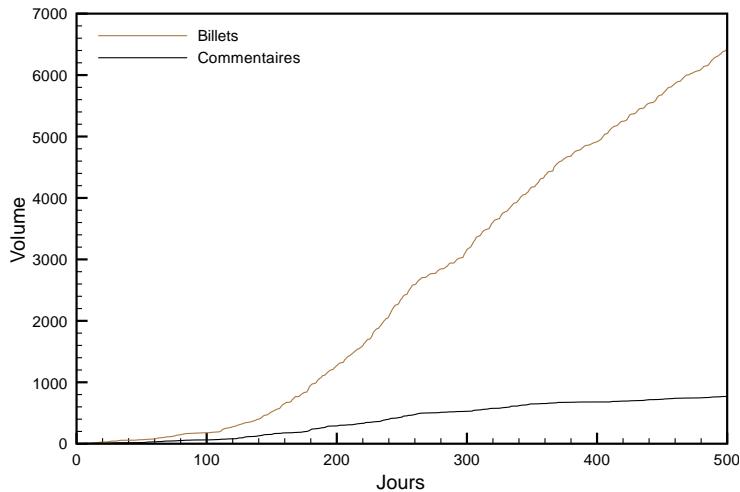


Figure 2.5: Évolution des billets et des commentaires sur la plate-forme

La situation des wikis est différente puisque près de 80 wikis pour plus de 4000 pages ont été créées, ce qui témoigne de la bonne adoption et prise en main de ce type d'outils parmi les utilisateurs. Cinq wikis notamment comptent plus de 300 pages. Bien que l'outil ait été initialement imaginé pour une utilisation à but encyclopédique au sein de la R&D, certaines communautés l'ont adopté spontanément dans une optique de gestion de projet afin d'y stocker les documentations techniques, les derniers comptes-rendus de réunion ou les contacts clients. Il est ainsi important de noter un parallèle qui s'est naturellement établi entre les cas d'utilisations des wikis sur le Web et leur utilisation en interne, malgré des utilisateurs jusque là novices vis-à-vis de ce mode de publication. Cette observation conforte le fait que le wiki est un outil dont les usages et pratiques dépendent fortement des besoins de la communauté qui se l'approprie (Section 1.2.2, page 35). Malgré tout, différents administrateurs ont décidé de restreindre l'édition de leur wikis (voire parfois la lecture) à des groupes prédéfinis. Il est intéressant de remarquer que dans ce cas, certains ont revu leur position en autorisant au final *a minima* la lecture et les commentaires, après avoir eu écho des retours positifs dont bénéficiaient les wikis des communautés ouvertes.

Plus généralement, qu'il s'agisse de blogs ou de wikis, les réticences à la publication et au partage d'information peuvent s'expliquer de différentes manières, tel que nous l'avons

constaté :

- comme nous l'avons déjà évoqué, la valeur de l'information reste essentielle pour celui qui la possède, notamment en termes de reconnaissance dans l'entreprise. Ainsi, il n'est pas toujours évident d'accepter de partager son temps ou ses connaissances ouvertement sans avoir l'assurance que l'on sera valorisé pour des actions de ce type ;
- *a contrario*, certains utilisateurs ne s'aventurent pas dans cette pratique, particulièrement pour les wikis, de peur que les informations qu'ils partagent soient modifiées dans une optique qui ne leur convienne pas. De plus, certains n'entrevoient justement pas l'intérêt de s'y investir, à partir du moment où d'autres seraient tout aussi en mesure d'effectuer cette démarche d'échange ou de capitalisation à leur place.

Rappelons que ces processus de diffusion ouverte d'informations ne faisaient pas jusqu'à présent partie de la culture d'entreprise. Ainsi, passer d'une vision fermée de la diffusion du savoir à un point de vue radicalement opposé mettant en avant l'aspect volontaire et spontané du partage d'informations n'est pas simple à accepter pour la majorité des utilisateurs. Si d'autres entreprises, notamment dans la culture anglo-saxonne ont pu mettre plus en avant cette ouverture comme le montre une étude menée chez Sun et IBM [Kolari *et al.*, 2007], on peut se demander jusqu'où la confidentialité de l'information et le désir de rester garant d'un certain niveau d'expertise prévaut sur le fait de partager celle-ci et d'en faire bénéficier ses pairs et les différentes strates de l'entreprise. C'est une question sociologique à laquelle nous ne tenterons pas de répondre, mais qui révèle bien les impacts que ce nouveau mode de participation et d'échange de savoir ont au niveau d'entreprises dont la culture a été toute autre pendant de longues années. Cette relation entre la culture d'entreprise et l'acceptation d'un système d'information d'Entreprise 2.0 se retrouve également dans l'étude d'AIIM évoquée précédemment qui indique que 41% des sondés n'ont pas de compréhension claire de la notion d'Entreprise 2.0, contre seulement 15% pour les entreprises orientées *Knowledge Management*. Ainsi, il est important de garder à l'esprit que, plus qu'un ensemble d'outils et de prérequis technique, l'Entreprise 2.0 est une philosophie qui peut parfois prendre du temps pour être acceptée. Comme le souligne également Dion Hinchcliffe²⁰, "*l'entreprise 2.0 est davantage un état d'esprit qu'un produit que l'on peut acheter*".

Malgré tout, les chiffres obtenus nous semblent encourageants pour la suite du projet et l'usage croissant des wikis laisse entrevoir de nombreuses communautés demandeuses de cette pratique à l'avenir. Un autre point qui nous semble favorable à une augmentation du nombre d'utilisateurs et de contributeurs aux outils est une combinaison des stratégies *top-down* et *bottom-up* pour faire entrer l'outil dans les mœurs, comme nous avons pu le constater et tel que suggéré par Suw Charman²¹ :

- d'une part, stratégie *bottom-up*, les outils mis en place et testés successivement auprès des différentes communautés ont permis de faire connaître la plate-forme par bouche-à-oreilles. Certains utilisateurs sont même devenus évangélistes de la plate-forme, participant aux actions de communication autour de celle-ci (interviews par exemple) ;
- d'autre part, stratégie *top-down*, le management et l'équipe projet ont régulièrement promu l'outil via différents canaux de communication, qu'il s'agisse de messages à

²⁰<http://blogs.zdnet.com/Hinchcliffe/?p=143>

²¹<http://strange.corante.com/2006/03/05/an-adoption-strategy-for-social-software-in-enterprise>

destination des différents départements ou de séminaires plus larges au sein de la R&D.

La combinaison de ces stratégies fait que l'outil est aujourd'hui connu dans l'ensemble de la R&D et que les demandes d'abonnement sont de plus en plus courantes. Nous pouvons ainsi penser que la philosophie associée à l'Entreprise 2.0 évoquée plus haut est en route, malgré certaines réticences initiales. De plus, l'adoption de ces nouvelles méthodes de travail devrait selon nous croître durant les prochains cycles de vie du projet, la masse critique de contributeurs étant maintenant en place et ayant permis de lancer la dynamique initiale. On peut également imaginer que de nouveaux outils viennent s'ajouter au système actuel, offrant une nouvelle dynamique d'échange qui pourrait intéresser de nouvelles communautés. Nous pensons par exemple à l'introduction d'outils de messagerie instantanée, de microblogging ou encore de partage de signets, proposant ainsi de nouvelles manières de répondre au paradigme *SLATES*.

2.2 LIMITES DE L'APPROCHE CLASSIQUE

Alors que nous venons de recenser quelques limites sociologiques et culturelles à la mise en place de ces outils dans un contexte d'entreprise, nous allons ici présenter différentes problématiques auxquelles nous avons été confrontés, aussi bien en implantant qu'en utilisant ces outils.

2.2.1 Fragmentation de l'information et hétérogénéité des formats

Comme nous l'avons présenté auparavant différentes suites logicielles sont disponibles pour introduire les outils Web 2.0 dans un contexte d'entreprise. Or, l'Entreprise 2.0 repose dans de nombreux cas sur un ensemble de services indépendants, plusieurs raisons peuvent conduire à cette disparité :

- les outils peuvent par exemple avoir été introduits par les employés eux-mêmes, sans consultation préalable des autres équipes ou de la direction. Une équipe va donc créer son wiki de gestion de projet, une deuxième va installer un autre wiki pour ses documentations logicielles, alors qu'une troisième va mettre en place son propre agrégateur de flux RSS ou sa plate-forme de blogs ;
- une autre cause peut simplement être due à la nature des outils, des services demandés et à l'évolution des besoins. On peut par exemple envisager une plate-forme uniquement dédiée aux blogs et aux wikis et se rendre compte, au moment d'introduire des outils de microblogging, que celle-ci ne permet pas une telle utilisation, un nouvel outil étant alors introduit.

Cette diversité des outils introduit en conséquence un problème de fragmentation de l'information. Comme nous l'avons évoqué dans le chapitre précédent, la notion de partage d'informations sur le Web 2.0 est en général centrée autour d'objets particuliers (Section 1.2.3, page 41). Il en est de même en entreprise où les échanges et requêtes sont généralement centrés autour d'une société, d'un projet, d'un domaine technologique. Or, en raison de la diversité des outils utilisés, l'information peut-être répartie au sein de plusieurs systèmes. Un utilisateur devra donc interroger plusieurs sources d'information puis recouper les résultats, le coût de ce processus étant proportionnel au nombre d'outils. Dans notre

contexte, il arrive fréquemment que l'information au sujet d'un domaine particulier soit répartie au sein de plusieurs wikis, blogs et flux RSS. Si le moteur de recherche plein-texte ou l'utilisation des tags permettent en partie d'assister l'utilisateur dans cette tâche, nous verrons sous peu qu'ils soulèvent également de nombreux problèmes.

Conséquence de cette disparité des applications, les formats de données sont également distincts. La tâche d'intégration se révèle donc fastidieuse pour le développeur, avec différentes structures de base de données, APIs ou formats d'échange à appréhender et intégrer. On retrouve cette même problématique sur le Web où les données sont encore plus hétérogènes et distribuées que dans ces systèmes d'entreprise où l'on peut en général identifier plus facilement les sources d'information.

2.2.2 Capitalisation des connaissances

Si les wikis sont abondamment utilisés (comme nos statistiques le montrent (Section 2.1.4, page 59)) pour la capitalisation d'information, ils souffrent de certains défauts qui ne permettent pas d'exploiter celle-ci de manière optimale. Malgré la puissance de l'outil (édition libre, archivage des versions, liens bidirectionnels ...) il est en effet difficile d'accéder rapidement à l'information souhaitée. En effet, de par la nature déstructurée et extensible des wikis, les informations au sujet d'une thématique particulière peuvent être réparties sur un grand nombre de pages. On retrouve ici une partie du problème évoqué précédemment, mais cette fois ci à l'échelle de l'outil.

Selon nous, le principal problème des wikis dans cet objectif de capitalisation efficace des connaissances est lié à leur nature plein-texte. Celle-ci fait qu'il est en effet difficile, à moins d'utiliser des algorithmes complexes de traitement des langues et/ou d'extraction d'entités nommées et de relations, d'interpréter et réutiliser automatiquement le contenu des différentes pages. En conséquence, comme nous l'avons déjà évoqué en amont, un moteur de recherche est uniquement capable de valider ou non la présence d'une chaîne de caractères dans une page wiki. La réponse à des questions comme "*Quelles sont les entreprises françaises s'intéressant au domaine des énergies renouvelables*" ou "*Lister toutes les entreprises présentes dans ce wiki*" est ainsi impossible. Le wiki ne modélise en effet que des documents textuels et des liens hypertextes et non pas par des entités typées liées par des liens eux aussi typés, d'où une différence de représentation entre ce qui est stocké au sein de l'outil et l'interprétation que le lecteur en fait.

2.2.3 Tags et recherche d'information

Un autre écueil des systèmes Web 2.0, notamment dans notre contexte, est du à l'utilisation abondante de tags pour annoter les différents contenus produits. Si les avantages des tags sont multiples en termes d'annotation, l'utilisateur pouvant notamment adapter les termes à ses souhaits particuliers – appelés également lignes de désir (*desire lines*²²) – cette ouverture complexifie la recherche d'information. [Mathes, 2004] estime ainsi qu'"*une folksonomie représente simultanément une partie du pire et du meilleur dans l'organisation de l'information*". En effet, contrairement à des systèmes de classification modélisant une vision

²²<http://www.adaptivepath.com/publications/essays/archives/000361.php>

hiérarchique du monde, comme par exemple le système décimal de Dewey²³ ou la classification scientifique proposée par l'ACM²⁴, une folksonomie n'est qu'un amas de tags chaotiques et non organisés. Il devient ainsi difficile d'accéder à l'informations si l'on ne se réfère pas directement au tag souhaité et il est encore plus complexe d'étendre ou de spécifier sa recherche. Ainsi, certains pensent que si le gain de temps est considérable en termes de publication, il est perdu en termes de recherche d'information et que la pratique de *tagging* perd ainsi de son intérêt²⁵. Nous allons maintenant détailler spécifiquement certaines problématiques liées à ces pratiques telles que nous avons pu les constater au sein du projet et également soulevées par [Mathes, 2004] ou [Golder et Huberman, 2006].

Problèmes d'ambiguïté

Un tag peut en effet être associé à plusieurs significations. Par exemple le mot-clé pac peut correspondre à *pile à combustible*, *politique agricole commune* ou encore *pompe à chaleur* selon le contexte de l'annotation et le contenu annoté. Une recherche sur ce terme récupérera cependant les contenus annotés par le mot-clé quelque soit son sens, induisant un problème de bruit. Les mots-clés ne portent en effet pas suffisamment de sémantique pour définir par eux-mêmes et sans ambiguïté l'entité qu'ils représentent.



Figure 2.6: Résultats d'une recherche associée au tag `apple` sur Flickr

En pratique nous n'avons pas particulièrement été confrontés à ce problème dans notre contexte (Section 4.4, page 182). Il nous est apparu cependant plus fréquent sur le Web. Par exemple, une recherche sur les contenus tagués *apple* sur Flickr identifie aussi bien des photos de fruits que de produits Apple, comme le montre la figure qui suit (Figure 2.6, page 64).

²³<http://www.oclc.org/dewey/>

²⁴ <http://www.acm.org/about/class/>

²⁵http://blogs.talis.com/panlibus/archives/2005/09/why_tagging_is_.php

Problèmes d'hétérogénéité

Si un tag peut avoir plusieurs significations, il est également possible que plusieurs tags soient utilisés pour représenter la même entité. C'est là toute l'ambiguïté des systèmes de tags et du choix de ces termes par les utilisateurs eux-mêmes. Cette hétérogénéité est souvent morphologique ou morphosyntaxique (synonymes, pluriels, variations de casse, multilinguisme ...). Par exemple les tags EDF, ElectriciteDeFrance et électricité de france identifient la même entreprise. Si des systèmes de suggestion ou d'autocomplétion peuvent permettre de restreindre cette hétérogénéité, il arrive cependant qu'elle soit motivée par des raisons liées à des choix plus personnels (on trouve par exemple sur Delicious un certain nombre de tags débutant par _ permettant leur placement en début de liste alphabétique).

Nous avons constaté cette hétérogénéité à plusieurs reprises dans notre contexte applicatif. En analysant notre système, nous avons en effet pu nous rendre compte que certains concepts étaient associés à plus de cinq tags différents et que certains utilisateurs employaient eux-mêmes plusieurs tags pour se référer à un même concept (Section 4.4, page 182).

À nouveau, on retrouve abondamment ce problème sur le Web. Pour exemple, nous avons observé que plus de dix tags distincts sont utilisés pour le concept de Web Sémantique sur Delicious, comme le montre le tableau qui suit (Tableau 2.5, page 65) et ce sans prendre en compte les termes connexes (e.g. SPARQL, RDFa, etc.), sujet que nous allons maintenant évoquer.

Tag	URL de la page associée
semanticweb	http://delicious.com/tag/semanticweb
semantic-web	http://delicious.com/tag/semantic-web
semaweb	http://delicious.com/tag/semaweb
semweb	http://delicious.com/tag/semaweb
websemantic	http://delicious.com/tag/websemantic
web-semantic	http://delicious.com/tag/web-semantic
websemantique	http://delicious.com/tag/websemantique
web-semantique	http://delicious.com/tag/web-semantique
websemantica	http://delicious.com/tag/websemantica
web-semantica	http://delicious.com/tag/web-semantica
websemantico	http://delicious.com/tag/websemantico
web-semantico	http://delicious.com/tag/web-semantico
websem	http://delicious.com/tag/websem

Tableau 2.5: Tags utilisés pour le concept de Web Sémantique sur Delicious

Absence d'organisation

Une dernière limite associée à ces pratiques de *tagging* est l'absence d'organisation entre tags. Une folksonomie n'est en effet qu'un amas de mots-clés désorganisés au sens où au-

cune relation n'est explicitement définie entre les termes utilisés. Ainsi, bien qu'il puisse exister une relation entre les concepts représentés par différents tags, celle-ci n'est prise en compte à aucun moment. Ces systèmes ne sont ainsi pas capables d'identifier la relation qui existe entre les tags **énergie des vagues** et **énergie marine** (ou plutôt entre les concepts correspondants) et en conséquence de prendre en compte cette relation au niveau de la recherche d'information et de la navigation. À nouveau, cette absence d'organisation est liée au manque de sémantique qui existe dans des organisations comme les folksonomies.

Si certaines méthodes statistiques permettent de pallier à ce manque d'organisation, nous allons maintenant montrer en quoi celles-ci sont limitées, notamment dans un contexte où le niveau d'expertise des différents utilisateurs est relativement hétérogène.

Approches de clustering et limites de celles-ci dans notre contexte

Pour pallier à ces limitations, des méthodes classiques de *clustering* peuvent être utilisées afin d'identifier des ensembles de tags proches ou similaires [Begelman *et al.*, 2006]. En se basant sur des stratégies de cooccurrence, on peut suggérer des tags à partir d'un tag particulier afin d'enrichir les possibilités de recherche d'information, comme le propose par exemple Delicious avec une liste de *related tags* (Figure 2.7, page 66).

Figure 2.7: Tags suggérés par cooccurrence sur Delicious

Il nous est cependant apparu que ces stratégies étaient difficilement applicables dans certains contextes, notamment dans notre système de *tagging* d'entreprise. En effet, une analyse plus complète de notre folksonomie, reposant sur un ensemble de 12257 tags utilisés au sein de 21614 billets de blog, nous a conduits à des résultats intéressants à ce sujet. Comme le montre la figure qui suit (Figure 2.8, page 67) et les statistiques associées (Tableau 2.6, page 67), la distribution de nos tags au sein de la folksonomie suit une distribution de Pareto²⁶ : un très grand nombre de tags sont utilisés très peu de fois. On voit par exemple que 68% d'entre eux sont utilisés au maximum deux fois, et seulement 10% plus de dix fois. Comme étudié par [Hayes *et al.*, 2007], ce type de distribution rend difficile l'application des

²⁶Egalement connue dans le monde Web 2.0 sous l'appellation de *long tail*. – <http://www.wired.com/wired/archive/12.10/tail.html>

méthodes de *clustering*, à moins de combiner celles-ci avec d'autres techniques, par exemple prendre en compte le contenu associé aux tags.

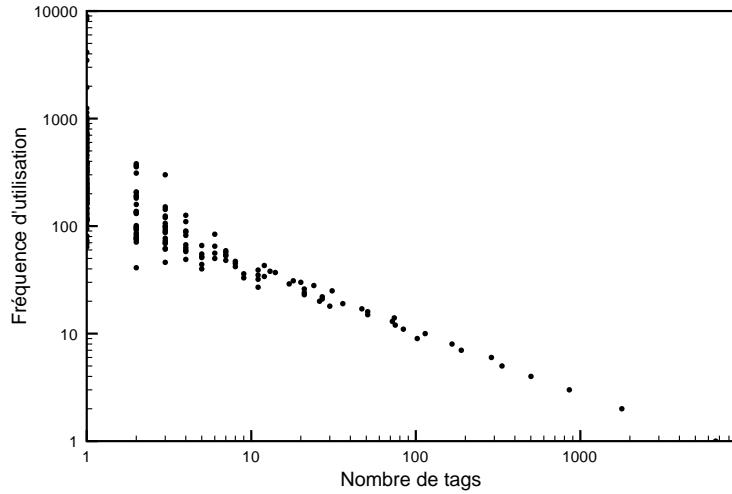


Figure 2.8: Distribution des tags au sein de notre folksonomie

Fréquence f	Nombre de tags	Pourcentage utilisé		
		f fois	f fois ou moins	$f + 1$ fois ou plus
1	6643	54.2	54.2	45.8
2	1787	14.58	68.78	31.22
3	857	6.99	75.77	24.23
4	501	4.09	79.86	20.14
5	334	2.72	82.58	17.42
6	288	2.35	84.93	15.07
7	189	1.54	86.47	13.53
8	166	1.35	87.83	12.17
9	102	0.83	88.66	11.34
10	114	0.93	89.59	10.41

Tableau 2.6: Distribution des tags au sein de la plate-forme Hermès

Cette analyse nous a également permis de constater que le niveau d'expertise des utilisateurs sur un domaine donné influait fortement sur la manière d'utiliser les tags. Par exemple, les experts en énergie solaire utilisent des tags tels que TF²⁷, alors que les non-experts vont utiliser des termes génériques comme *solaire*. Un problème particulier associé à ces différentes manières d'annoter les contenus est que les experts n'utilisent pas toujours les termes

²⁷Acronyme pour *Thin Film*, un type particulier de cellule solaire.

génériques, car évidents ou trop spécifiques pour eux. Il existe en effet une différence du niveau de base (*basic level*) pour un domaine donné entre experts et non-experts, comme l'ont souligné [Tanaka et Taylor, 1991], celle-ci se ressentant dans les principes d'annotation par tags. À ce sujet, [Golder et Huberman, 2006] ont d'ailleurs remarqué des comportements similaires au sein de Delicious. D'un point de vue de la distribution des tags associés et de leur cooccurrence, ceci conduit à un lien très faible entre le tag générique et les différents tags spécifiques associés. Nous avons ainsi constaté que seulement 1% des 194 billets tagués TF étaient également taggués *solaire*, alors que moins de 0.5% des 704 billets tagués *solaire* le sont avec TF. Ce faible rapport de cooccurrence rend à nouveau les algorithmes de *clustering* difficilement applicables pour identifier une similarité entre ces tags, comme l'ont montré [Begelman *et al.*, 2006]. En effet, la rapport entre tags est trop faible pour être pris en compte par de tels algorithmes, à moins de diminuer leur seuil d'acceptation, les rendant peu pertinents puisque suggérant alors un nombre de tags beaucoup trop élevé. En conséquence, les systèmes ne seront pas capable d'identifier certains tags comme proches bien qu'il soit évident qu'il existe un lien fort entre les concepts associés. Ceci complexifie d'autant plus la recherche de contenus annotés dès lors que l'utilisateur n'explicite pas le tag exact.

2.2.4 Synthèse des problèmes rencontrés

Nous pouvons ainsi synthétiser les différents problèmes rencontrés par rapport à la vision de l'Entreprise 2.0 définie par *SLATES* de la manière suivante (Tableau 2.7, page 68) :

Règle	Problème
<i>Search</i>	Pas de prise en compte des problèmes d'ambiguïté et d'hétérogénéité, information fragmentée, difficulté d'identifier les sources
<i>Link</i>	Production de lien hypertextes entre documents et non pas de relations typées entre les concepts qu'ils représentent
<i>Authoring</i>	Production de documents et non pas des concepts associés
<i>Tags</i>	Ambiguïté, hétérogénéité et absence d'organisation
<i>Extension</i>	Extension possible uniquement sur des méthodes statistiques ou de co-occurrences, limitées pour les raisons évoquées plus haut
<i>Signals</i>	Difficulté de suivi de l'information du à l'abondance de nouvelles issues de flux RSS

Tableau 2.7: Problématiques soulevés par l'approche *SLATES* classique au sein d'Hermès

Si les limites mentionnées s'appliquent à chacun des outils pris individuellement sur le Web (blogs, wikis ou agrégateur RSS), elles sont d'autant plus problématiques dans un contexte d'entreprise. En effet, un accès efficace à l'information est un prérequis dans un environnement tel que celui-ci. Les limites évoquées sont ainsi particulièrement problématique, dans le sens où l'utilisation de ces outils accentue la publication et le partage d'informations de valeur, mais ne permet pas de les identifier et les réutiliser de manière optimale. Il nous semble ainsi que l'analyse de [Mathes, 2004] au sujet des systèmes de tags peut s'appliquer à l'ensemble des applications Web 2.0. On peut considérer que si les outils clas-

siques de l'Entreprise 2.0 facilitent la publication d'information, la recherche peut s'avérer au contraire très complexe. A nouveau, cette complexité est proportionnelle au nombre de documents créés et d'outils utilisés.

2.3 ÉCOSSYSTEME SÉMANTIQUE POUR L'ENTREPRISE 2.0

2.3.1 Web Sémantique et méthodologie *SemSLATES*

Afin de répondre efficacement aux problématiques posées dans la section précédente, nous proposons d'appliquer les technologies du Web Sémantique (Section 1.1, page 12) à de tels systèmes d'information d'Entreprise 2.0. Notre proposition étend ainsi la vision classique des systèmes d'Entreprise 2.0 en se concentrant sur la modélisation d'annotations sémantiques associées à de telles architectures, proposant la mise en place d'un écosystème sémantique pour l'Entreprise 2.0 en support de l'existant, à la manière de ce que [Gandon, 2002] considère comme des *semantic intrawebs*.

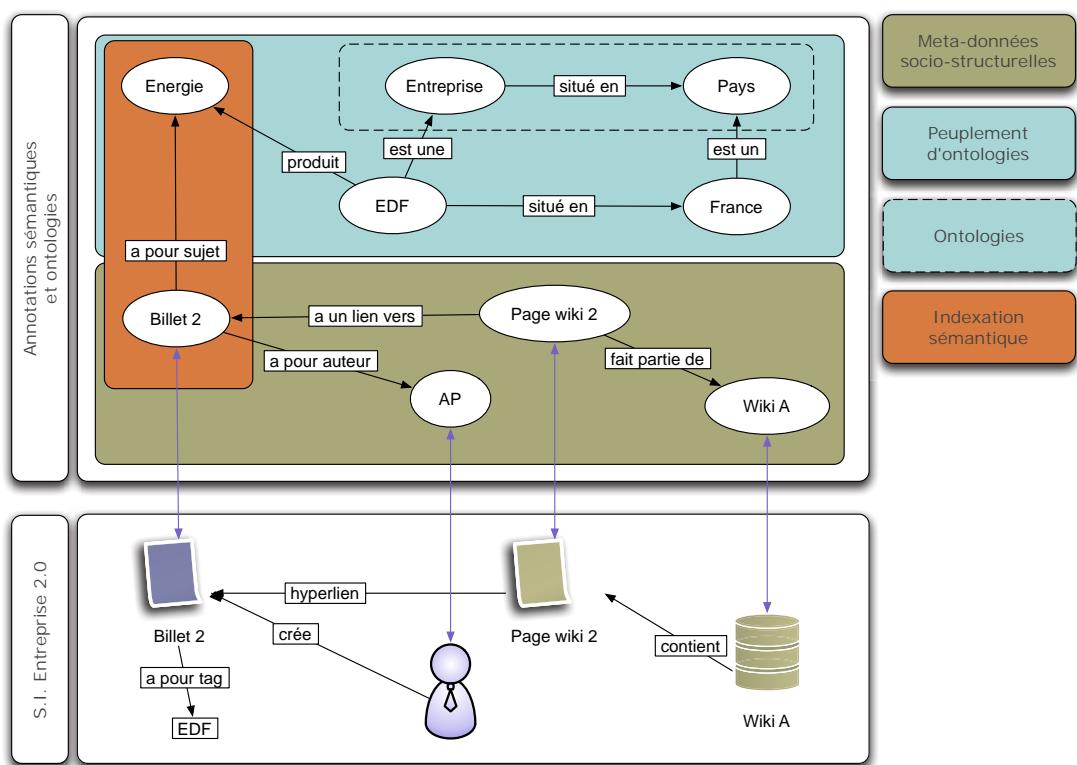


Figure 2.9: Annotations sémantiques en support d'un système d'Entreprise 2.0 existant selon trois niveaux d'annotations

Plus précisément, ces annotations se font avec trois objectifs complémentaires en tête (Figure 2.9, page 69) :

Règle	<i>SLATES</i>	<i>SemSLATES</i>
<i>Search</i>	Recherche plein-texte et/ou par tags	Recherche sémantique, <i>i.e.</i> par concepts
<i>Link</i>	Liens entre documents	Relations typées entre concepts
<i>Authoring</i>	Publication de contenus	Publication d'annotations sémantiques
<i>Tags</i>	Annotation de contenus par tags	Indexation sémantique avec des ontologies de domaine
<i>Extension</i>	Extension par hyperliens et systèmes de tags	Extension par parcours du graphe de connaissances induit par les annotations
<i>Signals</i>	Suivi de nouvelles par flux RSS	Indexation sémantique de flux RSS et création de flux dédiés

Tableau 2.8: Fonctionnalités comparées de *SLATES* et *SemSLATES*

- la modélisation de *métadonnées socio-structurelles* associées aux différents outils, *i.e.* la représentation des différentes activités établies au sein de ceux-ci. Ces annotations vont ainsi permettre de représenter qu'un billet de blog a été créé par tel auteur ou qu'une page wiki fait partie de tel wiki ;
- le *peuplement d'ontologies*, *i.e.* la création et le maintien d'instances et des propriétés associées. Ces annotations vont ainsi être utilisées pour modéliser des assertions comme le fait qu'EDF est une entreprise française ;
- l'*indexation sémantique* de contenu, *i.e.* l'indexation de documents avec des concepts d'ontologies, en pratique des instances d'ontologies de domaine. Ces annotations permettent donc de modéliser qu'un billet de blog a pour sujet EDF, identifié non pas comme simple chaîne de caractère mais comme instance d'une classe Entreprise, avec une URI propre.

Cet écosystème sémantique nous permet ainsi d'envisager *SemSLATES*, extension de *SLATES* basé sur les technologies du Web Sémantique et que nous définissons comme suit (Tableau 2.8, page 70).

Comme le montre la figure précédente (Figure 2.9, page 69), il est nécessaire de considérer ce niveau de représentation comme une *extension* venant en support de l'existant et non pas comme un système annexe, tout comme l'est le Web Sémantique par rapport au Web. De ce fait, deux points importants sont à retenir :

- l'utilisation des outils d'origine (blogs, wikis, agrégateur RSS) pour permettre la production des annotations sémantiques, sans pour autant complexifier leur utilisation. La pratique nous ayant montré que la simplicité des différents outils contribuait à leur réussite, conserver celle-ci est un prérequis à la réussite de notre proposition ;
- le rôle central joué par l'utilisateur final, par extension du point précédent. Les différentes annotations sont en effet le produit de la participation volontaire des utilisateurs à ces outils et plus généralement des interactions sociales qui en découlent. Le rôle de ces utilisateurs est en réalité double, puisque (1) d'une part notre système per-

met la représentation des actions utilisateurs (métadonnées socio-structurelles) et que (2) d'autre part les différentes annotations produites (pour le peuplement d'ontologies et l'indexation sémantique) sont le fait de ces interactions sociales.

2.3.2 Dé nition d'une architecture sociale de médiation sémantique

L'implémentation de l'écosystème précédent se traduit par la mise en place d'une architecture de médiation en complément du système d'information initial, permettant d'interconnecter et d'enrichir les différentes applications d'origine [Passant, 2008a].

[Rousset *et al.*, 2002] donne la définition suivante d'un médiateur : "*Un médiateur joue un rôle d'interface de requêtes entre un utilisateur et des sources de données. Il donne à l'utilisateur l'illusion d'interroger un système homogène et centralisé en lui évitant d'avoir à trouver les sources de données pertinentes pour sa requête, de les interroger une à une, et de combiner lui-même les informations obtenues*". S'il s'agit bien d'un prérequis aux objectifs que nous visons, notre implémentation diffère quelque peu de cette définition classique de médiation [Wiederhold, 1992]. Comme [Rousset *et al.*, 2002] le précise, les architectures de médiation ont généralement pour objectif de proposer des méthodes pour unifier les requêtes au dessus de sources de données hétérogènes et réparties. Ceci s'effectue via un système de distribution des requêtes puis de recomposition des résultats à partir de vues proposées par les outils sources. À l'opposé, notre approche consiste non pas à décomposer les requêtes pour interroger les différentes sources de données mais au contraire à modéliser les sources selon un ensemble d'ontologies prédéfinies en fonction d'annotations RDF associées. Ces graphes d'annotations sont ensuite immédiatement stockés au sein d'un entrepôt de données associé au médiateur, faisant de notre approche un modèle hybride entre les systèmes de médiation et les *datawarehouse* à la manière de Xylème [Xyleme, 2001]. Nous discuterons ce choix architectural en détail dans la suite du mémoire (Section 5.1, page 186), motivé essentiellement pour des raisons de performance devant le besoin réel de fournir aux utilisateurs des réponses rapides à leurs requêtes. Notons que nous emploierons par la suite simplement le terme *d'architecture de médiation* pour définir notre proposition, et considérons le système de stockage comme faisant partie intégrante de celle-ci. Malgré cette structure hybride, notre système conserve les différents niveaux d'une architecture de médiation, à savoir :

- des *sources* de données, *i.e.* les différents outils du système d'origine auxquels viennent se greffer différents *adaptateurs*, *plug-ins* permettant la production aisée d'annotations sémantiques à partir de ceux-ci ;
- un *médiateur* intégrant (1) les données RDF produites par ces différents *adaptateurs* et (2) les ontologies utilisées pour modéliser ces données, intégrant donc le système de stockage évoqué auparavant ;
- des *services* additionnels venant s'y greffer et permettant à l'utilisateur d'effectuer différentes requêtes et de naviguer simplement au sein des données du *médiateur*, *i.e.* de considérer de manière unifiée les sources hétérogènes d'origines.

De plus, de la même manière qu'un médiateur l'autorise, l'ajout d'une nouvelle source de donnée implique uniquement la mise en place d'un nouvel adaptateur pour celle-ci. Il est également important de noter que si le médiateur lui-même repose sur l'intégration de graphes d'annotations sémantiques, modélisées en RDF, les différents services additionnels

masquent totalement cette complexité. Nous utilisons notamment des interfaces issues des principes Web 2.0 pour représenter les données ainsi agrégées, proposant de cette manière une double complémentarité entre Web 2.0 et Web Sémantique (Figure 2.10, page 72) :

- d'une part, les données du médiateur modélisées selon les principes du Web Sémantique sont produites à partir des différents outils initiaux et des comportements utilisateurs. C'est en ce sens que nous parlons de médiation sociale, ces comportements étant également modélisés dans notre architecture de médiation ;
- d'autre part, les annotations peuvent être visualisées par l'intermédiaire d'outils simples, masquant la complexité de celles-ci à l'utilisateur et notamment inspirés de certains concepts introduits par le Web 2.0, comme la notion de *mash-ups* sémantiques.

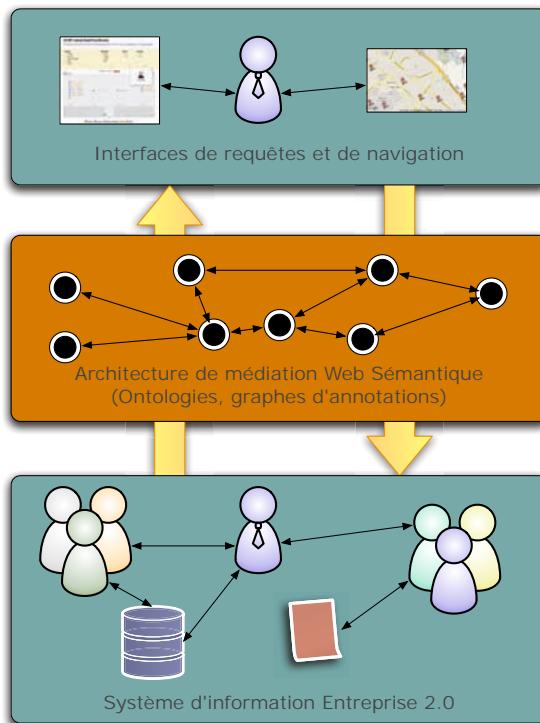


Figure 2.10: Architecture de médiation sémantique pour l'Entreprise 2.0

Comme nous l'avons indiqué, notre médiateur (M) repose sur un ensemble d'annotations sémantiques issues du système d'information initial (SI) et modélisées en RDF selon un ensemble d'ontologies RDF(S)/OWL. Nous modélisons ainsi les différents éléments pris en compte par le médiateur de la manière suivante :

$$Input(M) = (O, G) \quad (2.1)$$

$$O = \{O_{m(SI)}\} \cup \{O_{d(SI)}\} \quad (2.2)$$

$$G = \{G_{m(SI)}\} \cup \{G_{d(SI)}\} \cup \{G_{d(W)}\} \quad (2.3)$$

où

- $O_{m(SI)}$ est un ensemble d'ontologies relatives à la représentation des métadonnées socio-structurelles du *SI* ;
- $O_{d(SI)}$ est un ensemble d'ontologies relatives à la représentation des données métier évoquées dans les différents contenus du *SI* ;
- $G_{m(SI)}$ est un ensemble de graphes d'annotations RDF (Section 1.1.2, page 16) modélisant les métadonnées socio-structurelles du *SI* ;
- $G_{d(SI)}$ est un ensemble de graphes d'annotations RDF modélisant des données métier présentes dans les documents du système d'information, *i.e.* annotations relatives au peuplement d'ontologies ;
- $G_{d(W)}$ est un ensemble de graphes d'annotations RDF issus du Web et modélisant essentiellement des données métier, pouvant provenir notamment des efforts du projet *Linking Open Data* (Section 1.1.4, page 27).

Le médiateur est donc alimenté ($Input(M)$) par un ensemble d'ontologies (prédéfinies) et de graphes d'annotations sémantiques reposant sur celles-ci (créés depuis les différents adaptateurs). Comme nous l'avons évoqué dans la section précédente, ces annotations peuvent avoir plusieurs rôles : métadonnées socio-structurelles, peuplement d'ontologies et indexation sémantique mais sont toujours créées via les différents outils d'origines et adaptateurs associés (hormis celles issues des graphes $G_{d(W)}$, provenant du Web). Par exemple, comme nous allons maintenant le voir, des graphes d'annotations du type $G_{m(SI)}$ vont être produits à partir d'interactions sur les blogs alors que les wikis vont permettre le peuplement d'instances d'ontologies de domaine et en conséquence la production d'annotations du type $G_{d(SI)}$. Notons par ailleurs, pour des raisons de suivi de l'information et de traçabilité de celle-ci au sein du médiateur que nous détaillerons quand nous aborderons les wikis sémantiques (Section 4.2.1, page 148), les graphes d'annotations métier issus des outils internes sont liés aux graphes de métadonnées socio-structurelles et ne peuvent s'intégrer seuls au médiateur.

2.3.3 Modèles, adaptateurs et services

Production des métadonnées socio-structurelles à partir des outils d'origine

Notre premier besoin pour la réalisation de cet écosystème sémantique consiste en la définition d'un ensemble d'ontologies permettant une représentation unifiée des métadonnées socio-structurelles du *SI*. Nous souhaitons que de tels modèles permettent de représenter :

- les documents mais aussi les outils eux-mêmes en tant que conteneurs de données (un blog, un wiki donné) ;
- les utilisateurs en tant qu'entités virtuelles représentées dans le système (et non pas directement les personnes physiques) ;
- les liens entre ces différents composants permettant de prendre en compte la composante sociale évoquée ci-dessus.

C'est en raison de cette combinaison entre activités sociales et structures des différents outils et documents que nous utilisons l'appellation de métadonnées socio-structurelles. De tels modèles ($\{O_{m(SI)}\}$) vont ainsi venir en support de la production des graphes d'annotations associées ($\{G_{m(SI)}\}$) permettant de résoudre en partie le problème d'hétérogénéité des

sources d'information (Section 2.2.1, page 62). Cette sémantique commune permet à terme d'interroger les outils de manière unifiée, réduisant la problématique de fragmentation.

Pour satisfaire ces différents besoins, nous avons participé à la définition de l'ontologie SIOC – Semantically-Interlinked Online Communities [Breslin *et al.*, 2005] – que nous détaillerons par la suite (Section 3.1, page 84). SIOC offre un modèle destiné à la représentation des activités des communautés en ligne via une ontologie légère et modulaire . Ce modèle se compose d'un noyau et de différents modules dont un module Types permettant de définir de manière assez fine les différents objets manipulés dans le contexte du Web 2.0 (blog et billets, wiki et pages wiki...). SIOC réutilise également des vocabulaires existants et populaires (DublinCore, FOAF ...) pour définir certaines propriétés, évitant ainsi de redéfinir des besoins déjà satisfait par des modèles existants. La production de données RDF modélisées avec SIOC se fait de manière automatique depuis les différents outils mis en place dans le système d'information d'origine. Par le biais de différents adaptateurs, sous la forme d'exporteurs ou de traducteurs, ces métadonnées sont ainsi produites sans aucune intervention utilisateur, de manière totalement transparente. Nous détaillerons dans les chapitres qui suivent les différents outils nécessaires pour permettre cette traduction dans notre contexte mais aussi de manière générale sur le Web (Section 4, page 137). SIOC est en effet aujourd'hui utilisé et recommandé dans un grand nombre de services combinant principes du Web Sémantique et du Web 2.0.

La figure suivante illustre la modélisation uniforme de différentes sources de données hétérogènes grâce à SIOC (Figure 2.11, page 75). Elle montre ainsi l'intérêt d'une sémantique commune, les instances des classes représentant les documents héritant toutes de `sioc:Item` et utilisant la même propriété `sioc:has_container` pour les rattacher à leur conteneur. On obtient ainsi un modèle homogène, tout en conservant la spécificité de chacun des contenus grâce à l'utilisation du module Types. Cette unification se traduit par un ensemble d'annotations RDF similaires quelque soit l'outil d'origine et permet donc l'utilisation de requêtes SPARQL uniformes. Le système bénéficie ainsi d'un premier niveau de sémantique commune pour notre architecture de médiation, permettant par exemple d'identifier tous les contenus créés il y a plus de dix jours et ce quelque soit l'outil d'origine.

Capitalisation des connaissances et peuplement d'ontologies

Alors que le point précédent s'intéresse essentiellement à la structure des différents outils, notre second besoin concerne le contenu même des documents, dans un objectif de capitalisation des connaissances. C'est ici qu'intervient le second niveau de sémantique nécessaire à notre architecture, comprenant un ensemble d'ontologies de domaine ($\{O_{d(SI)}\}$) et les graphes d'annotations associés ($\{G_{d(SI)}\}$). Comme nous l'avons évoqué, la méthodologie *SemSLATES* repose fortement sur le rôle des utilisateurs dans ce contexte de médiation sémantique et sociale. Ainsi, si ce processus se rapproche du peuplement d'ontologies, nous avons fait en sorte de l'associer aux comportements des utilisateurs à travers les outils initiaux. Notre proposition en termes de capitalisation des connaissances pour l'Entreprise 2.0 repose donc sur l'utilisation de wikis sémantiques en tant qu'interfaces de peuplement d'ontologies de domaine (Section 4.2.1, page 148). Cette proposition permet ainsi de bénéficier des principes de la philosophie wiki (ouverture, collaboration ...) pour peupler une ou plu-

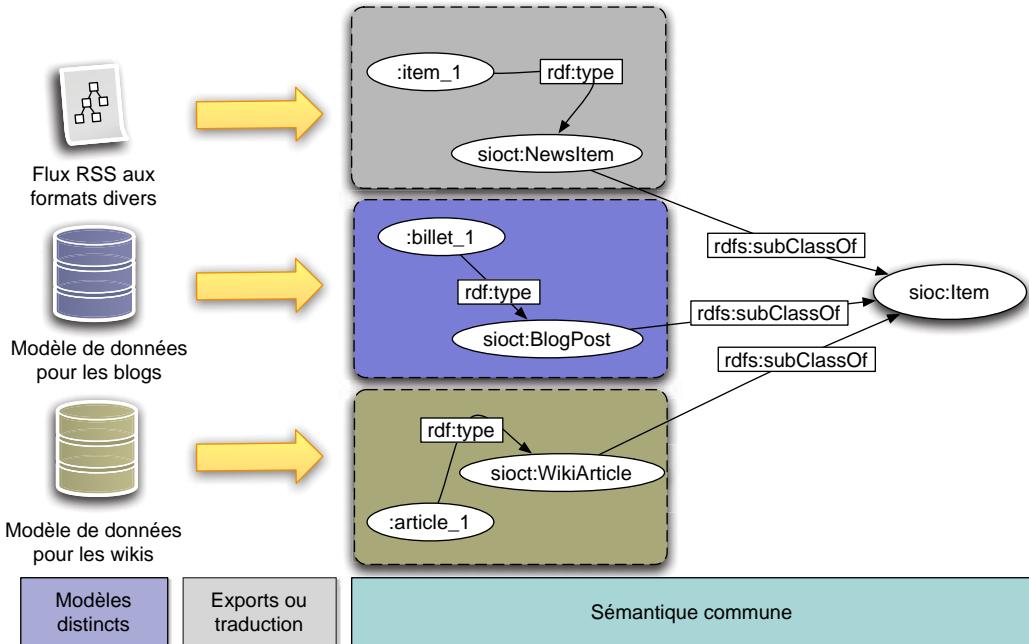


Figure 2.11: Représentation unifiée des métadonnées documentaires avec SIOC

sieurs ontologies de domaine, tout en masquant la complexité du processus aux utilisateurs.

Dans notre contexte applicatif, ce processus se traduit par un enrichissement de la plate-forme de wiki avec un système d'annotations guidées par un ensemble de formulaires reposant sur des ontologies de domaine, sous le nom d'UfoWiki (Section 4.2.2, page 154). Ces ontologies étant par nature dépendantes des besoins de modélisation, nous nous sommes concentrés, du fait de notre contexte industriel, sur des ontologies permettant de modéliser différents acteurs et technologies associées (Section 3.2, page 104). L'exemple suivant représente ainsi la modélisation de connaissances établies au sujet d'EDF produite via cet adaptateur. À partir d'une page wiki, on modélise ainsi en RDF le fait qu'il s'agit d'une entreprise évoluant en France dans le domaine de la production d'énergie nucléaire (Listing 2.1, page 75).

```

@prefix athena: <http://athena-project.eu/ontology#> .
@prefix foafplus: <http://purl.org/foafplus/ns#> .
@prefix role: <http://purl.org/role/ns#> .
@prefix geonames: <http://www.geonames.org/ontology#> .

@prefix athena:EDF a foafplus:Company ;
  role:hasRole [
    role:roleDomain athena:Nucleaire ;
    role:roleType athena:Production ;
    geonames:locatedIn <http://dbpedia.org/resource/France> .
  ] .

```

Listing 2.1: Représentation d'assertions au sujet d'EDF

Tags et indexation sémantique

Si les systèmes à base de tags souffrent de nombreux défauts en termes de recherche de documents annotés (Section 2.2.3, page 63), les processus d'indexation sémantique permettent selon nous d'y répondre efficacement. En associant ces documents non pas à de simples termes linguistiques (tags) mais à des concepts ou instances d'ontologies identifiés (via leur URI) de manière universelle et non ambiguë, les différents problèmes recensés précédemment peuvent être résolus :

- via l'utilisation d'URIs pour annoter les contenus, les problèmes d'ambiguïté et d'hétérogénéité sont résolus. Ces URIs sont en effet non-ambigus par rapport aux concepts qu'elles identifient et font abstraction du terme. Par exemple, pour identifier le terme Web Sémantique, nous pouvons utiliser l'unique URI dbpedia:Semantic_Web en lieu et place de différents termes ;
- la découverte de contenus proches peut se faire en utilisant les différentes relations entre instances, palliant ainsi aux différentes limites que nous avons évoquées quant aux méthodes statistiques poursuivant un but similaire via l'analyse de cooccurrence de tags. Par exemple, puisqu'il existe au sein de DBpedia une relation directe entre les URIs dbpedia:Category:Semantic_Web et dbpedia:RDFa, les documents au sujet de RDFa pourront être retrouvés à partir de ceux relatifs au Web Sémantique.

Cependant, il existe en général une marge assez large entre ces deux méthodes d'indexation documentaire. Alors que la première repose sur l'association de simples mots-clés aux documents à annoter, sans prérequis technique ni connaissance d'un vocabulaire prédéfini, la seconde fait appel à des connaissances plus poussées à la fois en termes de connaissance du (ou des) vocabulaire(s) disponible(s) pour l'indexation et en termes d'ingénierie des connaissances et de représentation des annotations sémantiques associées.

Pour ce faire, nous avons ainsi défini le modèle MOAT – *Meaning Of A Tag* –, dont l'objectif est de proposer une approche mixte entre la simplicité d'utilisation des tags et la complexité – mais la puissance – de l'indexation sémantique [Passant et Laublet, 2008b]. L'approche défendue au sein de notre méthodologie consiste ainsi à fournir aux utilisateurs un moyen simple de franchir le pas qui sépare ces deux principes d'indexation. Cette approche permet aux utilisateurs d'assigner à chaque tag utilisé sa signification correspondante, en utilisant une ressource du Web Sémantique. Nous entendons donc ici par signification d'un tag le sens qu'il porte, *i.e.* le concept auquel il se réfère.

Par exemple, ce modèle permet de représenter que dans un contexte particulier, le tag *apple* est utilisé pour représenter le fruit (identifié par dbpedia:Apple) mais qu'il est utilisé dans un autre contexte pour représenter l'entreprise informatique (dbpedia:Apple_Inc.). En particulier, dans ce contexte d'Entreprise 2.0, notre approche permet d'utiliser non pas uniquement des URIs de concepts disponibles sur le Web (via DBpedia par exemple) mais aussi des instances d'ontologies de domaine peuplées en interne, notamment via les wikis. Cette approche repose sur une ontologie légère (Section 3.3.3, page 128) et sur un processus participatif permettant de partager ces différentes significations au sein d'une communauté (Section 4.3, page 170). Le tout se fait de manière la plus transparente possible pour l'utilisateur de manière à rester proche de la simplicité des systèmes à base de tags.

Exploitation des données via de nouveaux services

Bien que produites via des outils distincts, les différentes annotations sémantiques s'organisent sous la forme d'un unique graphe d'annotations issues des différents outils du système initial d'Entreprise 2.0 et comprenant annotations socio-structurelles, données métier (interconnectées aux précédentes notamment via MOAT) et ontologies. Afin d'exploiter au mieux ces différents niveaux de représentation, interconnectés au sein de cet écosystème sémantique venant se greffer sur notre plate-forme d'origine, différentes applications peuvent-être envisagées. En effet, à partir du moment où nous disposons de graphes d'annotations RDF interconnectées, ceux-ci peuvent être manipulés de diverses manières. Ceci nous semble un point important à considérer dans l'approche du Web Sémantique : considérer les annotations RDF comme élément fondamental de l'approche, les applications n'étant ainsi que des systèmes de visualisation adaptables au dessus de ces données, comme nous l'avons introduit dans le chapitre précédent (Section 1.1.4, page 27) et le détaillerons par la suite (Section 5.4.3, page 221). Dans notre contexte, les services permettant de répondre à ces divers points peuvent donc être de nature assez diverses : indexation automatique de flux RSS entrants, moteur de recherche sémantique, *mash-ups*, navigation à facettes ... Par exemple, nous avons mis en place un service de *mash-up* sémantique combinant données internes et des données publiques proposées sur le Web par le projet Geonames²⁸ afin de géolocaliser les différentes instances d'ontologies de domaine produites par nos wikis (Section 5.2.3, page 206). Ce service met ainsi en valeur l'intégration de données publiques au sein d'une plate-forme d'entreprise, intégration rendue possible via l'utilisation de formats communs entre le Web et le système d'information interne (Section 4.2.4, page 163).

Nous présenterons en détail dans le dernier chapitre de ce mémoire les différents outils mis en place dans notre contexte et la manière dont ils offrent de nouvelles manières de visualiser et accéder à l'information dans ce contexte d'Entreprise 2.0 ainsi que les principes architecturaux associés (Section 5, page 185).

2.3.4 Situation de l'approche vis-à-vis de l'état de l'art

Alors que nous reviendrons dans les chapitres suivants sur des aspects particuliers de nos travaux et leur situation par rapport à l'état de l'art (SIOC, MOAT, UfoWiki), il nous semble pertinent de positionner l'approche *SemSLATES* dans son ensemble.

Nous pouvons tout d'abord situer celui-ci par rapport aux architectures de médiation basées sur les principes du Web Sémantique. [Wiederhold, 1992] justifie le besoin d'architectures de médiation en raison de la surcharge d'information, problème que nous avons également exposé dans ce chapitre²⁹ : "Without smart software we will gain access to more data but not improve access to the type and quality of information needed for decision making". En proposant des formats de structuration et d'échange de données standardisés avec RDF(S)/OWL, les technologies du Web Sémantique sont particulièrement adaptées pour la mise en place de tels systèmes de médiation et de gestion de l'information reposant sur des ontologies. De nombreux travaux ont ainsi été proposés dans cette direction, parmi lesquels :

²⁸<http://geonames.org>

²⁹Il est par ailleurs intéressant de constater que cette problématique datant d'une quinzaine d'années est toujours présente et s'est même accentuée avec l'explosion du Web et des modes de publications Web 2.0.

- PICSEL [Rousset *et al.*, 2002] qui s'attache notamment à l'intégration de sources de données dans le domaine du tourisme. Il repose sur l'utilisation de différentes ontologies (modélisées en CARIN-*ALN*) pour permettre l'intégration de sources de données distribuées et hétérogènes de manière transparente pour l'utilisateur ;
- Ontobroker [Decker *et al.*, 1999], l'un des premiers systèmes d'intégration de données reposant sur des principes précurseurs au Web Sémantique. Celui-ci est aujourd'hui commercialisé par la société Ontoprise³⁰ et repose sur les formalismes RDF(S)/OWL mais aussi \mathcal{F} logic). Il s'intéresse particulièrement à l'intégration de bases de données et dispose d'adaptateurs pour les principales solutions du marché ;
- SCORE – *Semantic Content Organization and Retrieval Engine* [Sheth *et al.*, 2002] – qui s'intéresse également à l'intégration de sources de données hétérogènes à l'aide d'ontologies. Dans cette approche, la phase d'extraction de connaissances et leur normalisation depuis les différentes sources de données joue un rôle majeur pour permettre la mise en place de nouveaux services, notamment en termes de recherche d'information.

Plus proche de nos travaux, [Maedche *et al.*, 2003] proposent également une vision des systèmes de gestion de connaissances en entreprise basés sur des ontologies avec OMKS – *Ontology-based Knowledge Management System*. Leur proposition de concentrate notamment sur l'intégration et l'alignement de différentes sources de données internes (bases de données, annuaires ...) via un système central de médiation. Plus particulièrement, une caractéristique de cette approche est la notion d'alignement entre différentes ontologies locales au sein du système de médiation. Les différents cas d'utilisation du Web Sémantique en entreprise recensés par le W3C³¹ regroupent également de nombreux scénarios de médiation reposant sur ces technologies. On les trouve ainsi utilisées pour l'identification de profils d'experts à la NASA³² ou la gestion de données biomédicales chez Eli Lilly³³. Enfin, si notre approche se concentre sur l'utilisation des technologies du Web Sémantique pour le bénéfice de l'utilisateur final, ces techniques de médiation peuvent également être utilisées pour faciliter les échanges directs entre applications dans un processus d'intégration d'applications d'entreprises ou *EAI – Enterprise Architecture Integration*. C'est par exemple ce que proposent [Anicic *et al.*, 2006] avec l'utilisation d'ontologies OWL et de scripts dédiés permettant d'aligner les entrées et sorties XML de différentes applications selon des modèles communs.

Cependant, ces approches ne prennent généralement pas en compte les notions d'utilisateurs et d'interactions sociales dans ces processus de médiation, se focalisant essentiellement sur des données métier provenant de bases de connaissances figées (annuaires, fonds documentaires, etc.). C'est selon nous une des originalités de notre approche, le rôle de l'utilisateur étant pris en compte de deux manières :

- d'une part à travers la prise en compte des interactions sociales auxquelles il participe avec la représentation en RDF de métadonnées socio-structurelles associées aux différents outils et documents créés. Le rôle de l'utilisateur est ainsi pris en compte

³⁰<http://ontoprise.de>

³¹<http://www.w3.org/2001/sw/sweo/>

³²<http://www.w3.org/2001/sw/sweo/public/UseCases/Nasa/>

³³<http://www.w3.org/2001/sw/sweo/public/UseCases/Lilly/>

en termes de comportements sociaux et d'annotations documentaires, principalement via les modèles SIOC et MOAT ;

- d'autre part, son rôle en tant qu'acteur principal du peuplement d'ontologie, via l'utilisation de wikis sémantiques. Alors que les approches classiques de médiation se basent généralement sur des ontologies peuplées par un nombre restreint d'utilisateurs ou reposant sur des bases de connaissances prédéfinies, les ontologies sont ici peuplées par les utilisateurs eux-mêmes, les bases de connaissances évoluant ainsi en fonction de leurs comportements.

C'est en ce sens que nous pouvons qualifier notre approche de système de médiation sociale, l'utilisateur final ayant un rôle important selon deux points de vue, distincts mais non disjoints, dans la mise en place de cet écosystème informationnel.

Bien que non axé sur une médiation de données Web, il nous paraît également important de mentionner ici l'initiative du *Semantic Desktop*, notamment au travers du projet Nepomuk³⁴ [Bernardi *et al.*, 2008]. Celui-ci vise à proposer un système de médiation pour le poste de travail, permettant une interopérabilité entre différentes applications (carnet d'adresses, outils bureautique, client e-mail, etc.) via l'utilisation d'ontologies communes et de systèmes d'extraction d'information combinée à l'annotation manuelle de documents par les utilisateurs. En établissant ainsi une sémantique commune entre les données produites par ces différents outils, qui peut être couplée à certaines interactions sociales (et les représentations RDF associées), cette approche propose ainsi une idée similaire à la nôtre, au niveau du poste de travail et non pas d'un système d'information d'entreprise.

En terme plus général d'architecture et puisque nous ne nous basons pas sur un système de vues et de requêtes distribuées mais sur l'annotation sémantique de sources existantes, le modèle que nous proposons se rapproche de ce que définit [Berners-Lee, 2005a] avec la notion de *RDF Bus* (Figure 2.12, page 80). Cette solution propose la mise en place d'une couche additionnelle de sémantique au dessus d'outils hétérogènes sans pour autant repenser ceux-ci mais via de simples ajouts traduisant les données source en RDF (en utilisant des modèles communs pour leur représentation) pour ensuite utiliser celles-ci avec SPARQL.

Une autre catégorie à considérer dans cet état de l'art est celle des solutions combinant principes du Web 2.0 et du Web Sémantique pour les systèmes d'information d'entreprise. Bien que non dédié spécifiquement aux contextes d'entreprise mais plus généralement à toute communauté en ligne, Openlink DataSpaces [Idehen et Erling, 2008] propose une plate-forme combinant notamment blogs, wikis et systèmes de favoris. Ce système bénéficie de certains de nos travaux, puisqu'il intègre notamment SIOC et MOAT en son sein pour proposer cette architecture sémantique intégrée. Nous pouvons également citer Talis Engage³⁵, plate-forme collaborative basée sur un certain nombre d'ontologies, dont à nouveau SIOC. Plus particulièrement dédié aux structures d'entreprise, notamment les PME, citons enfin le récent projet Européen Organik³⁶ [Bibikas *et al.*, 2008]. Celui-ci vise également à étendre la vision de l'Entreprise 2.0 proposée par SLATES : annotation par concepts, recherche sémantique, etc. Il semble cependant (dans l'état actuel) ne pas prendre en compte

³⁴<http://nepomuk.semanticdesktop.org/>

³⁵<http://talism.com/engage>

³⁶<http://www.organik-project.eu/>

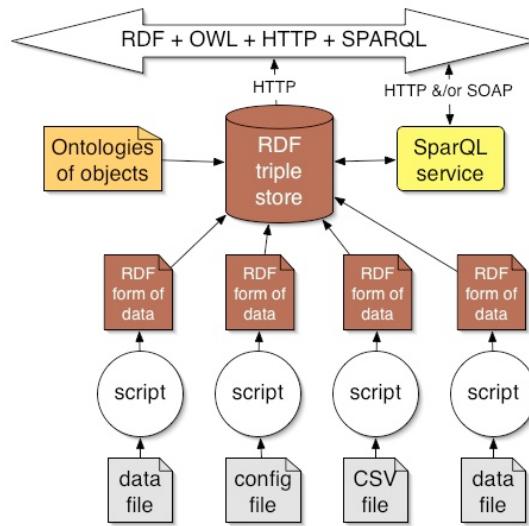


Figure 2.12: Architecture RDF Bus [Berners-Lee, 2005a]

le rôle de l'utilisateur pour le peuplement d'ontologies : contrairement à la vision que nous défendons avec l'utilisation de wikis sémantiques pour permettre ce peuplement d'ontologies par l'utilisateur, l'objectif est ici d'extraire ce type d'annotations avec des algorithmes dédiés.

Les approches pré-citées se basant sur des systèmes monolithiques, leur introduction dans un système déjà en place peut alors se révéler délicate. Bien qu'une migration des données existantes vers ce type de plate-forme soit envisageable, il faut garder à l'esprit le temps nécessaire à l'adoption de tels systèmes par les utilisateurs, comme nous l'avons évoqué plus tôt dans ce chapitre (Section 2.1.4, page 59). Ainsi, basculer vers de nouveaux outils est un risque qu'il est nécessaire d'évaluer, notamment dans des contextes où, comme nous l'avons vu, l'appropriation de tels outils collaboratifs et des principes associés peut prendre du temps.

Enfin, nous pouvons également citer les travaux autour de CoMMA, système également axé sur la notion d'écosystème sémantique pour l'entreprise mais reposant sur une approche différente pour parvenir à cet objectif, *i.e.* sur un système multi-agents [Gandon, 2002]. Celui-ci nous semble pertinent dans la manière où, bien que la prise en compte de l'utilisateur ne soit pas assurée (du moins en termes d'interactions sociales) au niveau de la production d'informations, celui-ci est pris en compte au moment de la diffusion de celle-ci. Les informations publiées au sein de cette mémoire sémantique d'entreprise sont en effet diffusées vers les utilisateurs en fonction des centres d'intérêts de chacun, centres d'intérêts définis via des profils utilisateurs.

CONCLUSION

Dans ce chapitre, nous avons tout d'abord introduit la notion d'Entreprise 2.0, notamment au travers du projet Athéna et de la plate-forme Hermès, mise en place au sein d'EDF R&D pour faciliter les échanges d'information entre ingénieurs et chercheurs. Nous avons présenté en quoi cette plate-forme répondait au paradigme *SLATES* mais restait limitée sur certains points. Nous avons ainsi introduit différents problèmes soulevés par les outils Web 2.0 classiques, à savoir l'hétérogénéité des modèles, l'absence de connaissances interprétables de manière autonome, et les écueils des systèmes à base de tags. Nous avons ensuite présenté en quoi il nous paraissait intéressant d'aller plus loin via l'utilisation d'une couche d'abstraction basée sur les technologies du Web Sémantique et la mise en place d'une architecture de médiation au dessus ces différents outils avec le paradigme *SemSLATES*. Les chapitres suivants détailleront les trois points principaux de notre approche, à savoir la définition des ontologies utilisées au sein de cette architecture, le fonctionnement de chacun des adaptateurs permettant la production de données homogènes et interprétables et enfin l'utilisation que nous faisons de ces données via différents services qui viennent se greffer au médiateur.

Notons pour finir que si notre approche est ici présentée dans un contexte fortement orienté entreprise, elle peut s'appliquer selon nous à toute communauté en ligne utilisant un ensemble d'outils Web 2.0 et souhaitant aller au-delà des fonctionnalités proposées traditionnellement par ceux-ci. Ainsi, cette vision *SemSLATES* nous semble aller plus loin que le contexte d'Entreprise 2.0 au sens où elle peut s'adapter à tout écosystème social d'utilisateurs et d'outils centré autour d'intérêts communs. De plus, via l'utilisation de modèles et d'URIs communs, il est également possible d'envisager une complémentarité entre différentes communautés de ce type pour parvenir à un Web de données interconnectés où chaque élément contribue à un écosystème général de la connaissance [Passant *et al.*, 2009c].

Chapitre 3

Rôle et définition d'un ensemble d'ontologies pour l'Entreprise 2.0

INTRODUCTION

Comme nous l'avons évoqué dans le chapitre précédent, notre système de médiation sémantique fait appel à des ontologies que l'on peut considérer d'une part comme des ontologies socio-structurelles et d'autre part des ontologies métier ou de domaine. Bien que ces ontologies aient des objectifs relativement distincts, nous avons fait à chaque fois le choix de développer des ontologies légères de manière à permettre une appropriation simple de celles-ci et leur réutilisation dans d'autres contextes. Nous présenterons ici ces deux types d'ontologies, répondant chacun à un besoin distinct, ainsi que la manière dont elles interagissent pour proposer un modèle complet pour la modélisation à la fois des contenus et des contenants produits au sein d'écosystèmes d'Entreprise 2.0.

Tout d'abord, nous expliciterons nos travaux autour de SIOC, ontologie dédiée à la représentation des communautés en ligne et de leurs activités. Nous détaillerons ainsi le modèle (son noyau et les modules associés) et la manière dont il interagit avec différentes ontologies déjà populaires sur le Web Sémantique. Nous présenterons également dans cette première partie différents vocabulaires précurseurs au travers d'un état de l'art consacré aux modèles de représentation d'activités Web 2.0 via les technologies du Web Sémantique. Nous étudierons également en guise d'évaluation la manière dont SIOC a pu être adopté à grande échelle, mettant ainsi l'accent sur ce qui nous semble être des bonnes pratiques pour l'acceptation d'un vocabulaire sur le Web Sémantique.

Nous présenterons ensuite les différentes ontologies de domaine que nous avons définies au sein de notre architecture de médiation. Puisque celles-ci sont par nature relatives aux domaines abordés, nous allons particulièrement nous intéresser aux choix de modélisation auxquels nous avons été confrontés et à la manière dont nous y avons fait face. Nous allons également voir en quoi, malgré le besoin d'un lien fort avec les domaines métier, le niveau d'abstraction que nous avons choisi permet d'en réutiliser une partie dans d'autres domaines. Ceci conduit à des ontologies que l'on pourrait considérer à mi-chemin entre des ontologies de domaine et des ontologies génériques au sens où l'adhérence avec le domaine se situe plus au niveau de la base de connaissances que du modèle lui-même. Plus particulièrement, nous présenterons l'utilisation et l'extension de modèles existants pour représenter la notion d'agent et ses différentes propriétés et expliciterons le choix du modèle

SKOS [Miles et Bechhofer, 2008] pour la définition d'une ontologie des rôles.

Ensuite, nous présenterons MOAT, modèle dédié à la formalisation de liens entre tags et concepts du Web Sémantique et permettant ainsi d'établir un pont entre les pratiques classiques de *tagging* et l'indexation sémantique. Nous nous attarderons également dans cette partie sur l'état de l'art relatif aux liens entre ontologies et folksonomies et situerons nos travaux dans ce contexte, à la fois dans le domaine de la définition d'ontologies à partir d'analyse de tags mais aussi dans celui de la représentation des tags (et des objets associés) via des modèles du Web Sémantique. Plus particulièrement, nous motiverons la définition de MOAT au travers de cet état de l'art et présenterons de quelle manière celui-ci s'intègre avec des modèles existants. Nous présenterons également de quelle manière cette vision que nous défendons, à savoir l'utilisation d'ontologies de domaine et de base de connaissance en support des tags pour résoudre certaines de leurs limites, a été acceptée sur le Web.

Enfin, nous détaillerons comment s'intègrent globalement ces différents niveaux d'ontologies pour proposer un modèle complet de représentation des connaissances pour l'Entreprise 2.0, combinant ainsi des aspects purement documentaires et sociaux et des aspects plus formels de connaissances métier. Outre les modèles définis, une des originalités de notre approche est ainsi de proposer une combinaison cohérente au sein d'un même système de médiation de ces différents niveaux de représentation, alors que la plupart des systèmes de médiation se concentrent uniquement sur la couche métier comme nous l'avons vu dans le chapitre précédent (Section 2.3.4, page 77).

3.1 MÉTADONNÉES SOCIO-STRUCTURELLES POUR LE WEB 2.0 AVEC SIOC

3.1.1 Identification des Besoins

Comme nous l'avons déjà mentionné dans ce mémoire, les échanges d'informations sur le Web et en entreprise sont généralement centrés autour d'objets particuliers (Section 1.2.3, page 41). Or, la diversité des services proposés (blogs, wikis, agrégateurs RSS, services de partage de contenus ...) introduit généralement une fragmentation des informations et des documents créés au sujet de ces objets. Par exemple, les informations relatives à un artiste particulier peuvent être réparties entre une éventuelle biographie sur Wikipedia, un profil sur Last.fm, des photos de concerts sur Flickr ou bien encore des billets de blogs distribués au sein de la blogosphère. En entreprise, le problème est sensiblement le même. Si l'on prend un projet particulier, il est fort probable que sa description soit publiée sur un wiki mais que des comptes-rendus de réunion soient postés sur différents blogs ou bien encore que les flux RSS contiennent des informations importantes sur les différents partenaires du projet. On peut même imaginer l'utilisation de canaux de messagerie instantanée ou de microblogging pour communiquer plus aisément au sujet de certains aspects du projet, fragmentant encore un peu plus les informations à son sujet. En conséquence, que cela soit sur le Web ou dans un contexte d'Entreprise 2.0, cette fragmentation de services complexifie la recherche d'information (Section 2.2.1, page 62). Il est en effet nécessaire d'interroger diverses sources de données pour obtenir une vue globale au sujet d'un objet ou d'un domaine particulier. En complément, il est de plus nécessaire de connaître l'existence et l'emplacement de ces différentes sources.

Cette hétérogénéité des applications se traduit également par l'absence de format commun pour représenter les documents et les interactions sociales produites depuis celles-ci. En particulier, les structures de bases de données ou les APIs proposées (pour peu que les services en possèdent) reposent généralement sur des modèles distincts, obligeant les développeurs et utilisateurs à adapter les requêtes à l'outil utilisé. Ainsi, une requête pour identifier les derniers documents publiés s'écrira différemment si l'on interroge un service comme Flickr, un blog sous Wordpress, un autre sous Drupal ou bien un wiki utilisant ce même outil.

Afin de répondre à cette double problématique, et permettre la défragmentation d'informations issues d'outils sociaux, il nous a semblé utile de proposer un modèle RDF de représentation commun de contenus Web 2.0 pour s'abstraire des formats de données initiaux et permettre la représentation standardisée de contenus créés à partir d'outils distincts. Un tel modèle offre également la possibilité de créer un lien entre outils Web 2.0, ceux-ci partageant alors une sémantique commune pour représenter une partie de leurs métadonnées. À l'aide de ce modèle, une même requête peut-être utilisée pour répondre à la question "*Quels sont les titres de tous les items créés en Janvier 2008 ayant reçu au moins un commentaire*" que l'outil d'origine soit un blog sous Wordpress, un wiki sous Drupal ou que l'on interroge un service de partage de photos, à partir du moment où il a été possible de représenter les données produites de manière uniforme. Surtout, cette sémantique commune permet d'unifier différents outils qui interagissent généralement comme des îlots de données complètement décorrélatés.

Nous avons ainsi activement participé au développement du projet SIOC – *Semantically-Interlinked Online Communities* [Breslin *et al.*, 2005] –, vocabulaire poursuivant ce but de représentation des activités des communautés en ligne. Pour reprendre le vocabulaire proposé dans le chapitre précédent, nous considérons donc SIOC comme une ontologie permettant la représentation de métadonnées socio-structurelles : celui-ci va permettre aussi bien de représenter les communautés et leurs activités que les documents produits et la façon dont ceux-ci sont structurés et interconnectés. Comme le montre le schéma qui suit (Figure 3.1, page 86), SIOC permet une représentation uniforme des outils collaboratifs et des documents qu'ils permettent de générer, permettant d'envisager une notion de forums virtuels en support de l'existant. Notre intérêt pour SIOC a débuté fin 2005, alors que nous étions nous même partis sur la définition d'un tel modèle et nous avons participé à son élaboration jusqu'à sa *Member Submission* au W3C en Juin 2007¹ en tant que coauteur de la spécification [Berrueta *et al.*, 2007] et éditeur de certains documents associés [Bojārs *et al.*, 2007a] [Fernández *et al.*, 2007b]. Ajourd'hui toujours, nous sommes des contributeurs actifs du projet, travaillant notamment sur d'autres cadres d'utilisation de SIOC, sur lesquels nous reviendrons plus tard (Section 3.1.6, page 101).

Avant de présenter SIOC plus en détail (Section 3.1.3, page 89), nous allons revenir sur un état de l'art relatif aux approches antérieures poursuivant ce même but de représentation des activités Web 2.0 via les technologies du Web Sémantique.

¹<http://www.w3.org/Submission/2007/02/>

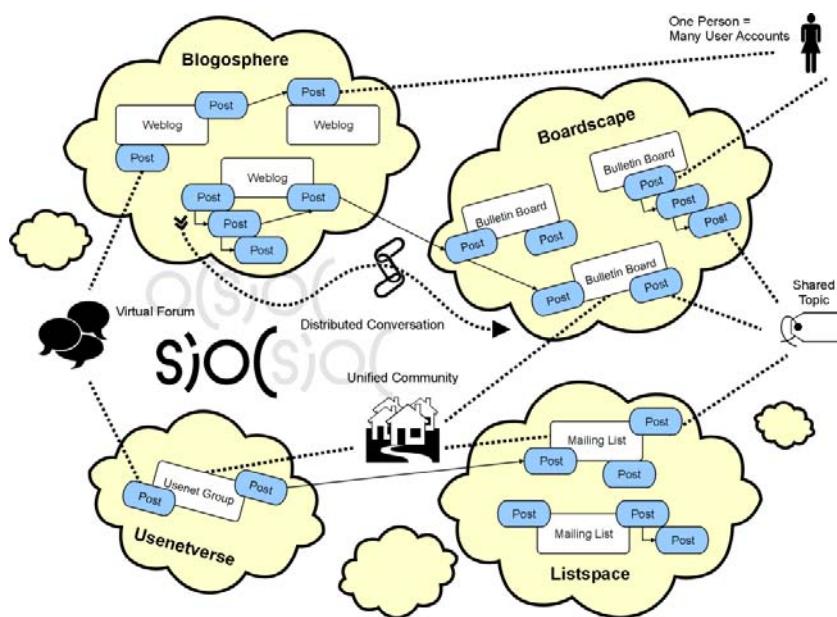


Figure 3.1: Intégration de données hétérogènes réparties avec SIOC [Bojārs *et al.*, 2006]

3.1.2 Positionnement par rapport à de l'art

Un des premiers modèles qui vient à l'esprit lorsque l'on parle de représentation uniforme de documents provenant d'outils sociaux est généralement RSS (Section 1.2.2, page 36). Celui-ci offre en effet un modèle commun pour la syndication de données, et dispose d'une version RDF² qui permet d'envisager son utilisation dans le contexte du Web Sémantique. Il souffre cependant de certaines limites :

- RSS représente les items (billet de blogs, pages wiki ...) mais ne permet pas de représenter les autres données relatives à la plate-forme. Ainsi, on n'exporte ni la description du site associé, ni les utilisateurs et les interactions sociales associées. Nous sommes donc ici dans un contexte de modélisation de métadonnées essentiellement documentaires plutôt que structurelles et socio-structurelles. De plus il n'est pas possible de faire la distinction entre les types de documents exportés (pour distinguer par exemple un billet de blog d'une page wiki), RSS modélisant uniquement la notion d'élément au sens large ;
- il s'agit d'un format de syndication et non d'export. En ce sens, il est possible de suivre en continu les mises à jour d'un site mais pas d'exporter l'ensemble des contenus publiés sur un site depuis sa création. On ne peut donc pas disposer d'un historique complet d'une application en utilisant RSS, à moins d'avoir mis en place un tel export dès le début, et conservé les différents flux exportés ;
- si RSS 1.0 est un format RDF – et peut donc être étendu avec d'autres vocabulaires pour par exemple affiner les types d'éléments publiés – , il n'est malheureusement pas

²<http://web.resource.org/rss/1.0/>

majoritaire en termes de présence sur le Web, et la plupart des lecteurs RSS ne savent interpréter que son modèle de base, et non pas ses éventuelles extensions. On peut cependant nuancer ce point à partir du moment où il s'agit plus d'un problème d'implémentation que théorique. Ceci met cependant en avant un problème d'adoption des technologies du Web Sémantique à grande échelle, tel que nous l'avons évoqué au début de ce mémoire (Section 1.1.4, page 27).

Hormis RSS, de nombreux travaux ont eu lieu plus spécifiquement autour de la modélisation des données de blogs, recensés pour la plupart dans un premier état de l'art sur le sujet [Cayzer, 2006]. [Cayzer et Shabajee, 2003] ont ainsi défini une première idée du *semantic blogging* en envisageant deux facteurs permettant d'augmenter le potentiel des outils de blogs grâce aux technologies du Web Sémantique à savoir (1) une structure riche (aussi bien au niveau des métadonnées des billets que de leurs thématiques avec des ontologies partagées) et (2) des requêtes plus puissantes (en termes de souscription, de découverte et de navigation de contenu). [Cayzer, 2004] revient plus tard sur ces propositions en définissant cette fois trois idées, auxquelles par ailleurs SIOC et les différentes applications utilisant ce modèle permettent d'aboutir :

- la *vue*, i.e. l'utilisation des métadonnées des différents billets pour enrichir les interfaces de visualisation et d'agrégation ;
- la *navigation*, i.e. l'utilisation des métadonnées pour faciliter la navigation, en fonction par exemple d'une thématique donnée ;
- les *requêtes*, i.e. l'utilisation des métadonnées pour répondre à des requêtes avancées, par exemple lister les billets d'une communauté donnée.

Un premier outil mettant en pratique ces trois propositions est également proposé, le *Snippet manager* [Cayzer et Castagna, 2005].

[Karger et Quan, 2004] définissent quant à eux les implications de bloguer sur le Web Sémantique. Ils précisent notamment que les outils capables de produire des contenus structurés et interprétables doivent pouvoir le faire de manière autonome, sans intervention supplémentaire de l'utilisateur. Nous reviendrons sur ce point dans le chapitre suivant en détaillant la manière dont nous automatisons la production d'annotations socio-structurelles dans nos outils (Section 4.1, page 138). [Karger et Quan, 2004] proposent également une architecture permettant de parvenir à cet objectif. En se basant sur RSS, ils définissent :

- d'une part des possibilités de passer des différents formats RSS à RSS 1.0 en utilisant des feuilles de style XSLT. Cette idée de transformation de formats sera reprise plus tard par des outils comme Triplr³ ou Babel⁴ ;
- d'autre part une ontologie étendant RSS avec la notion de réponse et de sujets partagés (via une classe Collection). Ce modèle reprend sur certains points leurs travaux précédents relatifs à la modélisation de messages en ligne (IRC ...) [Quan *et al.*, 2003a].

Un prototype basé sur Haystack [Quan *et al.*, 2003b] est également présenté et montre de quelle manière la sémantique commune offre de nouvelles méthodes d'agrégation et de navigation entre billets de blog. Dans cet outil, les billets sont essentiellement considérés comme des annotations au sujet d'autres contenus Web. Ils bénéficient ainsi en plus d'un

³<http://triplr.org>

⁴<http://simile.mit.edu/babel/>

modèle pour représenter les opinions de leurs auteurs par rapport aux contenus d'origine (opposition, accord ...), envisageant une première manière de modéliser la notion d'argumentation sur le Web Sémantique.

Si ces approches s'attachent principalement à représenter des métadonnées documentaires, d'autres travaux se sont consacrés à extraire ou modéliser le contenu même des documents. Si ceux-ci sont assez importants dans le monde des wikis sémantiques (Section 4.2.1, page 148) certains travaux sont à signaler dans le domaine des blogs. Par exemple, l'idée du *Structured Blogging*⁵ permet de définir des patrons de pages (ou *templates*) pour créer des contenus selon des champs prédéfinis. On peut par exemple définir un patron *avis* permettant de donner son opinion sur un objet donné avec un champ nom, une note et un commentaire⁶. En plus de permettre un affichage uniforme des documents, ces *templates* peuvent être associés à l'utilisation des microformats⁷, introduits dans le premier chapitre de ce mémoire. Dans cette même direction, [Cayzer, 2006] présente avec BlogAccord une manière de combiner structuration et production de données RDF à partir de billets de blog. Un composant supplémentaire vient s'y ajouter avec la possibilité d'intégrer automatiquement des informations externes au sein d'un billet en fonction de ressources disponibles sur le Web. Il est par exemple possible d'intégrer automatiquement des informations issues de la base de connaissance MusicBrainz⁸ lors de l'écriture d'un article sur un artiste donné. Nous reviendrons sur cette notion d'intégration de sources externes dans les chapitres suivants (Section 4.2.4, page 163). D'autres outils permettent également de réutiliser les données présentes sur le poste de travail utilisateur pour enrichir les informations publiées en ligne, cette vision s'intégrant dans l'idée du *Semantic Desktop* introduit dans le chapitre précédent. [Möller *et al.*, 2006] offrent ainsi avec semiBlog la possibilité d'intégrer au sein de billets de blog des informations issues par exemple d'une application de carnet d'adresses, le contenu étant ensuite publié en RDF et pouvant donc être utilisé par d'autres applications.

D'autres modèles spécifiques, non liées aux blogs, ont également été proposés dans cet objectif de représentation de données sociales avec les technologies du Web Sémantique. SAM [Franz et Staab, 2005] ou NABU [Osterfeld *et al.*, 2005] proposent tous deux des vocabulaires relatifs à la messagerie instantanée, réutilisant notamment FOAF et DublinCore. Plus récemment, des projets de modélisation des listes de diffusions e-mail comme SWAML [Fernández *et al.*, 2007a] ou mle [Rehatschek et Hausenblas, 2007] ont vu le jour. Tous deux sont basés entre autre sur SIOC, SWAML étant désormais entièrement intégré à ce dernier⁹. Pour les wikis, dont nous détaillerons plus loin d'autres aspects (Section 4.2.1, page 148), des modèles ont été proposés avec WIF et WAF [Völkel et Oren, 2006] respectivement pour l'échange et l'archivage des données entre wikis. WikiOnt [Harth *et al.*, 2005] suit cette même idée de modèle standard pour définir la structure des wikis, ce dernier réutilisant DublinCore et définissant les notions de pages, de catégories et de liens internes et externes. Une partie de WikiOnt est d'ailleurs aujourd'hui intégrée dans SIOC et nous avons récemment

⁵<http://structuredblogging.org>

⁶<http://structuredblogging.org/formats.php>

⁷<http://microformats.org>

⁸<http://musicbrainz.org>

⁹http://developer.berlios.de/forum/forum.php?forum_id=25510

approfondi l'utilisation de SIOC pour prendre en compte certaines caractéristiques particulières des wikis¹⁰. De plus, des outils comme SweetWiki ou IkeWiki (Section 4.2.1, page 148) définissent également leur propre format de représentation pour la structure des wikis.

3.1.3 Présentation du modèle de représentation SIOC

Classes et propriétés

SIOC est une ontologie volontairement légère (au sens des *lightweight ontologies* comme défini par [Gómez-Pérez et Corcho, 2002]) de manière à ce qu'elle puisse facilement être étendue selon les besoins de chacun. La cible visée par SIOC étant celle des communautés Web 2.0 et notamment des développeurs Web 2.0, nous avons souhaité que le modèle soit suffisamment simple pour être abordé par tous, tout en étant suffisamment expressif pour capter l'ensemble des activités des communautés en ligne. SIOC se compose d'un noyau¹¹ et de deux modules principaux : un module Types¹² et un module Services¹³, sa spécification complète étant disponible en ligne¹⁴. Le noyau se compose de 11 classes¹⁵, que l'on peut regrouper en deux parties : une première consacrée à l'aspect social (comptes utilisateur), une seconde à l'aspect structurel (contenus et conteneurs). L'ensemble permet ainsi de représenter la quasi-totalité des éléments d'une communauté en ligne. À la frontière de ces deux aspects se trouve justement la classe `sioc:Community` qui comme son nom l'indique permet de représenter une communauté d'intérêt. Elle peut ainsi regrouper un certain nombre d'éléments qui peuvent faire partie sans distinction des deux groupes précédents. Bien entendu, une communauté peut rassembler des éléments provenant d'espaces distincts sur le Web, et c'est là un des objectifs de SIOC, à savoir créer des passerelles entre différents outils Web 2.0.

Afin de représenter les comptes utilisateurs et le ou les rôles qui peuvent leur être associés au sein de différents services, SIOC définit trois classes :

- `sioc:User` – un compte utilisateur sur un service en ligne, auquel vont être rattachés les différents contenus produits. Il s'agit ici du compte au sens entité virtuelle et non pas de la personne physique associée, celle-ci étant modélisée avec FOAF (Section 3.1.4, page 94) ;
- `UserGroup` – un groupe d'utilisateurs (`sioc:User`), réunis (explicitement) car partageant par exemple des intérêts ou des rôles communs au sein d'un service en ligne ;
- `Role` – le rôle assigné à un utilisateur ou à un groupe. Ce rôle est typé (administrateur, modérateur ...) et contextualisé en fonction d'un espace de communication donné (un forum, un blog ...).

SIOC ne s'attache pas à modéliser le contenu des documents mais uniquement certaines de leurs métadonnées. Trois classes principales (et différentes sous-classes associées) sont proposées dans cette perspective de modélisation :

¹⁰http://groups.google.com/group/sioc-dev/browse_thread/thread/449b9f8fa2a71590

¹¹Espace de noms <http://rdfs.org/sioc/ns>, préfixe `sioc` par la suite.

¹²Espace de noms <http://rdfs.org/sioc/types>, préfixe `siocTypes` par la suite.

¹³Espace de noms <http://rdfs.org/sioc/services>, préfixe `siocServices` par la suite.

¹⁴<http://rdfs.org/sioc/spec/#sec-modules>

¹⁵Révision 1.30 du 9 Janvier 2009.

- **sioc:Space** – un espace communautaire. Le niveau d’abstraction est volontairement élevé, permettant de représenter aussi bien un système de fichier qu’un site Web. La classe **sioc:Site** définie dans le noyau de SIOC est ainsi une représentation concrète des sous-classes possibles ;
- **sioc:Container** : un conteneur de données communautaires, inclu dans le précédent espace. Ce Container représente le niveau où sont présentes les données. Plus précisément, **sioc:Forum** et **sioc:Thread** sont deux sous-classes de ce Container, le module Types en définissant d’autres ;
- **sioc:Item** – un élément présent dans un conteneur. **sioc:Post**, sous-classe, représentant un message au sens large (sur un forum, dans un blog, une page wiki ...). À nouveau le module Types définit des sous-classes plus spécifiques (Section 3.1.3, page 92).

À ces classes viennent s’ajouter un certain nombre de propriétés. Sans toutes les détailler, voici celles qui nous intéressent plus particulièrement dans notre contexte d’expérimentation en entreprise :

- **sioc:content** – le contenu textuel d’un **sioc:Item**. Celui-ci est représenté en texte brut et peut ainsi être utilisé dans une requête SPARQL avec une clause FILTER pour limiter les recherches aux documents devant contenir certains termes¹⁶ ;
- **sioc:container_of**¹⁷ – cette propriété permet de faire le lien entre une instance de **sioc:Container** et l’ensemble des instances de **sioc:Item** qu’il contient. C’est par son intermédiaire que l’on modélise par exemple qu’un billet appartient à un blog donné. Nous verrons plus en détail l’intérêt de cette propriété pour limiter la recherche d’information à un conteneur donné (Section 5.4, page 212) ;
- **sioc:creator_of** (inverse : **sioc:has_creator**) – permet d’établir un lien entre un utilisateur (**sioc:User**) et un **sioc:Item** afin d’identifier l’auteur d’un contenu. Une instance de **sioc:Item** peut avoir plusieurs propriétés de ce type qui lui sont assignées selon le type d’objet manipulé, par exemple une page wiki ;
- **sioc:reply_of** (inverse : **sioc:has_reply**) : permet d’établir des liens entre deux instances de **sioc:Item**, en considérant l’un comme réponse de l’autre. Rien n’oblige les deux items en question à être issus du même outil ce qui permet ainsi de représenter un système de commentaire décentralisé comme les *trackbacks*¹⁸ ;
- **sioc:num_replies** et **sioc:num_views** : indiquent respectivement le nombre de réponses associées à un **sioc:Item** et son nombre de lectures. Ces propriétés sont particulièrement utiles étant donné que SPARQL n’implémente pas nativement de fonctions agrégat (Section 1.1.3, page 25)¹⁹ ;
- **sioc:topic** : permet d’assigner un ensemble de sujets à un Item, chaque sujet correspondant à une ressource identifiée via son URI. Nous verrons successivement comment se modélise ainsi l’assignation d’un item à un élément de taxonomie défini par

¹⁶Notons que pour représenter un contenu encodé en HTML, nous suggérons avec SIOC l’utilisation de la propriété **content** du module RSS 1.0 du même nom. – <http://purl.org/rss/1.0/modules/content/>

¹⁷Propriété inverse : **sioc:has_container**

¹⁸http://www.sixapart.com/pronet/docs/trackback_spec

¹⁹On peut cependant regretter que des choix de modélisation soient orientés principalement en raison des outils associés.

exemple via SKOS (Section 3.1.4, page 94) mais aussi de manière plus large à toute ressource du Web Sémantique ou instance d'ontologie du domaine avec MOAT (Section 3.3, page 120).

Le schéma suivant (Figure 3.2, page 91) synthétise les différentes classes et propriétés du cœur de SIOC.

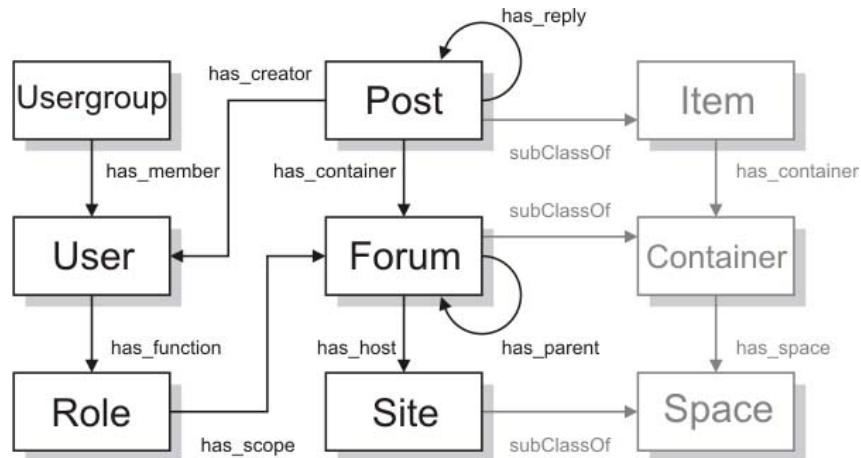


Figure 3.2: Le modèle de classes et propriétés de SIOC [Berrueta *et al.*, 2007]

Avant de terminer cette présentation du noyau de SIOC, prenons un exemple de document représenté avec ce modèle et une requête SPARQL associée. L'exemple qui suit représente donc un élément et sa réponse associée (Listing 3.1, page 91) alors que la requête correspond à la question "*Quel est le titre des items créés en Janvier 2008 ayant reçu au moins un commentaire*" (Listing 3.2, page 92) :

```

<http://example.org/blog/post/33> a sioc:Post ;
  dct:title "Mon billet exemple" ;
  sioc:content "Ceci est mon premier billet" ;
  sioc:has_creator <http://example.org/user/alex> ;
  sioc:num_replies 1 .

<http://example.org/blog/post/33/comment_1> a sioc:Post ;
  sioc:reply_of <http://example.org/blog/post/33> .
    
```

Listing 3.1: Exemple de contenu Web 2.0 avec SIOC

Enfin, signalons pour finir que SIOC est modélisé en OWL avec un niveau d'expressivité OWL-Lite. En conséquence, le modèle ne dispose pas de contraintes de cardinalité relatives aux différentes propriétés, même si cette question a été plusieurs fois abordée lors de sa définition. Si le modèle a longtemps été défini en RDFS, induisant notamment par certaines propriétés une expressivité OWL-Full, nos récents travaux autour de SWANSIOC (Section 3.1.6, page 101) nous ont amené à affiner celui-ci pour passer à un niveau OWL-Lite. SIOC

```

SELECT ?item ?title
WHERE {
  ?item rdf:type sioc:Post ;
    dct:title ?title ;
    dct:created ?date ;
    sioc:num_replies ?replies .
  FILTER ( ?date > "2008-01-01T00:00:00"^^xsd:dateTime ) .
  FILTER ( ?date < "2008-02-01T00:00:00"^^xsd:dateTime ) .
  FILTER ( ?replies >= 1 ) .
}

```

Listing 3.2: Exemple de requête SPARQL dédiée à SIOC

peut ainsi être utilisé au sein d'applications bénéficiant de capacités de raisonnement, tout en s'assurant que celles-ci peuvent s'effectuer en un temps fini.

Les modules de SIOC

Comme nous l'avons évoqué précédemment, la volontaire légèreté de SIOC fait que certaines caractéristiques des services Web 2.0 sont représentées au sein de modules additionnels et non pas directement dans le noyau de SIOC.

Le module Services²⁰ permet ainsi de représenter la présence (et l'emplacement) de services Web associées à des éléments Web 2.0, par exemple l'emplacement d'une API ou d'un point d'accès SPARQL. Il utilise pour cela une propriété `siocs:has_service` et une classe `siocs:Service` éventuellement associés à `siocs:has_format` pour représenter le format de celui-ci. Ce module est relativement léger et son objectif est de fournir un moyen simple de modéliser des services Web et APIs Web 2.0 sans s'aventurer dans des descriptions plus complexes avec des modèles comme WSDL [Christensen *et al.*, 2001] ou WSMO [Vitvar *et al.*, 2008]. Nous n'insisterons pas plus sur ce module, ne l'ayant pas mis en pratique dans nos travaux. Il peut cependant être utile à partir du moment où des services Web 2.0 exposent leurs données via SPARQL.

Le second module de SIOC est le module Types²¹ qui définit un certain nombre de conteneurs et d'items venant sous-classer `sioc:Container` (ou plus précisément `sioc:Forum`) et `sioc:Item` (ou `sioc:Post`). Ceci permet de typer plus finement les documents générés par des services Web 2.0 afin de prendre en compte la spécificité de certains d'entre eux. Par exemple, on peut à l'aide de ce module formellement différencier un blog (`sioc:Blog`) d'une liste de diffusion (`sioc:MailingList`), et une page wiki (`sioc:WikiArticle`) d'un message de microblogging (`sioc:MicroblogPost`). De plus, le fait de définir ces différents types comme sous-classes de `sioc:Container` et `sioc:Item`, couplé aux principes d'inférence associés à ces hiérarchies de classes (Section 1.1.2, page 21), permet au travers d'une requête telle que "*Lister les instances de sioc:Item*" d'identifier des instances qui n'ont pas été définies directement en tant que `sioc:Item` mais comme instances d'une des sous-classes proposées par le module. Ce module Types définit ainsi une vingtaine de classes

²⁰<http://rdfs.org/sioc/services>

²¹<http://rdfs.org/sioc/types>

distinctes dont certaines sont de plus alignées avec des vocabulaires existants. Le tableau suivant (Tableau 3.1, page 93) représente ainsi un ensemble de sous-classes de Container et les éventuelles classes Item associées, ainsi que dans certains cas les alignements avec des modèles existants, tels que défini dans ce module.

Conteneur	Item
sioct:AudioChannel	dcmitype:Sound ^a
sioct:BokmarkFolder	bookmark:Bookmark ^b
sioct:ImageGalery	exif:IFD ^c
sioct:MailingList	sioct:MailMessage
sioct:MessageBoard	sioct:BoardPost
sioct:Microblog	sioct:MicroblogPost
sioct:VideoChannel	dcmi:MovingImage ^d
sioct:Webglog	sioct:BlogPost
sioct:Wiki	sioct:WikiArticle

^a<http://purl.org/dc/dcmitype/Sound>

^b<http://www.w3.org/2002/01/bookmark#Bookmark>

^c<http://www.w3.org/2003/12/exif/ns#IFD>

^d<http://purl.org/dc/dcmitype/MovingImage>

Tableau 3.1: Eléments du module Types de SIOC

Le code qui suit reprend ainsi l'exemple précédent en réutilisant certaines classes définies dans le module Type pour spécifier les instances de sioct:Post utilisés (Listing 3.3, page 93). Comme on peut le constater, l'utilisation de ce module n'implique pas de modification vis-à-vis des autres éléments utilisés.

```
<http://example.org/blog/post/33> a sioct:BlogPost ;
  dct:title "Mon billet exemple" ;
  sioc:content "Ceci est mon premier billet" ;
  sioc:has_creator <http://example.org/user/alex> ;
  sioc:num_replies 1 .

<http://example.org/blog/post/33/comment_1> a sioct:Comment
  ;
  sioc:reply_of <http://example.org/blog/post/33> .
```

Listing 3.3: Exemple de billet de blog avec SIOC et son module Types

En permettant ainsi de typer finement les différents documents produits, ce module est ainsi un élément essentiel de notre proposition d'écosystème sémantique, aussi bien en termes de modélisation des contenus à partir des différents adaptateurs (Section 4, page 137) que pour les requêtes associées. En effet, comme nous le verrons par la suite, il permet de bénéficier d'une sémantique commune pour représenter les documents créés aussi bien

depuis les blogs, les wikis ou l'agrégateur de flux RSS mis en place dans notre système, tout en offrant la possibilité de distinguer ceux-ci au moment des requêtes.

3.1.4 Alignement avec des vocabulaires existants

Lors de la définition de SIOC, nous avons au maximum essayé de réutiliser des vocabulaires existants et déjà populaires sur le Web Sémantique, soit en alignant les classes et propriétés de SIOC avec celles de ces ontologies, soit en suggérant leur utilisation dans certains contextes (comme nous l'avons vu auparavant avec le module Types). Ceci nous semble nécessaire dans la mesure où nous souhaitons que les activités et documents représentés avec SIOC fassent partie intégrante du Web Sémantique et ne soient pas considérés comme faisant partie d'un écosystème disjoint de l'existant. Ces bonnes pratiques ont d'autre part été consignées dans un document associé à sa Soumission Membre au W3C [Bojārs *et al.*, 2007a] et nous allons ici présenter certains de ces alignements.

DublinCore

De nombreuses propriétés nécessaires à la modélisation des éléments visés par SIOC sont disponibles dans DublinCore²² [Dublin Core Metadata Initiative, 2006]. C'est par exemple le cas du titre (`dct:title`) ou de la date de création (`dct:created`) et de modification (`dct:modified`) d'un élément²³. SIOC suggère ainsi l'utilisation de ces propriétés, comme le montre l'exemple suivant (Listing 3.4, page 94).

```
<http://athena.der.edf.fr/blog/2006/08/09/104-sample-post> a
  sioc:BlogPost ;
  dct:title "Billet de test" ;
  dct:created "2006-08-03T22:50:32Z</dct:created">;
  dct:modified "2006-09-19T23:36:05Z</dct:modified">;
  dct:subject "EDF" .
```

Listing 3.4: Utilisation de propriétés issues du DublinCore avec SIOC

FOAF

Si SIOC définit la notion d'utilisateur d'un service Web en tant qu'entité en ligne, il ne s'attache pas à modéliser la personne physique associée à ce compte. Pour prendre en compte cet aspect, SIOC se base ainsi sur le vocabulaire FOAF (Section 3.2.2, page 104). Nous utilisons ainsi la propriété `foaf:holdsAccount` pour établir un lien entre une personne physique (en réalité une instance `foaf:Agent`) et son ou ses différents comptes en ligne (`sioc:User`) et introduisons également une propriété inverse `sioc:account_of`. Cette utilisation combinée de SIOC et FOAF rend donc possible le rattachement d'un ensemble de comptes en ligne à une même personne physique (Figure 3.3, page 95). Ces comptes peuvent bien entendu être distribués sur le Web, cette complémentarité prenant alors tout son sens

²²Espace de noms <http://purl.org/dc/terms/>, préfixe `dct` par la suite.

²³Les premières versions de SIOC ont défini des propriétés similaires, aujourd'hui déclarées comme obsolètes avec `owl:DeprecatedProperty`.

pour modéliser l'ensemble des activités sociales d'une personne selon différents services (Section 3.1.5, page 96).

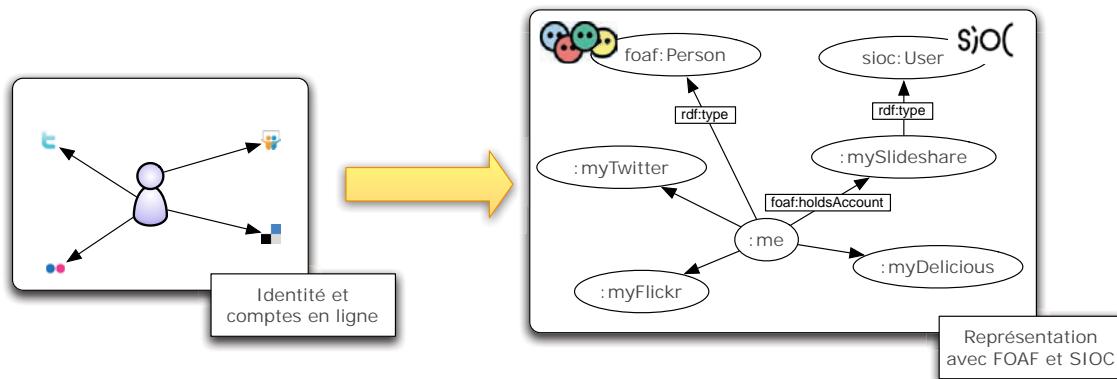


Figure 3.3: Comptes utilisateur et personne physique avec SIOC et FOAF

Nous suggérons également dans [Bojārs *et al.*, 2007a] l'utilisation de `foaf:maker` pour établir directement un lien entre documents et personne physique, et non pas uniquement via le couple `sioc:has_creator` / `sioc:User` qui permet d'établir un lien entre document et compte utilisateur. Cette relation directe entre un document et une personne peut cependant être inférée à partir de ce couple `sioc:has_creator` / `sioc:User` et de la règle d'inférence qui suit (Listing 3.5, page 95) :

```
{
  iii a sioc:Item ;
    sioc:has_creator uuu .
  uuu a sioc:User ;
    sioc:email_sha1 mmm .
  aaa a foaf:Agent ;
    foaf:mbox_sha1sum mmm .
} => {
  iii foaf:maker aaa .
}
```

Listing 3.5: Règle d'inférence pour lier SIOC et FOAF, représentée en N3

RSS 1.0

Comme nous l'avons signalé, SIOC réutilise la propriété `encoded` du module Content²⁴ de RSS 1.0 en suggérant son utilisation pour représenter le contenu encodé en (X)HTML d'une instance de `sioc:Item`, le contenu plein-texte étant lui représenté avec `sioc:content`. Notons également qu'il est possible de manière assez simple de passer d'un flux RSS à une

²⁴<http://web.resource.org/rss/1.0/modules/content/>

modélisation SIOC comme nous le verrons en détail dans le chapitre suivant (Section 4.1.2, page 140).

SKOS

SIOC peut également être combiné efficacement avec SKOS – Simple Knowledge Organisation Schema [Miles et Bechhofer, 2008] – dans un but d'indexation sémantique de contenus Web 2.0. Nous détaillerons SKOS plus loin dans ce mémoire (Section 3.2.4, page 109) mais signalons simplement ici que ce modèle permet la définition de vocabulaires contrôlés ou de taxonomies en RDF. SKOS permet en effet de définir des relations `skos:narrower` et `skos:broader` pour organiser hiérarchiquement différentes instances de `skos:Concept`. SKOS peut ainsi être utilisé par exemple pour définir une hiérarchie de catégories de blog, comme le propose un *plug-in* pour la plate-forme Wordpress²⁵, ou une taxonomie de concepts plus poussée qui peut être définie au sein d'une organisation.

La propriété `sioc:topic` permet ainsi de faire le lien entre les instances de `sioc:Item` et les instances de différents `skos:Concept` proposées par de telles taxonomies. Notons que par le passé, SKOS proposait une propriété similaire `skos:subject` dans son vocabulaire SKOS Core²⁶, aujourd'hui remplacé par le SKOS Vocabulary²⁷, ce premier vocabulaire et la propriété évoquée devenant ainsi obsolètes²⁸. On retrouve ici une optique similaire à ce qu'ont proposé [Cayzer et Shabajee, 2003] et [Karger et Quan, 2004] pour modéliser des thématiques partagés entre billets de blog, comme nous l'avons évoqué auparavant (Section 3.1.2, page 86). Nous verrons par la suite que des méthodes avancées de *tagging* couplées aux technologies du Web Sémantique permettent d'aller plus loin dans ce processus d'indexation sémantique (Section 3.3, page 120).

3.1.5 SIOC, FOAF et la portabilité des données Web 2.0

Comme nous l'avons évoqué en début de ce chapitre, si l'ascension du Web 2.0 a contribué à la publication spontanée de données et de réseaux sociaux sur le Web, elle entraîne également en contrepartie leur fragmentation. Les contributions sociales d'un utilisateur sont en effet souvent éclatées entre différents services agissant comme des îlots déconnectés, la communication et l'échange de données n'étant possibles qu'à l'intérieur d'une même plate-forme. Ainsi, l'inscription à différents services Web 2.0 implique :

- d'une part la nécessité de répliquer ses données si l'on souhaite qu'elles soient disponibles sur chacun des outils utilisés ;
- d'autre part de définir son réseau social sur chaque application, même si celui-ci a déjà été identifié sur un autre service.

Ce processus répétitif conduit à ce que certains appellent la *social network fatigue*²⁹. Si cela peut ne pas sembler problématique à première vue, l'analogie avec l'utilisation de l'e-mail permet de prendre conscience de ces limites : qui accepterait aujourd'hui de souscrire à un service où les e-mails ne peuvent être envoyés qu'à des utilisateurs du même service ?

²⁵<http://www.wasab.dk/morten/blog/archives/2004/09/01/skos-output-from-wordpress>

²⁶<http://www.w3.org/2004/02/skos/core.rdf>

²⁷<http://www.w3.org/2008/05/skos>

²⁸<http://www.w3.org/2004/02/skos/vocabs>

²⁹<http://factoryjoe.com/blog/2007/09/20/stop-building-social-networks/>

Ainsi, Brad Fitzpatrick définit mi-2007 sa vision d'un graphe social distribué et ouvert³⁰ en réponse à cette problématique. La charte *A Bill of Rights for Users of the Social Web*³¹ insiste quelque temps plus tard sur la notion de propriété relative aux données issues de ces différents sites. Alors que les conditions d'utilisation de la plupart d'entre eux stipulent que celles-ci appartiennent aux dits sites, cette charte défend la notion de propriété par leur auteur, *i.e.* l'utilisateur, afin d'en faire l'usage qu'il souhaite. Lancée plus récemment, l'initiative DataPortability³² s'inscrit également dans ce but d'échange transparent de données et de réseaux sociaux entre applications. En complément de ces efforts communautaires, des solutions propriétaires ont été proposées pour résoudre cette problématique (Google Open-Social³³, Facebook Connect³⁴ ...). Leur acceptation dépend cependant de notions politiques et économiques entre les différents acteurs de services de réseaux sociaux. Nous pensons que le Web Sémantique, notamment à travers SIOC et FOAF, permet de répondre à cette problématique d'interopérabilité entre données sociales de manière ouverte et qui plus est en se basant sur des formats et protocoles standards [Bojārs *et al.*, 2008a].

Tout d'abord, concernant les données, SIOC permet une modélisation uniforme des contenus quelque soit l'outil d'origine. Ainsi les contributions sociales d'un utilisateur, bien que produites via des outils distincts et distribués, sont unifiées au niveau de leur représentation sur le Web Sémantique. Cette sémantique commune permet en conséquence de standardiser les requêtes associées (avec SPARQL) mais surtout de faciliter les échanges de données entre services basés sur le même modèle. Pour exemplifier cette possibilité, un premier prototype d'import SIOC a été développé pour le système de blog WordPress³⁵.

Alors que SIOC permet de résoudre ce problème de portabilité et d'interopérabilité pour les données issues d'outils Web 2.0, un autre aspect important concerne les réseaux sociaux modélisés depuis ces applications. Ici, FOAF a un rôle important à jouer en tant que modèle de référence pour représenter l'identité personnelle et les réseaux d'accointance sur le Web Sémantique. Tout comme pour les contenus, le passage à un niveau de modélisation uniforme pour la représentation de ces réseaux permet de proposer une interopérabilité entre différentes applications. Différents exporteurs FOAF pour des services grand public (par exemple pour Flickr [Passant, 2008b]) permettent déjà de bénéficier de cette sémantique commune et des avantages qu'elle procure. Ainsi, en couplant cette notion de réseau social avec FOAF aux contenus modélisés avec SIOC, il est possible de représenter uniformément via un unique graphe RDF l'ensemble des contributions sociales et des accointances d'un individu au sein de différentes plates-formes (Figure 3.4, page 98).

Malgré tout, un problème d'unification d'identité se pose avec l'utilisation des exporteurs FOAF mentionnés précédemment. Ceux-ci redéfinissent en effet chacun une URI particulière pour l'individu modélisé. Brutes, ces données ne permettent donc pas d'identifier qu'une personne présente sur Flickr (identifiée par exemple par l'URI <http://apassant.net/home/2007/12/flickrdf/people/33669349@N00>) est la même que telle autre sur

³⁰<http://bradfitz.com/social-graph-problem/>

³¹<http://opensocialweb.org/2007/09/05/bill-of-rights/>

³²<http://dataportability.org>

³³<http://code.google.com/apis/opensocial/>

³⁴<http://developers.facebook.com/fbconnect.php>

³⁵http://wiki.sioc-project.org/w/SIOC_Import_Plugin

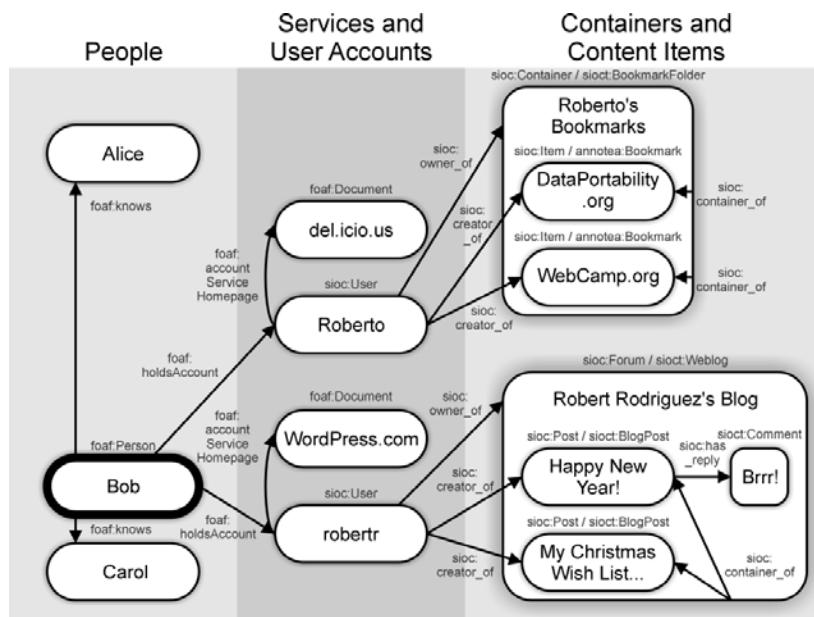


Figure 3.4: Interopérabilité entre données sociales avec SIOC et FOAF [Bojārs *et al.*, 2008b]

Last.fm (<http://dbtune.org/last-fm/terraces>). Il est donc nécessaire d'unifier ces représentations de manière explicite ou implicite :

- *explicitement*, une bonne pratique consiste en l'utilisation des principes d'identité OWL reposant sur la propriété `owl:sameAs`. Définir une relation de ce type entre deux instances de `foaf:Person` va ainsi permettre à un raisonneur d'établir que les deux URIs, bien que distinctes, identifient la même ressource, en l'occurrence la même personne physique ;
- *implicitement*, et toujours en utilisant les possibilités de raisonnement offertes par le Web Sémantique, la solution consiste à se baser sur les propriétés inverses fonctionnelles (`owl:InverseFunctionalProperty`)³⁶. FOAF définit un certain nombre de propriétés de ce type comme `foaf:mbox` et `foaf:openid`. Ainsi, associer un même e-mail à deux instances de `foaf:Person` va permettre d'identifier qu'il s'agit de la même personne.

Qu'elle soit implicite ou explicite, cette unification va permettre d'agréger les réseaux distribués d'un même individu, conduisant à la définition d'un réseau social distribué et ouvert (Figure 3.5, page 99).

À partir de celui-ci, il est relativement aisé de développer des applications de visualisation associées, comme nous l'avons fait avec l'application *FOAAGear*³⁷ (Figure 3.6, page 100). Celle-ci permet de visualiser de manière uniforme un ensemble de réseaux sociaux distribués et modélisés avec FOAF. De plus, le code permettant d'effectuer cette agrégation de

³⁶Pour rappel, deux ressources partageant une même valeur pour une propriété de ce type sont considérées comme identiques.

³⁷<http://apassant.net/home/2008/01/foafgear>

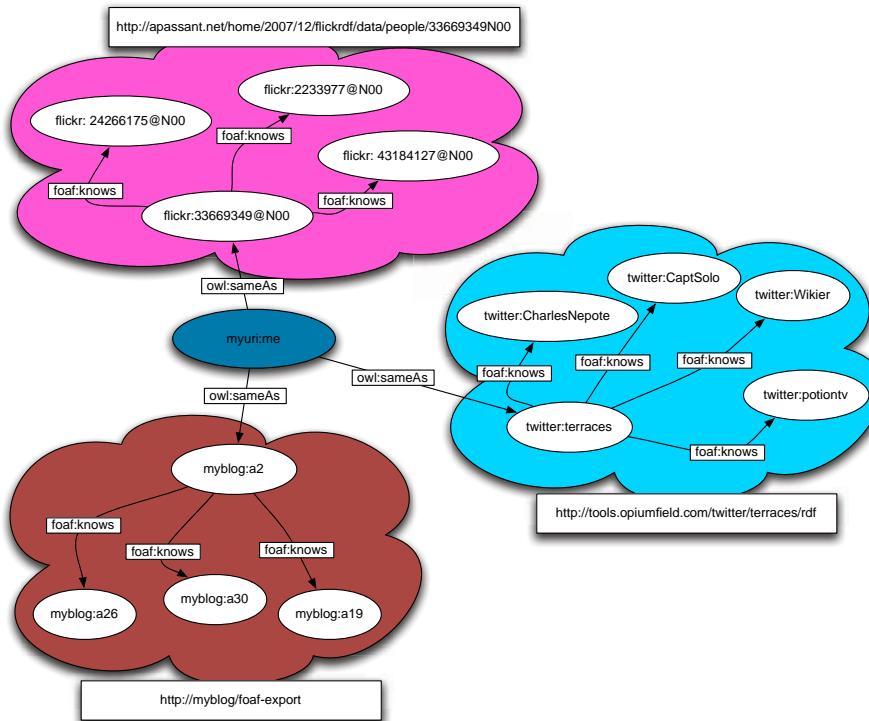


Figure 3.5: Unification de réseaux sociaux distribués avec owl :sameAS

réseaux sociaux ne compte qu'une centaine de lignes, et deux requêtes SPARQL, mettant ainsi en avant ces processus d'interopérabilité avec les technologies du Web Sémantique du point de vue du développement d'applications Web 2.0, et cette complémentarité entre les deux mondes.

Pour aller plus loin dans cette interopérabilité entre applications Web 2.0, on peut également considérer l'utilisation d'OpenID³⁸. Ce système d'authentification décentralisé permet de se connecter (sur les sites qui le supportent) avec un même login (en l'occurrence une URL) et un mot de passe unique là où il est en général nécessaire de créer un nouveau compte utilisateur. Un point intéressant est la manière dont OpenID et FOAF peuvent être connectés. D'une part, il est possible de lier une URL OpenID à un profil FOAF. Ceci peut se faire soit via un lien dans l'entête du document (X)HTML vers le profil FOAF, soit directement via l'inclusion du profil au sein du fichier avec RDFa ou eRDF. D'autre part, FOAF permet de définir l'URL OpenID d'un agent avec la propriété foaf:openid. En pratique, ce couplage peut être utilisé lorsqu'un utilisateur s'authentifie sur un site avec OpenID. On peut ainsi découvrir le profil FOAF associé puis récupérer à partir de là l'ensemble des données sociales de l'utilisateur connecté. C'est par exemple ce que nous avons mis en place au

³⁸<http://openid.org>

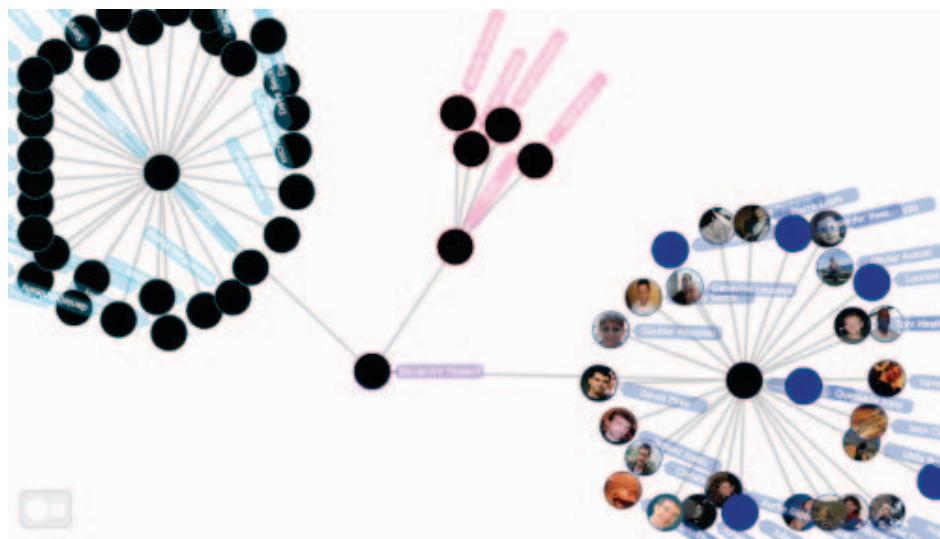


Figure 3.6: Visualisation uniforme de réseaux sociaux distribués

sein de SparqlPress³⁹, un *plug-in* pour WordPress permettant l'import et l'export de données RDF. Ici, lorsqu'un utilisateur se connecte, on affiche différentes informations le concernant (profils en ligne, nom, photo ...) sans aucune intervention de sa part⁴⁰ (Figure 3.7, page 100). Cette complémentarité entre initiatives communautaires et Web Sémantique nous paraît ainsi idéale pour répondre à ces différentes problématiques d'interopérabilité entre applications Web 2.0.

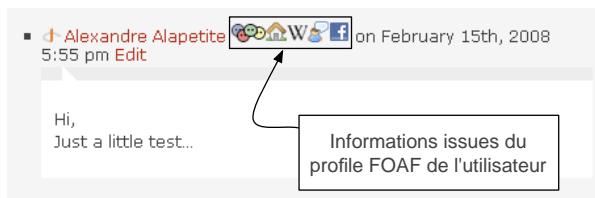


Figure 3.7: Utilisation combinée de FOAF et OpenID pour identifier un profil utilisateur avec SparqlPress

Des challenges importants restent selon nous à prendre en compte dans cette optique de données sociales ouvertes et interopérables, à savoir ceux de la protection des données privées et de la fragmentation volontaire d'identité entre services Web 2.0. À nouveau, nous pensons que le Web Sémantique a un rôle important à jouer par rapport à ces problématiques.

³⁹<http://wiki.foaf-project.org/SparqlPress>

⁴⁰<http://apassant.net/blog/2008/02/16/foaf-hacks-day>

D'une part, concernant la protection des données privées, nous pensons que cette ouverture des réseaux sociaux et des différents contenus créés ne signifie pas moins de protection mais au contraire, la possibilité de gérer plus finement les droits d'accès à ses données en ligne [Passant *et al.*, 2009b]. En effet, en offrant la possibilité de combiner les informations issues de plusieurs services, il est possible de définir des polices d'accès relativement pointues. Par exemple, en combinant des informations modélisées avec FOAF et SIOC depuis différents services, on peut imaginer un système de droits d'accès où l'utilisateur spécifie que son CV n'est accessible qu'à des personnes qui font partie à la fois de son réseau social LinkedIn et Twitter (utilisation de FOAF) et qui ont commenté au moins deux fois son blog (utilisation de SIOC). Puisque modélisées en RDF, ces polices d'accès peuvent également utiliser d'autres données présentes sur le Web Sémantique. On peut ainsi étendre la règle précédente en indiquant que la personne ne peut accéder à ce CV que si elle fait partie d'une entreprise considérée comme non-concurrente de celle de l'utilisateur, ce type d'information pouvant être extrait de DBpedia. Ces différentes pistes font en outre partie de travaux qu'il nous semble important d'approfondir par la suite dans ce contexte d'accès aux données sociales et de complémentarité entre Web Sémantique et Web 2.0 (Section 5.4.3, page 226).

D'autre part, si la fragmentation d'identité est le plus souvent un effet de bord de l'utilisation de différents services (en fonction du type de contenu à partager), il nous faut garder en mémoire qu'elle est parfois volontaire. Certaines personnes vont ainsi utiliser LinkedIn pour leurs contacts professionnels et MySpace pour leurs amis, ne souhaitant pas que les deux identités en ligne puissent être associées. Un rapport du cabinet Fabernovel rappelle en outre cette notion de fragmentation volontaire sur le Web⁴¹. Malgré tout, certains principes de raisonnement proposés par le Web Sémantique (notamment les propriétés inverses fonctionnelles que nous avons évoquées auparavant) vont conduire à cette fusion d'identité. Il est donc selon nous nécessaire de prendre en compte ces problématiques et de n'exposer certaines données (par exemple la propriété foaf : openId) qu'avec l'accord de l'utilisateur ou bien encore de prendre en compte des notions d'inférence avec autorité [Hogan *et al.*, 2008] en effectuant par exemple des raisonnements que si le demandeur fait partie du réseau social de l'utilisateur.

Plus généralement, ces problèmes relatifs à la protection de données sociales ne sont bien entendu pas seulement techniques et il est selon nous également nécessaire d'informer et d'éduquer les utilisateurs de services Web 2.0 afin de faire prendre conscience des risques possibles associés aux informations qu'ils dévoilent.

3.1.6 Adoption du modèle et évaluation

Etant donnés la nature et l'objectif de SIOC, il nous semble peu pertinent d'utiliser des métriques formelles comme celles recensées dans [Hartmann *et al.*, 2004] pour évaluer l'ontologie. Cependant, un point qui nous paraît important à prendre en compte est son acceptation sur le Web Sémantique. On peut ainsi parler d'évaluation par l'acceptation, proposition qui nous semble pertinente pour des ontologies de ce type qui ont pour but de devenir des modèles de référence pour la représentation de données sur le Web Sémantique. À partir

⁴¹<http://www.fabernovel.com/news/research-paper-social-network-websites/>

du moment où l'objectif de SIOC est de permettre une interopérabilité entre applications Web 2.0 et leur intégration au sein du Web Sémantique, on peut en effet considérer que la réussite d'un tel modèle repose sur le nombre de données ainsi représentées. En considérant logiquement l'ensemble de ces données interconnectées comme un graphe, il est évident que la valeur de ce graphe dépend du nombre de nœuds et d'arcs qui le composent, comme le rappelle la loi de Metcalfe⁴². Ainsi, plus le nombre de données représentées avec ce même modèle croît, plus la valeur inhérente de SIOC est importante. Cette observation sur la valeur d'un graphe est également valable pour le Web Sémantique dans son ensemble, notamment dans cet objectif de représentation et d'unification de données sociales comme le rappelle [Hendler et Golbeck, 2008].

En ce sens, on peut considérer l'adoption de SIOC comme un succès. *Ping The Semantic Web*⁴³ (service sur lequel nous reviendrons plus tard (Section 5.1.3, page 192)) recense plus de 127000 documents utilisant le noyau de SIOC et plus de 115000 utilisant son module Types⁴⁴, ce qui en font respectivement les quatrième et cinquième espaces de nom les plus utilisés. La figure qui-suivit indique également un nombre croissant de données ainsi représentées sur le Web Sémantique (Figure 3.8, page 102), toujours d'après ce même service.

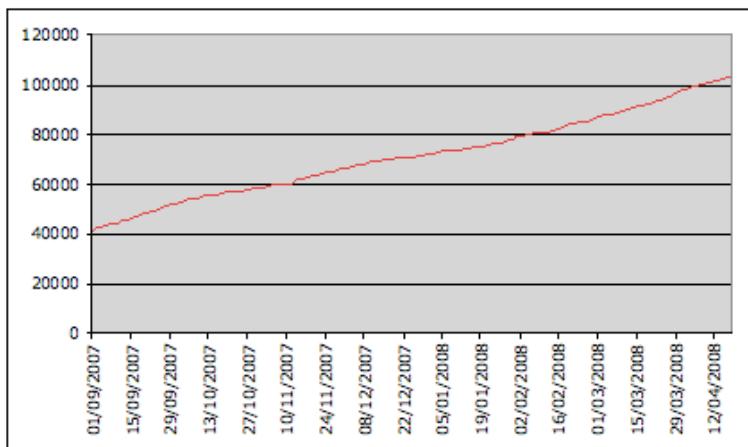


Figure 3.8: Statistiques de production de données SIOC sur le Web [Bojārs *et al.*, 2008b]

De plus, il nous semble important de considérer cette acceptation non pas uniquement en nombre de données, mais en observant la diversité des systèmes utilisant SIOC et l'activité de la communauté associée. Ainsi, la Soumission Membre de SIOC au W3C en Juin 2007 a réuni plus d'une dizaine d'organisations différentes, aussi bien académiques (LaLIC, DERI

⁴²Selon Bob Metcalfe, co-inventeur du protocole Ethernet, la valeur d'un réseau s'accroît avec son nombre de connections et est proportionnelle au carré du nombre de ses utilisateurs. – http://en.wikipedia.org/wiki/Metcalfe%27s_law

⁴³<http://pingthesemanticweb.com>

⁴⁴Janvier 2009, cf. <http://pingthesemanticweb.com/stats/namespaces.php> pour une mise à jour.

Galway, Fundation CTIC⁴⁵ ...) qu'industrielles (Opera Software⁴⁶, OpenLink⁴⁷ ...). Cette soumission comporte en outre trois documents qui servent aujourd'hui de référence à SIOC et pour lesquels nous avons eu à chaque fois un rôle particulier :

- *SIOC Core Ontology Specification* [Berrueta *et al.*, 2007], spécification du cœur de l'ontologie SIOC (coauteur) ;
- *SIOC Ontology : Applications and Implementation Status* [Fernández *et al.*, 2007b], document listant un ensemble d'applications utilisant SIOC au moment de la soumission (coéditeur) ;
- *SIOC Ontology : Related Ontologies and RDF Vocabularies* [Bojārs *et al.*, 2007a], document listant les relations entre SIOC et d'autres ontologies populaires sur le Web Sémantique comme FOAF ou DublinCore (coéditeur).

Elle a de plus a été favorablement reçue, comme en témoignent les commentaires du W3C à son égard⁴⁸ "SIOC has the potential to become one of the foundational vocabularies that make Semantic Web applications useful, alongside DOAP, FOAF, Dublin Core, etc." ou "The SIOC vocabulary is a useful component of the Semantic Web", malgré certaines remarques sur l'absence de considérations relatives au respect de la vie privée, sujet que nous avons évoqué auparavant.

SIOC est considéré aujourd'hui comme une brique fondamentale du *Social Semantic Web* et plus d'une cinquantaine d'applications l'utilisant sont aujourd'hui disponibles, la plupart étant des applications *open-source* (Section 4.1, page 138). À nouveau, c'est selon nous la légèreté d'un tel modèle qui a permis une telle acceptation, celui-ci pouvant être simplement appréhendé, du moins dans ses termes principaux. Concernant ces outils, si les premiers ont logiquement été développés par des membres actifs de la communauté SIOC, on trouve aujourd'hui une importante diversité concernant les partenaires et les domaines d'applications utilisant SIOC. Ainsi, Yahoo! SearchMonkey⁴⁹, moteur de recherche tirant bénéfices des annotations sémantiques disponibles sur le Web pour enrichir la présentation de ses résultats, suggère l'utilisation de SIOC comme modèle de référence pour représenter les activités Web 2.0 sur le Web Sémantique⁵⁰. Un autre exemple pertinent est selon nous l'utilisation de SIOC dans des contextes autres que celui d'applications purement Web 2.0, en plus du cas d'utilisation que nous décrivons dans cette thèse. Par exemple, un de nos récents efforts se concentre autour du projet SWANSIOC⁵¹, qui vise à intégrer les vocabulaires SIOC et SWAN – *Semantic Web Applications in Neuromedicine* [Ciccarese *et al.*, 2008] – dans un objectif de représentation du discours scientifique et des argumentations associées autour du traitement de la maladie d'Alzheimer. Ce projet mené dans le cadre du groupe d'intérêt *Health Care and Life Science* du W3C⁵² montre bien selon nous le potentiel que peut avoir SIOC pour des secteurs non-relatifs au Web 2.0 mais où l'aspect social prédomine. Pour plus de détails sur SIOC, on pourra se référer à la thèse [Bojārs, 2009]

⁴⁵<http://www.fundacionctic.org/>

⁴⁶<http://www.opera.com/>

⁴⁷<http://www.openlinksw.com/>

⁴⁸<http://www.w3.org/Submission/2007/02/Comment>

⁴⁹<http://developer.yahoo.com/searchmonkey/>

⁵⁰http://developer.yahoo.com/searchmonkey/smguide/profile_vocab.html

⁵¹<http://esw.w3.org/topic/HCLSIG/SWANSIOC>

⁵²<http://www.w3.org/2008/05/HCLSIGCharter>

3.2 MODÉLISATION DES ONTOLOGIES MÉTIER

3.2.1 Besoins en termes de représentation métier

Étant donné le contexte du projet Athéna, le niveau de représentation que nous souhaitons atteindre au sein de notre système de médiation doit nous permettre de modéliser des assertions comme :

- Électricité de France est une entreprise française du secteur de l'énergie ;
- l'énergie solaire est une énergie renouvelable ;
- les énergies des marées et les énergies houlomotrices sont deux types d'énergies marines ;
- Pierre Gadoneix est le président d'Electricité de France ;
- EDF a différents partenaires autour des énergies renouvelables.

Les besoins de modélisation métier se situent donc principalement autour des acteurs (au sens personnes physiques et morales), de leurs domaines d'activité et des propriétés associées (relations entre ces entités, localisation ...). En conséquence, cette partie du mémoire est sans doute celle où les travaux présentés auront le plus d'adhérence avec les besoins exprimés par l'entreprise. En effet, alors que les autres modèles définis dans ce chapitre (SIOC présenté précédemment et MOAT par la suite) ont été conçus de manière générique, la modélisation d'ontologies métier implique un rapprochement avec les domaines abordés par l'entreprise. Cependant, si ces modèles ont été conçus de manière *ad hoc*, il nous semble utile de les détailler notamment parce que nous avons été confrontés à des choix de modélisation qu'il nous paraît intéressant d'argumenter et de partager. Ceux-ci ont en effet influencé la modularité et la réusabilité des modèles développés et nous semblent être de bonnes pratiques quant au développement d'ontologies légères.

Comme nous allons le voir, nous avons fait le choix de modéliser un certain nombre d'ontologies interconnectées plutôt que de proposer une unique ontologie globale permettant de modéliser des choses aussi diverses que des zones géographiques, des secteurs d'activité ou des types d'agents. En effet, ce choix d'ontologies légères (tout comme nous l'avons explicité en présentant SIOC auparavant) nous semble plus pertinent dans une optique de réutilisation des modèles dans d'autres contextes mais aussi dans une optique de passage à l'échelle de certaines de nos propositions. De plus, comme nous le verrons, nous avons fait le choix dans certains cas d'étendre des modèles existants, ce qui nous a d'une part permis de bénéficier de l'existant mais aussi de permettre à la communauté de bénéficier de nos réflexions en termes de nouvelles classes ou propriétés (Section 3.2.3, page 107). Nous allons maintenant détailler les différents modèles mis en œuvre dans notre contexte.

3.2.2 FOAF pour la représentation des personnes physiques et morales

Concernant la représentation des personnes physiques et morales, nous avons considéré différents modèles avant d'établir notre choix de départ. Tout d'abord, des ontologies génériques comme Proton [Terziv et al., 2005] (et notamment ses modules Upper⁵³ et Top⁵⁴), Cyc

⁵³<http://proton.semanticweb.org/2005/04/protonu>

⁵⁴<http://proton.semanticweb.org/2005/04/protont>

[Lenat *et al.*, 1990] et son équivalent *open-source* OpenCyc⁵⁵, O'CoMMA⁵⁶ (associée au projet CoMMA [Gandon, 2002]) ou Yago [Suchanek *et al.*, 2007], plus récente, voire même à un certain niveau Wordnet [Fellbaum, 1998] et sa représentation RDF/OWL⁵⁷ pour sa taxonomie de classes. Nous avons rapidement mis ces choix de côté pour deux raisons majeures :

- du fait de leur caractère général (*i.e.* couvrant un large spectre de domaines) et de leur abondance de classes (jusqu'à plusieurs centaines) et de propriétés, celles-ci sont délicates à aborder. C'est ici tout le problème de la modularité des ontologies et de l'équilibre entre usabilité et réusabilité qui se pose [Klinker *et al.*, 1991]. Il s'agit donc de trouver le juste milieu entre un modèle complet mais trop complexe et un modèle plus léger et réutilisable ;
- dans leurs hiérarchies de classes, ces modèles considèrent généralement le rôle joué par une entité comme une sous-classe de l'entité elle-même, par exemple Student est définie en tant que sous-classe de Person dans O'CoMMA, ou Bank de Entreprise dans Proton (Figure 3.9, page 106). Or d'un point de vue du formalisme logique mais aussi de l'utilisation et l'évolution du modèle, il nous semble plus pertinent de considérer la notion de rôle comme un concept indépendant de l'entité à laquelle il est rattaché. Nous détaillerons plus loin notre approche à ce sujet (Section 3.2.4, page 109).

En conséquence, nous avons considéré des modèles plus légers, focalisés essentiellement sur ces notions d'agents, en particulier de personnes et de groupes. Nous avons donc étudié la *Portal Ontology*⁵⁸ d'AKT - Université de Southampton (encore trop riche pour nos besoins), la *Person Ontology*⁵⁹ d'eBiquity - Université du Maryland (UMBC) ou encore SWRC – *Semantic Web Research Community* [Sure *et al.*, 2005] –, ces deux dernières étant plus adaptées à la modélisation du monde universitaire. Nous avons finalement considéré FOAF – *Friend Of A Friend* [Brickley et Miller, 2004b] – comme modèle de base pour la description de ce domaine.

FOAF a pour objectif de représenter la notion d'agent (foaf:Agent) et de différentes sous-classes liées : personnes (foaf:Person), groupes d'agents (foaf:Group) et organisations (foaf:Organization), ainsi qu'un certain nombre de propriétés associées à ces concepts : nom (foaf:name), accointance (foaf:knows), appartenance (foaf:member)... Comme évoqué en amont, une des raisons de ce choix est due à la simplicité du modèle et le fait qu'il se concentre essentiellement sur la notion d'agent sans s'étendre sur des aspects complémentaires comme les rôles. Ceci nous permet de disposer d'un noyau simple et extensible sans s'encombrer d'une structure ontologique trop riche. Une autre raison de ce choix est l'utilisation abondante de ce vocabulaire sur le Web⁶⁰. Si FOAF a longtemps été utilisé essentiellement pour définir des profils personnels, son intégration comme ontologie de référence au projet *Linking Open Data* (comme le suggèrent [Bizer *et al.*, 2007b])

⁵⁵<http://www.opencyc.org/>

⁵⁶<http://pauillac.inria.fr/cdrom/ftp/ocomma/comma.rdf>

⁵⁷<http://www.w3.org/2001/sw/BestPractices/WNET/wn-conversion.html>

⁵⁸<http://www.aktors.org/ontology/portal>

⁵⁹<http://ebiquity.umbc.edu/ontology/person.owl>

⁶⁰Une étude menée en Août 2006 sur les données stockées par Swoogle [Ding *et al.*, 2004] a montré que FOAF était le quatrième espace de noms le plus utilisé. – <http://ebiquity.umbc.edu/resource/html/id/196/Most-common-RDF-namespaces>

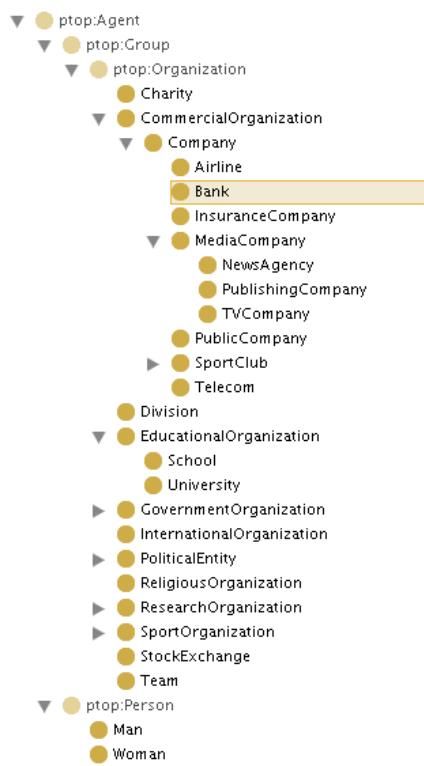


Figure 3.9: Taxonomie des sous-classes d'Agent dans Proton

en fait un vocabulaire que l'on retrouve désormais fréquemment pour la représentation de données relatives aux agents sur le Web Sémantique. Par exemple, les personnalités de DBpedia sont déclarées en tant qu'instances de foaf:Person et utilisent des propriétés comme foaf:name ou foaf:depiction (cf. http://dbpedia.org/resource/Albert_Einstein). De ce fait, utiliser ce vocabulaire en interne nous permet d'utiliser les mêmes outils pour les données produites au sein de notre écosystème que pour les données agrégées depuis l'extérieur (Section 4.2.4, page 163).

Tel quel, FOAF permet de modéliser des assertions comme "*Alexandre Passant, personne, est membre du LaLIC, organisation*" ou bien "*Electricité De France est une organisation*" mais ne permet pas de prendre en compte d'autres notions qui nous intéressent comme "*Electricité de France est une entreprise et a pour acronyme EDF*" ou bien encore "*Le LaLIC est basé à Paris*". Nous avons donc étendu FOAF au sein d'un modèle OWL-DL que nous avons nommé FOAFplus et introduisant différentes classes permettant de prendre en compte ces spécificités (Listing 3.6, page 107). Cette utilisation de FOAF nous a également permis de suggérer des évolutions du modèle en termes de taxonomie de classes et de domaine et codomaine de certaines propriétés⁶¹.

⁶¹<http://lists.foaf-project.org/pipermail/foaf-dev/2007-January/008396.html>

```

foafplus:Company rdf:type owl:Class ;
  rdfs:subClassOf foaf:Organization .
foafplus:ResearchInstitute rdf:type owl:Class ;
  rdfs:subClassOf foaf:Organization .
foafplus:Institution rdf:type owl:Class ;
  rdfs:subClassOf foaf:Organization .
foafplus:Association rdf:type owl:Class ;
  rdfs:subClassOf foaf:Organization .

foafplus:acronym rdf:type owl:DatatypeProperty ;
  rdfs:domain foaf:Agent ;
  rdfs:range rdfs:Literal .

```

Listing 3.6: Extension de FOAF pour la gestion de différents types d'agents

Afin de modéliser les relations entre entreprises plus finement qu'avec la simple relation `foaf:knows`, nous avons également introduit un modèle léger pour représenter la notion de partenariat entre différentes entités autour d'un domaine donné. Une classe `Partenariat` permet donc de représenter une relation entre différents agents autour d'un domaine donné, représenté via la classe (`role:Domain`) sur laquelle nous reviendrons par la suite (Section 3.2.4, page 109). Notons que cette classe peut simplement s'aligner avec la classe `Relationship` du vocabulaire du même nom⁶², nous conduisant ainsi au modèle suivant (Listing 3.7, page 107).

```

part:Partenariat rdf:type owl:Class .
  rdfs:subClassOf relationship:Relationship .

part:hasMember rdf:type owl:ObjectProperty ;
  rdfs:domain part:Partenariat ;
  rdfs:range foaf:Agent .
part:hasDomain rdf:type owl:ObjectProperty ;
  rdfs:domain part:Partenariat ;
  rdfs:range role:Domain .

```

Listing 3.7: Modélisation de partenariats entre agents

3.2.3 Localisation avec Geonames

Avant d'évoquer la notion de rôle associée aux agents, un autre aspect qui nous intéresse est leur localisation. Il est en effet pertinent de pouvoir localiser ceux-ci, par exemple pour étudier l'émergence d'une technologie sur un domaine donné ou identifier géographiquement le réseau (membres ou partenaires) gravitant autour d'un acteur.

Une première possibilité est l'utilisation de la propriété `foaf:based_near` proposée par FOAF. Celle-ci permet de lier deux instances de `SpatialThing` du vocabulaire `Geo`

⁶²<http://vocab.org/relationship/>

Vocabulary⁶³ [Brickley, 2003] proposé par le groupe d'intérêt Web Sémantique du W3C et basé sur la spécification *World Geodetic System 1984*. L'utilisation de cette propriété permet de modéliser une relation entre une instance de `foaf:Agent` (ou sous-classe) et un simple point (`geo:Point`) associé à ses coordonnées de latitude et de longitude. L'exemple qui suit modélise de cette manière qu'EDF est basé à Paris (Listing 3.8, page 108).

```
athena:EDF a foafplus:Company ;
  foaf:based_near [
    a geo:Point ;
    geo:lat "48,5144" ;
    geo:long "2,213" .
]
```

Listing 3.8: Localisation d'une entreprise avec FOAF et le Geo Vocabulary

Modéliser la géolocalisation des acteurs de cette manière pose deux principaux problèmes dans notre contexte :

- les points ainsi définis sont en général des nœuds anonymes, *i.e.* n'ont pas d'URI propre. Ceci complexifie les requêtes SPARQL destinées à identifier des entités localisées en un lieu donné. Il est en effet nécessaire d'utiliser une clause FILTER basée sur des valeurs de coordonnées pour identifier que deux agents sont basés au même endroit, ce type de requête est généralement plus complexe qu'une simple comparaison d'URIs. De plus, dans le cas où les coordonnées ne sont pas identiques, il est nécessaire d'utiliser un moteur SPARQL gérant l'imprécision ou la logique floue [Pan *et al.*, 2008] ou bien de déléguer cette comparaison à une application externe pour permettre cette identification ;
- l'utilisation directe de coordonnées rend complexe l'annotation sémantique, obligeant à indiquer explicitement celles-ci dans les outils produisant ces annotations. Si cela est envisageable pour des spécialistes de bases de données géographiques ou des personnes dont c'est le cœur de métier, ça l'est plus difficilement dans notre contexte, les utilisateurs étant plus prompts à utiliser simplement des noms de zones géographiques pour localiser les entités.

Fin 2005, le projet Geonames a vu le jour avec pour objectif de fournir une base de données géographique de référence sous licence Creative Commons, comptant aujourd'hui plus de six millions d'entités⁶⁴. Le point qui nous intéresse particulièrement ici est l'intégration de celle-ci au sein du Web Sémantique à la mi-octobre 2006⁶⁵. En particulier, les points suivants ont retenu notre attention :

- la définition d'une ontologie⁶⁶ définissant la notion de zone géographique avec une classe `geonames:Feature` ;

⁶³Espace de noms http://www.w3.org/2003/01/geo/wgs84_pos#, préfixe `geo` par la suite.

⁶⁴<http://geonames.org/about.html>

⁶⁵<http://geonames.wordpress.com/2006/10/14/semantic-web/>

⁶⁶Espace de noms <http://www.geonames.org/ontology#>, préfixe `|geonames|` par la suite.

- la mise à disposition d'URIs pour identifier chaque zone, et surtout l'association à chaque URI – déréférençable – de la description RDF de l'entité, notamment ses coordonnées avec le *Geo Vocabulary* défini précédemment ;
- la définition de relations entre entités, en particulier la présence d'une propriété pour indiquer le parent immédiat d'une zone donnée (`geonames:parentFeature`) ;
- la place de plus en plus importante de Geonames au sein du projet *Linking Open Data*, notamment son intégration avec DBpedia et de fait sa mise en avant naturelle comme ontologie et base de connaissances de référence pour la localisation d'entités sur le Web Sémantique.

Il nous donc a paru pertinent d'utiliser ce modèle dans notre contexte pour représenter la géolocalisation des différents agents. Nous pouvons ainsi bénéficier de la base de connaissance Geonames et de son service web⁶⁷ pour simplement produire des annotations RDF relatives à la localisation de différents concepts (Section 4.2.4, page 163). Si la propriété `foaf:based_near` peut-être envisagée pour lier chaque entité à une instance de `geonames:Feature`, sa sémantique est assez faible puisqu'elle indique simplement "We do not say much about what 'near' means in this context; it is a 'rough and ready' concept"⁶⁸. Nous avons ainsi proposé l'ajout d'une relation `locatedIn` permettant d'indiquer qu'une ressource est située dans une zone géographique précise (Listing 3.9, page 109). Celle-ci a été ajoutée au modèle Geonames dans sa version 2.0 d'Avril 2007⁶⁹.

```
geonames:locatedIn rdf:type owl:ObjectProperty ;
  rdfs:domain rdfs:Resource ;
  rdfs:range geonames:Feature .
```

Listing 3.9: Définition de la propriété `locatedIn` de Geonames

Un autre avantage quand à l'utilisation de ce modèle est la transitivité de la relation `geonames:parentFeature`. La figure suivante (Figure 3.10, page 110) représente ainsi (1) des relations entre des agents et leur zone géographique pouvant être définies au sein de notre système, (2) des relations entre zones géographiques modélisées au sein de la base de connaissances Geonames et (3) une des relations inférées par transitivité. Il est donc possible d'identifier que deux acteurs sont basés dans une zone similaire (par exemple un même continent) même si l'annotation au sein du système spécifie des localisations à un niveau de finesse plus précis (pays ou ville). Dans l'exemple qui suit, on peut donc inférer qu'EDF et Gazprom sont basés (dans une zone située) en Europe à partir du fait qu'EDF est localisée à Paris et Gazprom en Russie.

3.2.4 Ontologies des rôles et utilisation de SKOS

Notion de rôles associés aux agents

Après cet aparté géographique, revenons sur la notion de rôle associé aux différents acteurs. Un de nos besoins est en effet de représenter les différentes activités de ceux-ci,

⁶⁷<http://geonames.org/export>

⁶⁸http://xmlns.com/foaf/spec/#term_based_near

⁶⁹<http://lists.w3.org/Archives/Public/public-xg-geo/2007Jan/0001.html>

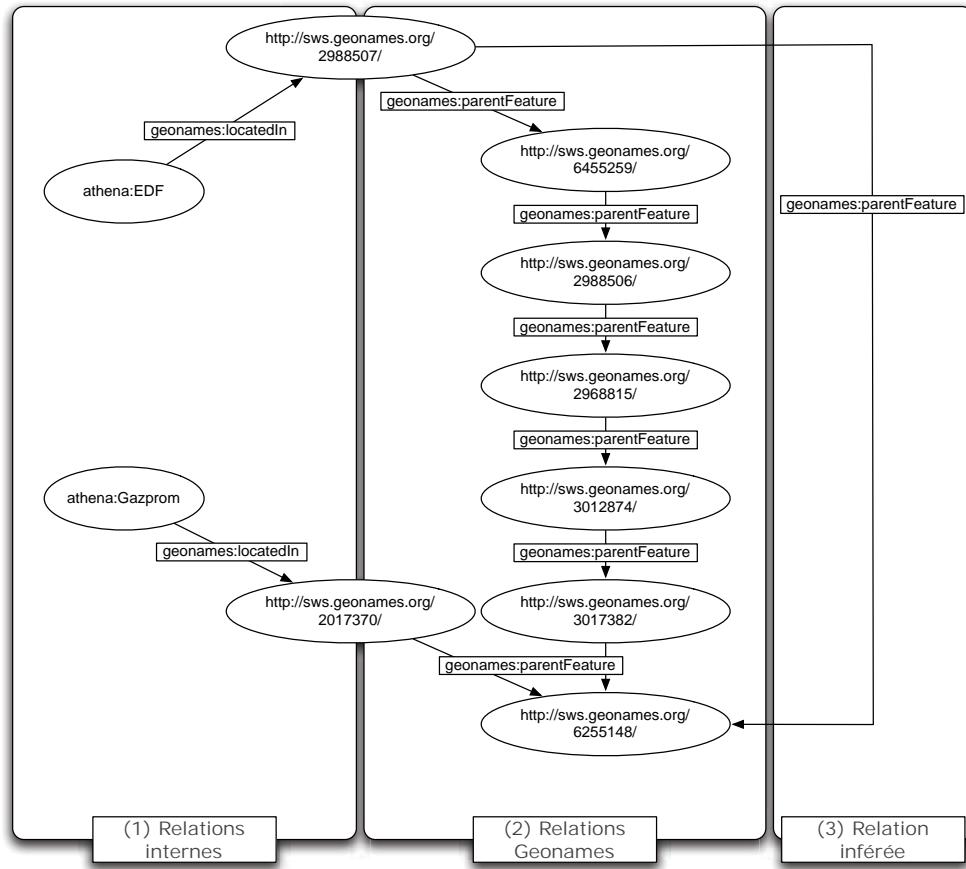


Figure 3.10: Relations géographiques entre entités et transitivité de la propriété `locatedIn` de Geonames

par exemple indiquer que telle entreprise est productrice d'énergies marines en Angleterre ou que telle autre commercialise des panneaux solaires en France. Nous considérons ces différentes activités comme des rôles associés aux agents et non pas comme la nature même de ceux-ci. Cette vision reprend les principes définis par [Sowa, 1984] qui distingue dans sa définition des réseaux sémantiques les *types naturels* – qui sont liés à l'essence même des entités – et les *rôles* – qui dépendent d'une relation accidentelle avec d'autres entités. En effet, contrairement au fait d'être une entreprise ou une personne, statut qui fait partie de l'essence même de l'entité en question, produire des énergies marines dépend d'un certain contexte (autres acteurs, marché industriel ...). De plus, le rôle – contrairement au type naturel – est considéré comme anti-rigide [Welty et Guarino, 2001], une entreprise pouvant changer de domaine d'activité tout en restant la même entreprise, *i.e.* la même entité. Pour exemplifier les propos précédents, on considère ainsi que la notion de personne fait parti de l'essence même d'un individu, mais que des notions aussi diverses que celles d'étudiant, de chercheur ou de mari sont des rôles que cet individu peut jouer à un moment ou un autre de son

existence sans pour autant que sa nature change.

Contrairement à [Sowa, 1984] qui considère d'un point de vue du modèle les rôles comme des sous-classes des types naturels, nous préférons la vision de [Guarino, 1992] qui définit ceux-ci comme des entités indépendantes et associées aux types naturels avec des propriétés dédiées. On retrouve cette modélisation notamment dans DOLCE [Claudio *et al.*, 2005] avec la notion de rôles fonctionnels qui se rapprochent des nôtres (producteur d'énergie, ingénieur ...). Nous avons pu également pu nous rendre compte d'un point de vue plus pratique en utilisant Proton (qui suit l'idée de [Sowa, 1984] en considérant les rôles comme des sous-classes) que les modèles RDF(S)/OWL de ce type entraînent des relations taxonomiques assez complexes dès lors qu'on les étend pour permettre à des types d'entités différents de jouer un même rôle. On se retrouve en effet confronté à des treillis de classes complexes, sujets à explosion combinatoire, l'ajout d'un rôle particulier qui peut être assigné à n classes distinctes entraînant la création de n nouvelles classes. Nous avons ainsi volontairement limité le nombre de types naturels dans notre modèle d'acteurs (comme défini précédemment) pour concentrer les rôles dans un modèle indépendant. Ainsi, nous avons tout d'abord défini un modèle très léger pour la définition des rôles (préfixe `role`) comprenant une simple classe `Role` et une propriété permettant de faire le lien entre un agent et ses différents rôles (Listing 3.10, page 111).

```
role:Role rdf:type owl:Class .

role:hasRole rdf:type owl:ObjectProperty ;
  rdfs:domain foaf:Agent ;
  rdfs:range role:Role .
```

Listing 3.10: Modèle simple pour la représentation des rôles

Une fois ce premier modèle établi, la modélisation du rôle en lui-même a demandé à nouveau réflexion. Une première possibilité est de considérer les rôles comme des concepts à part entière, par exemple `athena:ProdEnMarAng`⁷⁰ ou `athena:ComPanSolFra`. Ce choix implique également la définition d'un treillis de concepts complexe si l'on souhaite pouvoir identifier des acteurs associés à des thématiques similaires. Il est en effet nécessaire de modéliser que `athena:ProdEnMarAng` est lié à la fois à `athena:EnMarAng`, `athena:ProdEnMar` et `athena:ProdAng`. Plutôt que partir dans cette direction, nous avons choisi de considérer la classe rôle (`role:Role`) comme modélisant un triptyque entre⁷¹ :

- un type de métier, par exemple *Producteur*, modélisé avec une classe `role:Type` ;
- un domaine, par exemple *Energies Marines*, modélisé avec une classe `role:Domain` ;
- et une zone géographique, logiquement représentée avec `geonames:Feature` et la propriété `geonames:locatedIn`. Elle n'est donc pas représentée dans la description

⁷⁰URI fictive pour définir le concept de *Producteur d'Energies Marines en Angleterre*, nous ne détaillerons pas les autres URIs de cet exemple qui suivent le même principe.

⁷¹Malgré leur nom, ces différentes classes n'ont aucun lien avec les notions de type et de domaine que l'on retrouve en RDFS et OWL.

qui suit puisque le domaine (`rdfs:domain`) de `geonames:locatedIn` est volontairement ouvert et n'est donc pas sujet à redéfinition.

Le modèle précédent (Listing 3.10, page 111) se redéfinit donc de la manière suivante (Listing 3.11, page 112). En conséquence, notre exemple précédent représentant le fait qu'une entreprise soit productrice d'Energies Marines en Angleterre se traduit en RDF comme suit (Listing 3.12, page 112).

```
role:Role rdf:type owl:Class .
role:Domain rdf:type owl:Class .
role:Type rdf:type owl:Class .

role:hasRole rdf:type owl:ObjectProperty ;
  rdfs:domain foaf:Agent ;
  rdfs:range role:Role .
role:hasDomain rdf:type owl:ObjectProperty ;
  rdfs:domain foaf:Role ;
  rdfs:range role:Domain .
role:hasType rdf:type owl:ObjectProperty ;
  rdfs:domain foaf:Role ;
  rdfs:range role:Type .
```

Listing 3.11: Modèle pour la représentation des rôles avec prise en compte du métier et du domaine

```
athena:entreprise1 a foafplus:Company ;
  role:hasRole [
    role:hasDomaine athena:EnergiesMarines
    role:hasType athena:Production
    geonames:locatedIn <http://sws.geonames.org/6269131/>
  ] .
```

Listing 3.12: Association d'un rôle à un agent

Pour en revenir à la modélisation des rôles au sens large, [Fukazawa *et al.*, 2006] ont également montré qu'il était parfois nécessaire de prendre en compte le contexte social d'un rôle : famille, loisirs, travail, etc. Si nous n'avons pas pris en compte cette contextualisation dans notre modèle, on peut considérer que par défaut, tous nos rôles se situent dans un contexte de *rôle industriel*. Par ailleurs, pour une analyse plus complète de la littérature sur cette notion de rôle en Ingénierie des Connaissances, on pourra consulter [Steimann, 2000]. Notons aussi que si nous ne les avons pas pris en compte, d'autres modèles plus légers peuvent être considérés pour modéliser les rôles associés à des agents, comme par exemple DOAC⁷² – Description Of A Career -, ou le vocabulaire ResumeRDF⁷³ [Bojārs et Breslin, 2007].

⁷²<http://ramonantonio.net/doac/>

⁷³<http://rdfs.org/resume-rdf/>

Organisation des différents domaines et métiers associés aux rôles

Si nous utilisons Geonames pour la localisation d'un rôle, le problème d'organiser entre eux les domaines industriels et les métiers reste ouvert, toujours dans cette optique d'établir à terme des relations entre acteurs à partir de leurs activités. Nous avons ainsi décidé d'organiser hiérarchiquement ces informations et d'établir deux hiérarchies distinctes de domaines et métiers. Pour justifier ce choix de hiérarchie de concepts (avec uniquement une relation de subsomption) et non pas d'ontologie plus précise (qui ferait par exemple la distinction entre une relation `utiliseLeMateriau` et `permetDeProduire`), il est important de comprendre que nous modélisons des domaines d'activité ou d'expertise (notions abstraites) et non pas les objets physiques en eux-mêmes (notions concrètes). Nous faisons ainsi la distinction entre l'objet panneau solaire et le domaine des panneaux solaires. Ainsi, le fait qu'une entreprise commercialise des panneaux solaires en France se traduira par "*a un rôle de commercialisation dans le domaine des panneaux solaires*". C'est bien une relation de subsomption classique qui peut exister entre les domaines, indiquant par exemple que le domaine des panneaux solaires est plus spécifique que celui de l'énergie solaire. Si l'on s'était attaché au contraire à modéliser l'objet panneau solaire, il aura fallu une relation autre qu'un lien hiérarchique entre celui-ci et la notion d'énergie solaire (e.g. `permetDeProduire`).

Cette représentation nous permet donc de modéliser ces domaines au niveau d'un modèle taxonomique comportant une unique classe (`Domain`) et une seule relation (*plus spécifique que*) plutôt qu'une ontologie plus poussée avec différentes classes et propriétés comme le montre la figure qui suit avec cet exemple de panneaux solaires et d'énergie solaire (Figure 3.11, page 114)⁷⁴. La production des annotations associées à ce modèle étant en outre laissée à discréption des utilisateurs via l'utilisation de wikis sémantiques (Section 4.2.4, page 160), c'est une autre raison qui nous a amené à utiliser un modèle simple avec une unique relation pour structurer ces domaines et métiers.

À partir de cette vision, une première manière d'organiser ces domaines et métiers est naturellement de penser à une taxonomie de classes définies en tant qu'`owl:Class` et organisées avec `rdfs:subClassOf` sous les classes principales `role:Domain` et `role:Type`. Par exemple, on peut considérer la classe `role:PanneauxSolaires` comme une sous-classe de `role:EnergieSolaire`, elle-même sous-classe de `role:ENR`, à son tour sous-classe de `role:Domain`. Cependant, cette classe `role:PanneauxSolaires` sera également considérée comme étant instance de `Domain` du fait de l'utilisation de la propriété `role:hasDomain` pour associer ce domaine à un rôle et du codomaine de cette propriété (*i.e.* `role:Domain`) (Figure 3.12, page 114). Nous basculons alors dans le dialecte OWL-Full ce que nous ne souhaitons pas pour des raisons d'indécidabilité⁷⁵. Même si notre architecture n'utilise pas pour le moment de raisonneur OWL, nous préférions nous assurer que le modèle ne devra pas être repensé pour cette raison dans le futur.

Une autre solution est de toujours considérer une taxonomie de classes mais d'associer à chacune une instance de référence qui sera utilisé pour la modélisation des rôles au niveau des assertions, distinguant ainsi classes et instances et permettant de rester à un niveau OWL-Lite (ou OWL-DL en fonction des autres axiomes de l'ontologie) (Figure 3.13,

⁷⁴Il en est de même pour la représentation des métiers avec la classe `Type`.

⁷⁵Rappelons que nous évoquons uniquement OWL1 dans ce mémoire.

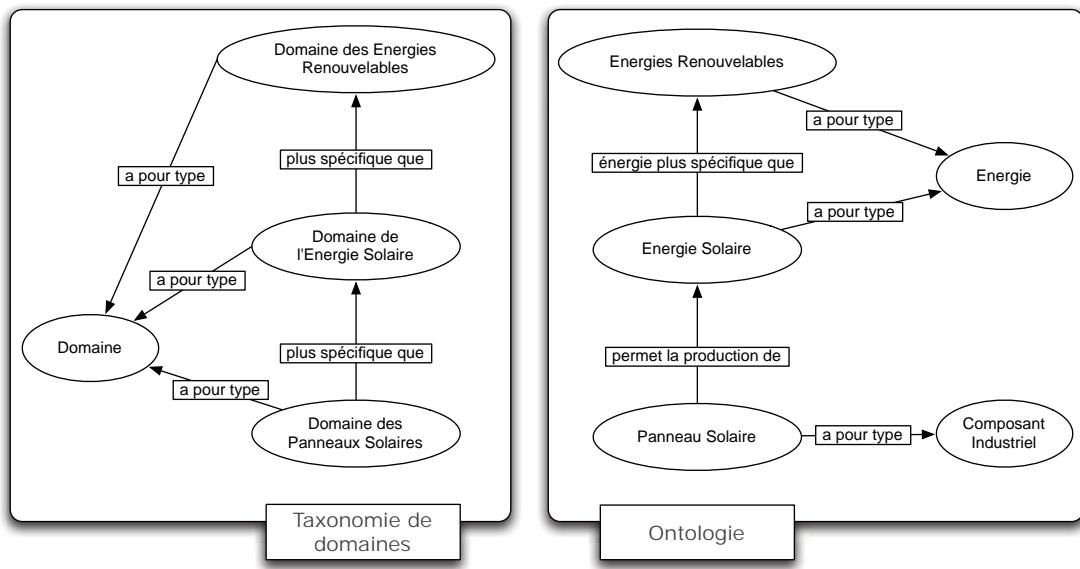


Figure 3.11: Distinction entre taxonomies et ontologies

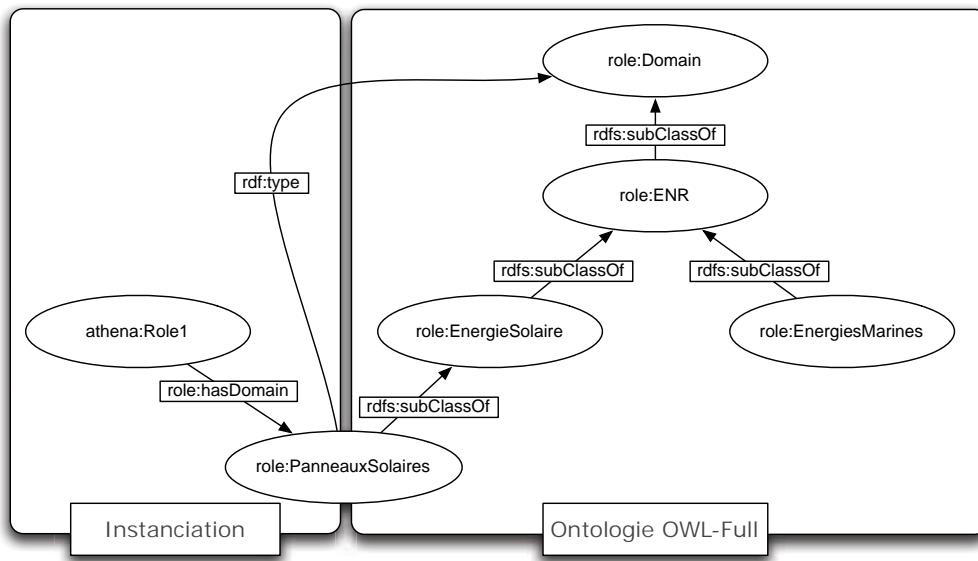


Figure 3.12: Taxonomies de domaines en OWL-Full

page 115). Ceci complique cependant inutilement le modèle et introduit des relations supplémentaires qui alourdissent les requêtes, puisque le parcours de graphe se complexifie avec ce noeud supplémentaire.

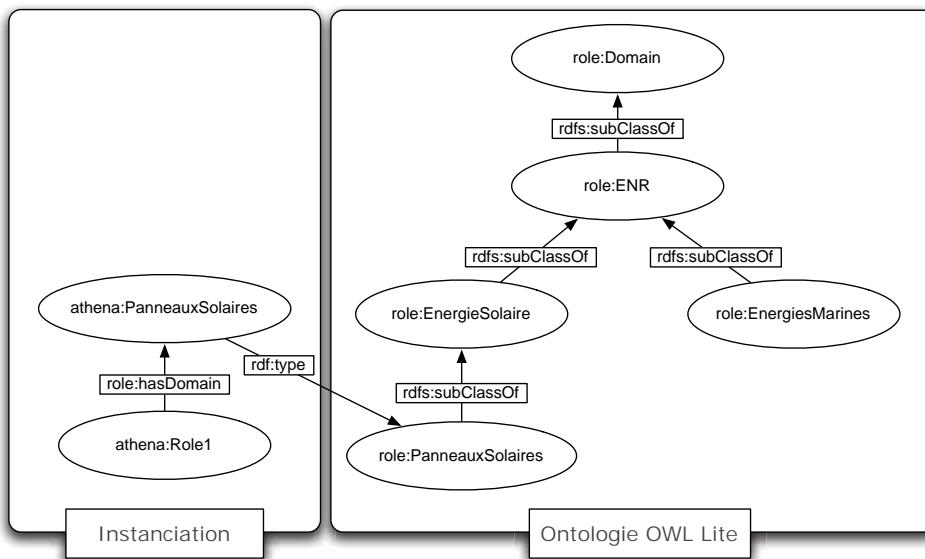


Figure 3.13: Taxonomies de domaines en OWL-Lite

Enfin, une solution est d'utiliser non pas une taxonomie de classes représentée au sein d'un modèle OWL classique, mais de se baser sur SKOS. SKOS – Simple Knowledge Organisation Schema [Miles et Bechhofer, 2008] – permet en effet de définir une *hiérarchie d'instances*, au sens organisation taxonomique d'instances et non plus de classes. Plus exactement et comme nous l'avons brièvement explicité auparavant, SKOS définit une classe `skos:Concept` et considère les relations `skos:narrower` et `skos:broader` (sous-propriétés d'une relation plus générique `skos:semanticRelation`) ainsi qu'une relation `skos:related` pour établir des liens entre différentes instances de cette classe. L'objectif de SKOS est ainsi de permettre la définition sur le Web Sémantique de modèles de représentation des connaissances plus légers que des ontologies comme des thesaurus ou des taxonomies. La sémantique des relations proposées par SKOS est en effet volontairement faible là où des ontologies plus poussées vont typer et distinguer différentes relations comme nous l'avons montré dans une figure précédente (Figure 3.11, page 114). De plus, les relations hiérarchiques proposées par SKOS ont une sémantique différente de celles proposées par RDF-S/OWL puisque l'on se situe au niveau des instances et non plus des classes. Ceci permet donc dans notre contexte de définir qu'une entreprise est active dans un domaine considéré comme plus spécifique qu'une autre en modélisant uniquement des relations entre instances et ce sans basculer dans un niveau d'expressivité OWL-Full (Figure 3.12, page 114) ni introduire des instances associées à chaque classe (Figure 3.13, page 115).

Ainsi, nous avons utilisé SKOS pour modéliser les notions de domaine et de métier en

définissant `role:Domain` et `role:Type` comme sous-classes de `skos:Concept`, en utilisant la relation `skos:broader` pour identifier les relations de hiérarchie qui existent entre les instances associées. L'ontologie se trouve donc ainsi réduite à un simple modèle OWL-DL basé sur SKOS et définissant comme suit les deux classes précitées en plus de la notion principale de rôle (`role:Role`). En conséquence, les différents domaines et métiers ainsi que leurs relations sont de ce fait modélisés au travers d'instances et de relations entre instances conformément avec SKOS (Figure 3.14, page 117), le modèle complet étant défini comme suit (Listing 3.13, page 116).

```
role:Role rdf:type owl:Class .
role:Domain rdf:type owl:Class ;
  rdfs:subClassOf skos:Concept .
role:Type rdf:type owl:Class ;
  rdfs:subClassOf skos:Concept .

role:hasRole rdf:type owl:ObjectProperty ;
  rdfs:domain foaf:Agent ;
  rdfs:range role:Role .
role:hasDomain rdf:type owl:ObjectProperty ;
  rdfs:domain foaf:Role ;
  rdfs:range role:Domain .
role:hasType rdf:type owl:ObjectProperty ;
  rdfs:domain foaf:Role ;
  rdfs:range role:Type .
```

Listing 3.13: Modèle complet pour la représentation des rôles

Ce choix d'utilisation de SKOS et d'une délégation des domaines, métiers et relations au niveau des instances a également été motivé par notre contexte d'annotations sémantiques guidées par des applications sociales comme nous l'avons brièvement évoqué auparavant. Nous ne souhaitons pas en effet que les utilisateurs modifient le modèle mais aient uniquement à gérer des instances. L'évolution des domaines et métiers peut donc être assurée par les utilisateurs finals, via les wikis (Section 4.2.4, page 160), sans que le modèle ne soit affecté. L'utilisation de SKOS nous semble ainsi être une bonne pratique dès lors qu'on souhaite modéliser des hiérarchies de concepts et d'une part avoir un modèle stable et d'autre part rester à un niveau OWL décidable. D'autres cas d'utilisation de SKOS confirment en outre cette bonne pratique [Isaac *et al.*, 2007].

Rappelons enfin que l'argumentation précédente se base sur l'utilisation d'OWL1 et que OWL2 permet de contourner les problèmes précédents. Cette évolution d'OWL (en cours de standardisation au moment de la rédaction de ce mémoire) introduit en effet la possibilité de définir une taxonomie de classes et d'utiliser ces classes comme instances sans pour autant basculer dans un modèle OWL-Full. Ceci s'effectue grâce au *punning* (ou métamodélisation), qui permet d'utiliser une même URI pour représenter à la fois une classe et une instance tout en restant décidable en temps fini⁷⁶.

⁷⁶<http://www.w3.org/TR/owl2-semantics/>

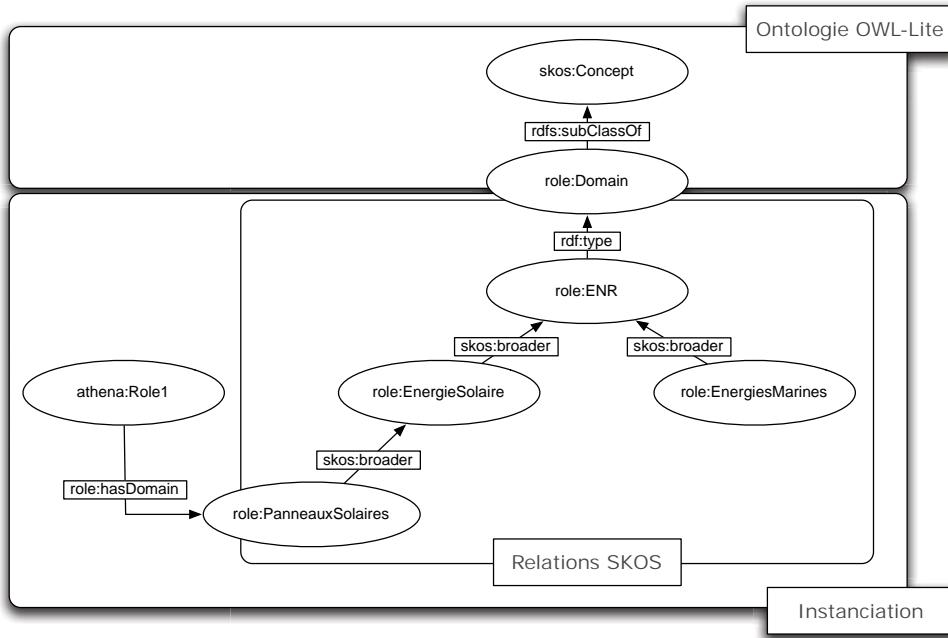


Figure 3.14: Taxonomies de domaines avec SKOS

3.2.5 Articulation globale des différentes ontologies métier

Les différents modèles présentés ci-avant sont donc chacun dédiés à un domaine particulier :

- FOAF permet de définir la notion d'agents (notamment de personnes et d'organisations) et certaines de leurs propriétés ;
- FOAFplus étend FOAF avec de nouvelles classes et propriétés ;
- la classe Partenariat du module du même nom permet de représenter et contextualiser les relations entre acteurs autour de domaines particuliers ;
- notre modèle pour la représentation des rôles permet de définir les différentes activités associées aux agents tout en séparant le domaine du métier ;
- SKOS nous permet de définir une taxonomie de domaines et métier non pas au niveau du modèle (classes), mais de la base de connaissance (instances) ;
- Geonames permet de localiser les entités, aussi bien avec un lien direct que via leur(s) rôle(s), les deux n'ayant évidemment pas la même valeur en termes de représentation.

Du fait de la faible adhérence des modèles en eux-mêmes avec notre contexte applicatif (cette adhérence se situant principalement au niveau des bases de connaissances associées comme nous l'avons vu), cet écosystème d'ontologies nous semble pertinent pour tout système d'Entreprise 2.0 à partir du moment où l'on souhaite disposer de modèles simples et extensibles pour définir un contexte industriel particulier.

L'exemple suivant (Listing 3.14, page 118) représente ainsi différentes assertions au sujet d'EDF utilisant les modèles précités, assertions que l'on retrouve par la suite représentées de manière graphique (Figure 3.15, page 119). Ce schéma permet de plus de faire apparaître les diverses relations qui peuvent exister entre modèles et instances définies aussi bien en interne via nos outils (relations `skos:broader` entre instances de `role:Domain`) qu'en externe via des données présentes sur le Web (relations `geonames:parentFeature` entre instances de `geonames:Feature`).

```
athena:EDF a foafplus:Company ;
  role:hasRole [
    role:hasType athena:Constructeur ;
    role:hasDomain athena:CentraleNucleaire ;
    geonames:locatedIn <http://sws.geonames.org/3017382/>
  ] ;
  role:hasRole [
    role:hasType athena:Producteur ;
    role:hasDomain athena:EnergieNucleaire ;
    geonames:locatedIn <http://sws.geonames.org/3017382/>
  ] ;
  geonames:locatedIn <http://sws.geonames.org/2988507/> ;
  foaf:member athena:PierreGadonneix .

athena:PierreGadonneix a foaf:Person ;
  geonames:locatedIn <http://sws.geonames.org/2988507/> .
```

Listing 3.14: Ensemble d'assertions au sujet d'EDF à l'aide de différents modèles

Enfin, si ces modèles forment le noyau de représentation métier au sein de notre médiateur, d'autres ontologies peuvent être utilisées, notamment en termes de propriétés :

- celles pour lesquelles le domaine (`rdfs:domain`) ou le codomaine (`rdfs:range`) n'est pas restreint et peut donc être adapté à n'importe quelle classe de nos modèles. On peut par exemple utiliser `dct:description` pour ajouter une description complète à chaque instance ;
- celles pour lesquelles le domaine ou le codomaine, bien que défini, est consistant avec nos vocabulaires. Par consistant, nous entendons qu'il ne va pas à l'encontre des axiomes définis à la fois dans nos modèles et dans le modèle des propriétés en question. Ceci nécessite cependant l'utilisation d'un raisonneur pour valider leur utilisation et la consistance du modèle, qu'il s'agisse de simple raisonnement RDFS sur les classes / sous-classes ou de raisonnement OWL plus poussé prenant en compte les éventuelles unions (`owl:unionOf`), intersections (`owl:intersectionOf`) ou disjonctions (`owl:disjointWith`).

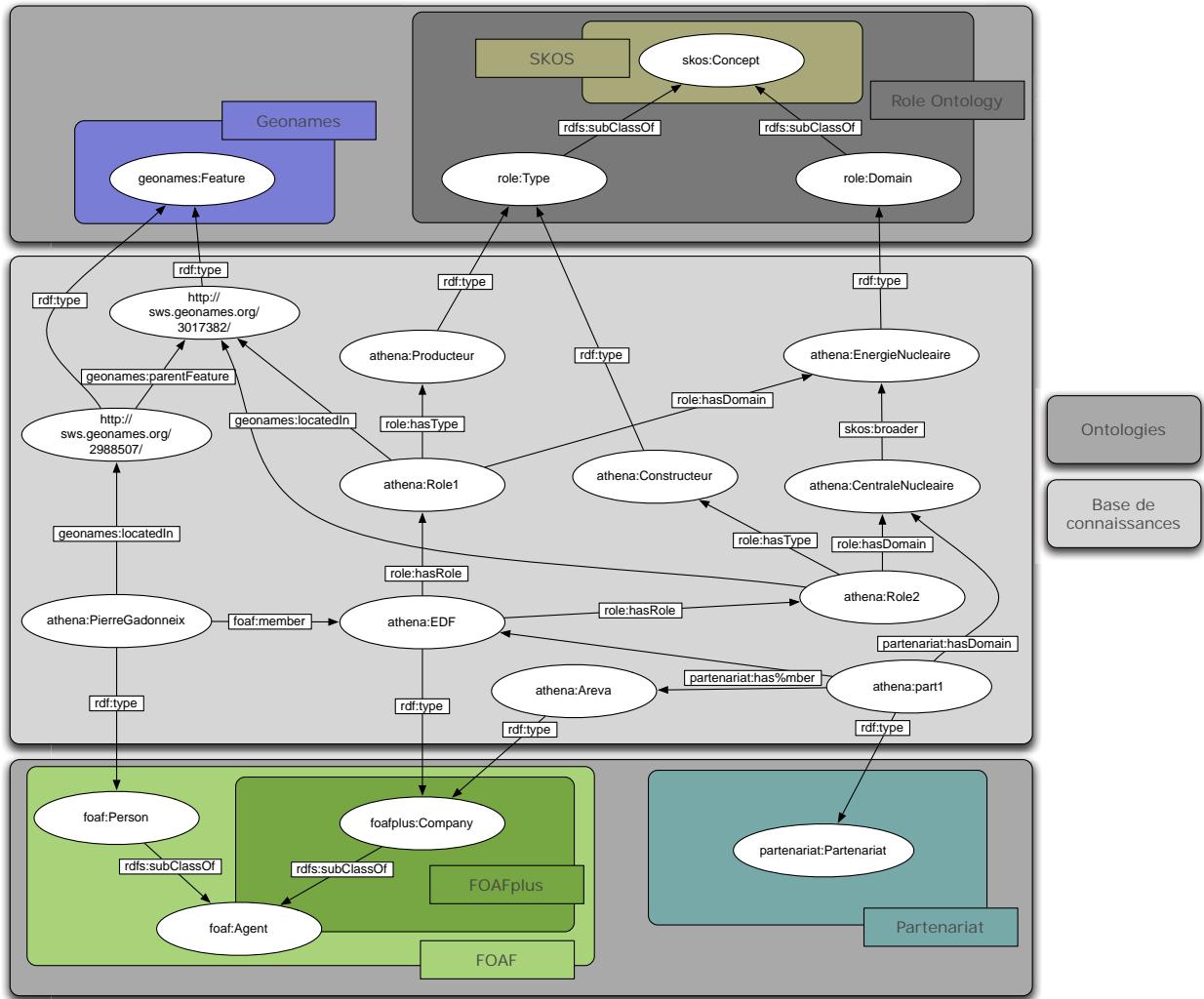


Figure 3.15: Combinaison d'ontologies et base de connaissance associée pour définir des assertions au sujet d'EDF

3.3 MOAT POUR LIER TAGS ET ONTOLOGIES

3.3.1 Tags, folksonomies et ontologies : un état de l'art

Folksonomies et ontologies ont régulièrement été confrontées, le plus souvent à tort. Un point de vue fréquent est ainsi de considérer la folksonomie comme une classification *bottom-up* orientée utilisateurs qui s'oppose à l'ontologie considérée comme une approche *top-down* centralisée. Cette opposition va d'ailleurs dans le sens d'une confrontation globale entre Web 2.0 et Web Sémantique que l'on retrouve souvent sur le Web et qui nous semble stérile comme nous l'avons déjà évoqué (Section 1.3, page 42). Comme nous le montrons tout au long de ce mémoire, c'est selon nous une complémentarité et non une distinction qu'il faut envisager entre Web Sémantique et Web 2.0 et il en est de même pour les relations entre folksonomies et ontologies. Ainsi rien ne s'oppose à la complémentarité des deux approches puisque l'on a d'un côté une pratique utilisateur et un modèle émergent (folksonomie) de l'autre un mode de représentation formelle (ontologie) comme l'ont souligné [Gandon et Giboin, 2008] : "*Les ontologies se définissent par le type de leur contenu. Les folksonomies se définissent par leur moyen d'obtention*". Différents travaux s'intéressent ainsi aux rapprochements et convergences possibles entre ces deux approches et l'on peut les classer en deux grandes familles, qui peuvent également se rejoindre sur certains points :

- les travaux cherchant à identifier une sémantique émergente depuis les folksonomies, voire à extraire des modèles taxonomiques ou ontologiques à partir de celles-ci ;
- les travaux visant à proposer des modèles de représentation pour les tags, les folksonomies et les objets associés (actions de *tagging*, nuages de tags ...) avec les technologies du Web Sémantique.

Avant de présenter nos travaux sur le sujet (Section 3.3.2, page 127), nous allons nous intéresser à l'état de l'art associé à ces deux courants et présenter certaines applications qui s'y rapportent.

Extraire une sémantique émergente depuis les tags

De nombreux travaux poursuivent l'objectif d'extraire des modèles structurés – taxonomies ou ontologies – depuis les folksonomies, principalement dans l'objectif de résoudre les problèmes classiques des systèmes à base de tags (Section 2.2.3, page 63). L'objectif est alors d'expliquer la sémantique qui peut exister dans ces systèmes là où celle-ci n'est qu'impliquée en raison de la nature même des folksonomies. La plupart d'entre eux se basent sur la notion de sémantique émergente [Staab, 2002] où l'usage collectif fait apparaître une sémantique contrôlée par la base (approche *bottom-up*) en opposition aux approches où la sémantique est définie en amont (*top-down*). [Mika, 2005] évoque à ce sujet "*ontologies would thus become an emergent effect of the system as opposed to be a fixed, limited contract of the majority*". L'ontologie (ou la taxonomie) émerge ainsi par effet de bord de l'architecture participative des systèmes à base de tags. Ce processus permet également de diminuer le goulot d'étranglement lié à l'acquisition d'un modèle structuré, étape généralement coûteuse, puisque ce modèle est ici issu des actions utilisateurs et des tags utilisés.

Afin d'identifier cette sémantique émergente au sein des folksonomies, [Mika, 2005] propose ainsi une approche sociale de constitution d'ontologies. Il définit alors l'ontologie comme

un modèle tripartite basé sur celui des folksonomies et ne considère plus uniquement les notions de classes et d'instances mais fait intervenir une composante sociale pour établir un modèle entre concepts (qu'il considère ici comme étant les tags), instances (les contenus tagués) et acteurs (les responsables des actions de *tagging*). En quelque sorte, il réifie la notion d'instanciation des ontologies au travers de la structure sociale des systèmes à base de tags. À l'aide de ce modèle et en se basant sur des approches de *clustering* et de cooccurrence combinées avec des techniques d'analyse de réseaux sociaux, il observe l'émergence de modèles taxonomiques à partir de folksonomies. En appliquant son approche à différents jeux de données, il identifie également un parallèle entre la subsomption d'un concept par un autre et l'inclusion de la communauté utilisant le tag le plus spécifique au sein de la seconde. Cette constatation nous semble en outre liée à la notion d'expertise au sein des systèmes à base de tags que nous avons constatée dans notre système et également évoquée par [Golder et Huberman, 2006] (Section 2.2.3, page 63). [Halpin *et al.*, 2006] se basent quand à eux sur une approche de cooccurrences réciproques entre tags pour extraire des relations taxonomiques, modélisées en RDFS avec `rdfs:subClassOf`, à partir d'une étude des *bookmarks* annotés par différents utilisateurs sur Delicious. [Schmitz, 2006] propose également une approche basée sur les cooccurrences de tags et sur un modèle statistique de subsomption (proposé par [Sanderson et Croft, 1999]) pour établir une hiérarchie de tags depuis Flickr. Tout comme les deux approches précédentes, la sémantique des relations se résume à une unique relation de subsomption et mélange ce que l'on aurait probablement distingué entre classes et instances dans une approche de constitution classique d'un tel modèle. Il nous semble par ailleurs que SKOS serait ici plus approprié qu'une hiérarchie de classes RDFS/OWL pour la modélisation de tels exports.

La méthodologie FLOR⁷⁷ – *Folksonomy Ontology enRichment* [Angeletou, 2008] – définit quand à elle une méthode totalement non-supervisée (se basant notamment sur les résultats obtenus par [Specia et Motta, 2007]) permettant d'expliciter la sémantique des tags et surtout des relations entre tags. Contrairement aux travaux précédents qui se limitent à des relations taxonomiques, leur approche permet d'extraire des relations typées entre concepts. Cette méthodologie repose notamment sur des notions de filtrage linguistique et d'expansion de termes et utilise différents outils proposés par le moteur sémantique Watson⁷⁸ [d'Aquin *et al.*, 2008]. Avec FolksOntology, [Van Damme *et al.*, 2007] proposent une approche semblable, l'utilisateur ayant en plus la possibilité de définir explicitement la sémantique des tags pour lesquels le système n'a pu trouver d'ontologie adaptée, *i.e.* de spécifier s'il s'agit d'une classe, d'une instance ou d'une propriété. On peut ainsi, plus qu'aligner la folksonomie avec des ontologies existantes, créer de nouveaux concepts. Malheureusement, contrairement à FLOR, cette approche se contente d'extraire un modèle mais n'applique pas celui-ci aux contenus tagués, ce qui nous semble pourtant être un des avantages de l'ontologie ainsi générée.

⁷⁷<http://flor.kmi.open.ac.uk/>

⁷⁸<http://watson.kmi.open.ac.uk/>

Modéliser les tags avec les technologies du Web Sémantique

Nous allons dans cette section présenter un certain nombre de travaux visant à modéliser les différents éléments des systèmes à base de tags (tags, actions de *tagging* ...) avec les technologies du Web Sémantique. De tels modèles, que l'on peut considérer comme des *ontologies pour les folksonomies*, permettent ainsi d'envisager les systèmes à base de tags comme partie intégrante du Web Sémantique, puisque représentés en RDF(S)/OWL.

[Gruber, 2007] propose un premier modèle⁷⁹ étendant la notion tripartite classique d'une action de *tagging* (Section 1.2.3, page 38) et où il définit celle-ci comme une relation faisant intervenir quatre éléments :

- un *Objet*, i.e. la ressource annotée quelque soit son type (billet de blog, photo, etc.) ;
- un *Tag*, i.e. le tag annotant la ressource ;
- un *Agent*, i.e. l'agent – en règle général une personne – qui crée la relation ;
- une *Source*, i.e. l'espace où est effectuée cette action (e.g. Flickr). C'est cette dernière propriété qui enrichit la relation initiale et qui permet de distinguer deux actions de *tagging* d'un même auteur pour la même ressource mais sur deux espaces distincts.

Un cinquième élément peut également intervenir dans cette relation, à savoir une polarité permettant d'assigner une valeur positive ou négative à la relation, dans l'objectif de résoudre des problèmes de *spam*. [Gruber, 2007] introduit également la notion d'identité d'un tag. Il considère ainsi que des tags peuvent être définis comme identiques malgré des labels distincts, établissant un premier pas vers l'unification de tags hétérogènes et la notion de sens associés aux tags (Section 3.3.2, page 127). Ce modèle qui sera à la base du projet Tag-Commons⁸⁰ reste cependant purement théorique et ne propose pas d'ontologie RDFS/OWL prête à utilisée.

[Newman *et al.*, 2005] définit avec la *Tag Ontology*⁸¹ une ontologie OWL-Full reprenant certains des principes définis par [Gruber, 2007]. Cette ontologie définit une classe `tag:Tag` pour modéliser les tags, sous-classe de `skos:Concept`, chaque tag disposant d'un ou plusieurs labels (via la propriété `tag:tagName`). Ce label représente le tag en tant que simple chaîne de caractères, tel que vu par l'utilisateur. L'utilisation d'une classe spécifique pour représenter les tags implique que chaque tag est défini par une URI et non plus par une simple chaîne de caractères. En conséquence, il est possible d'établir des assertions RDF entre tags et notamment de les organiser en créant des relations de proximité entre eux. Le modèle définit ainsi une propriété `tag:relatedTag` (sous-propriété de `skos:semanticRelation`) pour représenter les relations possibles entre différents tags. Malheureusement, cette relation-ci ne porte pas suffisamment de sémantique pour définir si deux tags sont liés par proximité linguistique (ex : un tag est le pluriel d'un autre) ou parce qu'ils évoquent des domaines plus ou moins proches (ex : EDF et énergie). Un autre défaut de ce modèle est l'absence de cardinalité vis-à-vis de la relation `tag:tagName`. Ainsi une instance de `tag:Tag` peut avoir deux labels totalement distincts qui lui sont rattachés. Par exemple un même tag peut être associé aux labels RDF et énergie, entraînant une incohérence évidente qu'un raisonneur

⁷⁹Publié originellement sur le Web en 2005 – <http://tomgruber.org/writing/ontology-of-folksonomy.htm>

⁸⁰<http://tagcommons.org>

⁸¹Espace de noms <http://www.holygoat.co.uk/owl/redwood/0.1/tags/>, préfixe `tag` par la suite.

ne pourra cependant détecter puisque non représentée dans le modèle. Ce modèle définit également la notion d'action de *tagging* via une classe `tag:Tagging` et des relations à partir de celle-ci vers l'utilisateur `tag:taggedBy` (cette propriété étant définie avec un codomaine `foaf:Agent`), le tag `tag:taggedWithTag` et la ressource `tag:taggedResource` associés. Il propose de plus une classe `tag:RestrictedTagging`, sous-classe de `tag:Tagging`, permettant de représenter une action de *tagging* pour un unique tag (via l'utilisation d'une restriction de cardinalité sur la propriété `taggedWithTag`), considérant ainsi l'action de *tagging* comme une relation tripartite stricte où un unique tag entre en jeu. La figure qui suit montre ainsi la représentation d'une telle action où un tag `apple` est ici assigné à un billet de blog représenté avec SIOC (Figure 3.16, page 123). Si la source et la polarité de chaque action ne sont pas prises en compte dans ce modèle contrairement à [Gruber, 2007], une composante temporelle peut-être ajoutée via la propriété `tag:taggedOn`, sous-propriété de `dc:date`⁸².

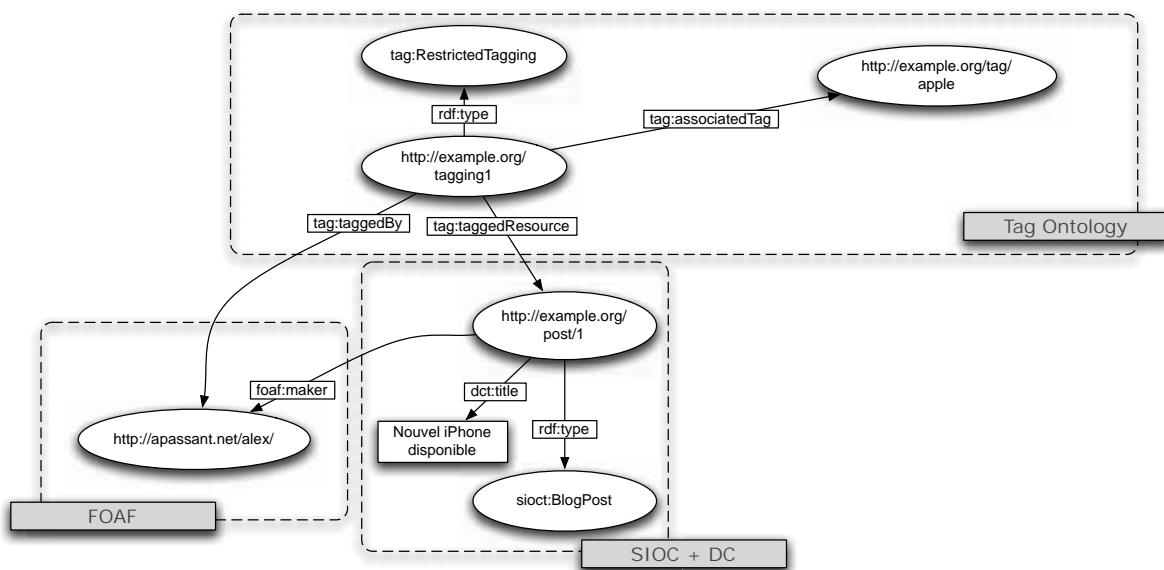


Figure 3.16: Tags et actions de *tagging* avec la *Tag Ontology*

Partant d'un besoin d'interopérabilité des tags entre applications, SCOT – *Semantic Cloud Of Tags* [Kim *et al.*, 2007] – se base sur les travaux précédents pour définir un modèle relatif à la représentation des nuages de tags. L'objectif est notamment de permettre l'export de l'ensemble des tags d'un utilisateur et leur fréquence d'utilisation d'un service vers un autre, toujours dans cette idée de portabilité des données sociales (Section 3.1.5, page 96). Pour ce faire, SCOT⁸³ introduit différentes classes et propriétés pour modéliser entre autres les cooccurrences entre tags au sein d'un système particulier (propriété `scot:coocurs_in`

⁸²<http://purl.org/dc/elements/1.1/date>

⁸³Espace de noms <http://scot-project.org/scot/ns>, préfixe `scot` par la suite.

et classe `scot:Cooccurrence`). SCOT permet également de représenter plus finement que dans la *Tag Ontology* les relations entre tags, avec une dizaine de propriétés distinctes comme `scot:acronym` ou `scot:plural` mais ne résout malheureusement pas le problème de cardinalité évoqué précédemment.

Toujours dans cette optique d'ontologies pour représenter les tags, [Knerr, 2006] propose TagOnt⁸⁴ qui reprend le modèle de [Newman et al., 2005] en y ajoutant la notion de visibilité d'une action de *tagging*. Malheureusement, ce modèle redéfinit ses propres classes et propriétés au lieu d'étendre la *Tag Ontology*, et bien que disponible en ligne⁸⁵ il ne semble être utilisé dans aucun projet. Un modèle similaire est proposé par [Echarte et al., 2007]⁸⁶ mais ne semble également pas avoir été utilisé en pratique. NEPOMUK propose via le vocabulaire NAO – *NEPOMUK Annotation Ontology* [Scerri et al., 2007] – ⁸⁷ une classe `nao:Tag` et une propriété `nao:has_tag` pour identifier les tags rattachés à une ressource quelconque, sans pour autant considérer l'action de *tagging* en tant que modèle tripartite. SIOC quant à lui définit une simple classe `Tag` qui peut être utilisée en complément avec `sioc:topic` pour représenter les tags associés à un item. Il est également possible d'utiliser SKOS pour représenter des tags via la classe `skos:Concept` (`sioc:Tag` hérite d'ailleurs de cette classe), les instances associées pouvant ensuite être associées aux contenus tagués via la propriété `sioc:topic` où jusqu'à peu via `skos:subject`, aujourd'hui obsolète comme nous l'avons signalé en évoquant les relations entre SIOC et SKOS (Section 3.1.4, page 94). Enfin, il est important également de signaler le modèle Bookmark⁸⁸ [Koivunen et al., 2001] proposé par Annotea⁸⁹ [Kahan et Koivunen, 2001]. Bien qu'il ne fasse pas explicitement référence à la notion de tags telle que popularisée ces dernières années, ce vocabulaire permet de faire le lien entre une ressource et un ensemble de termes annotants, représentés via une classe `bookmark:Topic` et une propriété `bookmark:Bookmark`. Ce modèle permet également d'organiser hiérarchiquement différentes instances de `bookmark:Topic` avec une propriété `bookmark:subTopicOf`, proposant un processus similaire à la propriété `skos:broader` définie dans SKOS. Citons également le module Taxonomy de RSS 1.0⁹⁰ qui propose une propriété pour représenter les différents sujets associés à un élément de flux RSS. Pour finir, notons l'existence du microformat `rel:tag`. Même s'il ne s'agit pas d'un modèle RDF, l'utilisation de GRDDL, que nous avons présenté dans le premier chapitre de cette thèse (Section 1.1.2, page 16), permet de transformer des données XHTML utilisant ce microformat en données RDF en utilisant par exemple un des modèles présentés en amont.

Le tableau qui suit (Tableau 3.2, page 125) synthétise les différents modèles étudiés précédemment et les compare selon différents critères. Ici, nous considérons uniquement ce que les modèles en eux-mêmes permettent de modéliser et non pas la manière dont ils utilisent

⁸⁴<http://code.google.com/p/tagont/>

⁸⁵<http://tagont.googlecode.com/files/tagont.owl>

⁸⁶<http://www.eslomas.com/tagontology-1.owl>

⁸⁷Espace de noms <http://www.semanticdesktop.org/ontologies/2007/08/15/nao#>, préfixe `nao` par la suite.

⁸⁸Espace de noms <http://www.w3.org/2003/07/Annotea/BookmarkSchema-20030707>, préfixe `bookmark` par la suite.

⁸⁹<http://www.w3.org/2001/Annotea/>

⁹⁰<http://web.resource.org/rss/1.0/modules/taxonomy/>

des vocabulaires externes (notamment en définissant des sous-classes ou sous-propriétés). Par exemple, bien que la *Tag Ontology* permette l'utilisation de la propriété `foaf:maker`, elle ne définit pas elle-même la notion d'agent ayant annoté une ressource (mais se base sur FOAF), ce qui explique que ce critère soit ici considéré comme non satisfait, tout comme l'est la notion de modèle tripartite pour SCOT, qui utilise pour se faire la *Tag Ontology*.

Notons également que ce que nous appelons *tagging (simple)* se réfère à la modélisation d'une relation directe entre une ressource et ses tags annotant et que *tagging (tripartite)* évoque la modélisation d'une actions de *tagging* en tant que modèle (*a minima*) tripartite. Le dialecte de chaque ontologie OWL a en outre été validé par Pellet.

Ontologie	Format	Supporte la modélisation de				
		Tag	Tagging (simple)	Tagging (tripartite)	Agent	Nuage de tags
Gruber	N/A					
<i>Tag Ontology</i>	OWL-Full					
SCOT	OWL-Full					
NAO	RDFS					
TagOnt ^a	OWL					
Echarte	OWL-DL					
SKOS Core	OWL-Full					
SIOC	OWL-Lite					
Annotea	RDFS					
Taxonomy	RDFS					
rel-tag	Microformat					

^aCe modèle n'a pu être validé par Pellet.

Tableau 3.2: Comparaison de différentes ontologies pour la représentation des tags et des objets associés⁹¹

Outils combinant tagging et technologies du Web Sémantique

De nombreux outils combinent système de tags et technologies du Web Sémantique et nous allons ici présenter certains d'entre eux.

Tout d'abord, citons Annotea [Kahan et Koivunen, 2001] qui propose dès 2001 un système d'annotations et de partage de ressources Web ouvert et reposant sur les technologies du Web Sémantique. Cet outil permet à chaque communauté de disposer de son propre serveur d'annotations, les différentes annotations produites étant ensuite disponibles en RDF utilisant un modèle particulier d'annotations⁹² combiné au vocabulaire Bookmark présenté auparavant. On peut ainsi considérer Annotea comme une des premières applications sociales de partage de contenus basé sur les technologies du Web Sémantique. Dans cette même idée de représenter des contenus annotés avec les technologies du Web Sémantique,

⁹¹✓ correspond aux critères satisfaits, ✗ aux critères non satisfaits.

⁹²<http://www.w3.org/2000/10/annotation-ns#>

Revyu⁹³ [Heath et Motta, 2007] est un service de revues entièrement basé sur les standards du Web Sémantique. Il repose notamment sur des heuristiques permettant de lier automatiquement les revues à des ressources déjà existantes, par exemple des livres en vente sur Amazon.com auxquels une URI propre a été assignée [Bizer *et al.*, 2007a]. L'ensemble des annotations produites au sein de cet outil est en outre disponible en RDF et utilise la *Tag Ontology* pour la représentation des tags. Toujours dans une approche de partage de contenus, Faviki⁹⁴ propose un service de gestion de favoris où les tags sont des identifiants DBpedia [Miličić, 2008]. Il prend ainsi en compte la notion de multilinguisme associée aux tags, puisqu'une même URI peut être associée à plusieurs termes.

D'autres outils sont axés plus spécifiquement sur la gestion des tags, et plus particulièrement sur la manière de les organiser pour pallier à leurs limites (Section 2.2.3, page 63). Ainsi, les outils de *bookmarking* SemanticScuttle⁹⁵ [Huynh-Kim-Bang et Dané, 2008], Gnizr⁹⁶ et Semanlink⁹⁷ [Servant, 2006] permettent de définir des relations hiérarchiques entre tags, le second offrant un export RDF des contenus annotés en utilisant certaines des ontologies présentées plus haut (notamment la angTag Ontology, SIOC et SKOS), le dernier étant basé sur son propre modèle de représentation des tags reposant sur SKOS⁹⁸⁹⁹. Dans une approche différente, GroupMe¹⁰⁰ propose aux utilisateurs de regrouper les tags par catégories pour faciliter la recherche d'information, représentant le tout avec sa propre ontologie [Abel *et al.*, 2007]. Sweetwiki [Buffa *et al.*, 2008] permet également l'organisation de tags (et utilise son propre modèle de représentation), cette fois-ci au sein d'un wiki (Section 4.2.1, page 148). L'approche reste fidèle à la philosophie wiki en permettant à tous les utilisateurs du système de gérer cette organisation commune de l'ensemble des tags du wiki de manière ouverte et collaborative.

S'ils n'utilisent pas explicitement les technologies du Web Sémantique, d'autres outils permettent manuellement de structurer ou d'enrichir les systèmes à base de tags et de bénéficier de ces enrichissements au moment de la recherche d'information. Ainsi, toujours dans une approche permettant de dériver des relations taxonomiques à partir de folksonomies, [Jäschke *et al.*, 2008] proposent aux utilisateurs de Bibsonomy¹⁰¹ (outil collaboratif de gestion de références bibliographiques issu du projet TagOra¹⁰²) de définir eux-mêmes des relations hiérarchiques entre tags. [Tanasescu et Streibel, 2007] proposent une autre solution pour structurer les folksonomies avec l'*Extreme Tagging*. Les utilisateurs ont ici la possibilité de typer les tags et les relations entre ces tags, à nouveau en utilisant des tags. Par exemple, on va pouvoir taguer le tag *apple* par *fruit* dans une action de *tagging*, et par *mac*

⁹³<http://revyu.com>

⁹⁴<http://faviki.com>

⁹⁵<http://sourceforge.net/projects/semanticscuttle/>

⁹⁶<http://code.google.com/p/gnizr/>

⁹⁷<http://www.semanlink.net/s1/home>

⁹⁸<http://www.semanlink.net/2001/00/semanlink-schema#>

⁹⁹Puisque basé sur SKOS et n'apportant pas de spécificité particulière en terme de fonctionnalités par rapport à celui-ci, nous ne l'avons pas inclus dans le comparatif précédent.

¹⁰⁰<http://groupme.org/GroupMe/>

¹⁰¹<http://bibsonomy.org>

¹⁰²<http://www.tagora-project.eu/>

dans une autre. Si l'idée est intéressante, l'utilisation de simples tags pour définir ces types nous semble conduire aux mêmes problèmes que ceux qu'elle souhaite résoudre. Enfin, on peut également citer les *machine tags* de Flickr¹⁰³, où les utilisateurs peuvent définir des tags *prédicat=objet*, par exemple *dct:description=New-York* ou *geo:lat=42.33*. S'ils ne sont nativement pas modélisés en RDF, ces *machine tags* peuvent être traduits comme tels via l'API Flickcurl¹⁰⁴.

3.3.2 Représentation de la signification des tags avec MOAT

Si les ontologies étudiées dans la section précédente modélisent les notions de tag et d'activité de *tagging*, aucune ne permet de prendre en compte la *signification* qui peut être associée à un tag dans le cadre d'une action de *tagging* particulière. Nous considérons en effet que lorsqu'un utilisateur associe un tag à une ressource, il lui assigne une *signification* particulière qu'il est nécessaire de prendre en compte pour interpréter correctement cette annotation. Comme nous l'avons déjà évoqué, plusieurs problèmes se posent dans l'assignation de tags en tant que simple libellés. On peut ainsi voir les limitations des tags en tant que simples mots-clés de la manière suivante :

- d'une part, comme le souligne [Bachimont, 2000] en évoquant la notion de libellés et d'ontologies "*si ces libellés sont interprétables, rien n'impose qu'ils soient interprétés de la même manière ou à tout le moins de manière cohérente et compatible entre plusieurs spécialistes*".
- si l'interprétation est possible, celle-ci dépend également du contexte interprétatif : un tag *apple* associé à une photo de fruits aura un sens différent de celui associé au même tag *apple* annotant un billet de blog sur l'iPhone. Si l'utilisateur est conscient de la signification qu'il donne à son tag au moment de l'annotation, celle-ci ne peut être prise en compte au moment de la recherche d'information, la machine ne considérant qu'une simple chaîne de caractères *a-p-p-l-e* sans aucune sémantique ;
- hors contexte, la sémantique est donc multiple et ambiguë. Si l'on prend le précédent tag *apple* tel quel, *i.e.* non associé à une ressource, il peut référencer aussi bien une entreprise qu'une maison de disque ou un fruit.

À partir de ce constat, il nous a semblé nécessaire de formaliser (1) la signification d'un tag dans le contexte d'une action de *tagging* particulière et (2) l'ensemble des significations potentielles que celui-ci peut avoir dans une folksonomie, *i.e.* selon un service ou une communauté donnée. On retrouve dans ce besoin de formalisation certains fondamentaux du Web Sémantique à savoir la notion d'identifiants référents et partagés pour représenter les choses (au travers d'URIs) et le passage de simples termes à ces identifiants (Section 1.1.2, page 16). Notre proposition, que nous allons maintenant détailler, a donc pour objectif de créer un pont entre cette notion souple de folksonomies et d'annotation contrôlée par l'utilisateur et les notions plus formelle du Web Sémantique et notamment l'indexation sémantique, en indexant donc les documents non plus par de simples termes (*i.e.* les tags) mais par des URIs de concepts. On passe ainsi d'une indexation par mot-clé à une indexation par concept (ou instance) d'ontologie, permettant de résoudre les différents problèmes posés par

¹⁰³<http://www.flickr.com/groups/mtags/>

¹⁰⁴<http://librdf.org/flickcurl/>

les systèmes à base de tags (Section 2.2.3, page 63). L'ontologie devient alors un support à la folksonomie, permettant d'associer souplesse de l'annotation par tag et langage formel pour représenter sans ambiguïté et de manière interprétable les significations associées à ces tags. Ce processus nous permet ainsi d'établir un lien fort entre ces différents degrés de formalisation, offrant chacun des perspectives différentes en terme d'annotation et de recherche d'information comme le souligne [Zacklad, 2007],

Si l'approche courante consiste à considérer une action de *tagging* comme une relation tripartite entre un utilisateur, un tag et la ressource annotée (Section 1.2.3, page 38), relation qui peut-être enrichie par des considérations temporelles ou d'espace social (cf. [Newman *et al.*, 2005] ou [Gruber, 2007]), nous y ajoutons un paramètre supplémentaire, à savoir la signification du tag dans ce contexte. Plus particulièrement, nous distinguons :

- la signification *locale* d'un tag, *i.e.* la signification particulière et non ambiguë d'un tag au sein d'une action de *tagging*;
- les significations *globales* d'un tag, *i.e.* l'ensemble des significations qui peuvent lui-être associées si l'on considère le terme seul, hors contexte.

Nous étendons ainsi de la manière suivante le modèle de représentation tripartite d'une action de *tagging* en un modèle quadripartite où la signification (*Signification*) est ici considérée comme *locale* :

$$\text{Tagging}(\text{Utilisateur}, \text{Ressource}, \text{Tag}, \text{Signification}) \quad (3.1)$$

D'autre part, modéliser l'ensemble des significations potentielles d'un tag dans une folksonomie donnée revient à considérer qu'à chaque tag est associé un ensemble de significations, que nous considérons ici comme *significations globales*. Chaque signification globale est de plus associée à la liste des utilisateurs l'ayant ainsi définie, afin de conserver un côté social dans cette association. Nous modélisons donc les *significations globales* d'un tag comme suit :

$$\text{Significations}(\text{Tag}) = \{(\text{Signification}, \{\text{Utilisateur}\})\} \quad (3.2)$$

Ainsi, à partir de ces deux définitions, nous pouvons représenter une folksonomie non plus comme composée de trois ensembles mais de quatre – Utilisateurs, Tags, Ressources et Significations – associés à un ensemble de relations (*i.e.* les actions de *tagging*) de la manière suivante :

$$\text{Folksonomie}(\text{User}^*, \text{Resource}^*, \text{Tag}^*, \text{Signification}^*, \text{Tagging}) \quad (3.3)$$

On peut également représenter ce modèle quadripartite au travers de la figure qui suit, identifiant ici au sein d'une folksonomie deux actions de *tagging* distinctes qui portent sur le même tag pour deux significations distinctes (Figure 3.17, page 129).

3.3.3 Modèle de représentation MOAT

L'introduction de la notion de signification au sein des systèmes à base de tags nous amène au problème de représentation de celle-ci. Si l'on utilise un simple label, le problème est simplement déporté, puisqu'à nouveau celui-ci peut-être ambigu et est sujet à

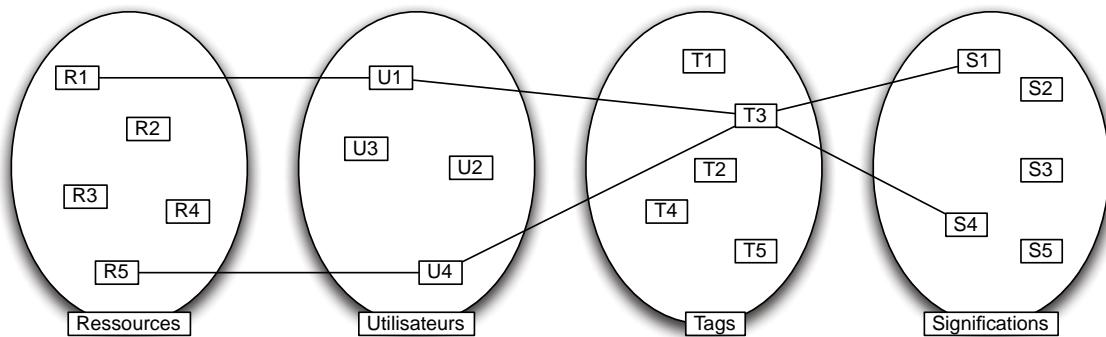


Figure 3.17: Modélisation quadripartite de deux relations de *tagging* au sein d'une folksonomie

son interprétation par le lecteur. Comme le souligne à nouveau [Bachimont, 2000] en évoquant les ontologies, "il est nécessaire de contraindre l'interprétation spontanée que fait tout spécialiste des libellés pour que, respectant ces contraintes d'interprétation, tout spécialiste associe les mêmes significations que ces confrères à un libellé". En allant plus loin, et puisque nous nous situons dans le contexte du Web Sémantique, nous souhaitons que non seulement tout spécialiste mais surtout tout agent logiciel interprète ces significations de la même manière. La notion d'interprétation par une machine est elle-même sujet à débat comme nous l'avons souligné au début de cette thèse (Section 1.1.1, page 12) et ici nous nous referons aux notions d'interprétations des données dans le contexte du Web Sémantique avec l'utilisation d'URIs et d'ontologies associées. Pour ce faire, nous représentons donc les significations non pas avec de simples labels (ce qui ne ferait que déplacer du tag à la signification les problèmes que l'on souhaite résoudre), mais via l'utilisation d'URIs de concepts du Web Sémantique, qu'il s'agisse d'instances d'ontologies de domaines (qui peuvent alors être internes à une organisation) ou provenant de bases de connaissances comme DBpedia, Geonames ou autres ressources du projet *Linking Open Data*. Les significations associées aux tags sont donc ainsi représentées par identifiants non-ambigus référençant des concepts interprétables par des agents logiciels. Pour en revenir à l'exemple précédent, on peut ainsi assigner au tag `apple` les significations globales `dbpedia:Apple` (identifiant pour le fruit) et `dbpedia:Apple_Inc.` (identifiant pour l'entreprise) permettant de distinguer ensuite, via une signification locale, le sens qu'un utilisateur a voulu donner à son tag au moment d'une action de *tagging* particulière. Si cette signification est destinée en premier aux machines, on peut malgré tout simplement en proposer une interprétation humaine en utilisant les différentes propriétés associées à ces URIs, notamment leur label (`rdfs:label`).

Nous avons ainsi proposé un premier modèle relativement simple permettant de considérer des ontologies de domaine (et les instances associées) en support des tags pour définir ces significations [Passant, 2007c]. Dans un objectif de formaliser plus finement ces relations, nous avons par la suite défini MOAT¹⁰⁵ – *Meaning Of A Tag* [Passant et Laublet, 2008b]. L'ob-

¹⁰⁵<http://moat-project.org>

jectif de MOAT est ainsi de permettre la représentation formelle de ces différentes significations, aussi bien locale que globales, pour modéliser des faits tels que "Dans le contexte de cette photo, j'utilise le tag `apple` représentant le concept identifié par dbpedia :`Apple`, i.e. le fruit alors que pour ce billet de blog, j'annote avec le même tag `apple` mais cette fois-ci avec une signification associée à dbpedia :`Apple_Inc`, i.e. l'entreprise". Les ontologies et bases de connaissances associées viennent donc ici en support des folksonomies, permettant de définir la sémantique de chaque tag. En se référant aux notions de termes, notions et concepts proposées par [Kassel et Perrette, 1999], MOAT permet donc le passage du terme (le tag `apple`) à la notion (la pomme en tant que fruit) et finalement au concept (identifié par une URI référente).

MOAT propose ainsi une ontologie OWL-DL¹⁰⁶ et introduit une classe `moat : Tag`, sous-classe de `tag : Tag`. La raison de cette redéfinition est la présence dans notre modèle d'une contrainte de cardinalité sur la relation `tag : name` pour résoudre les problèmes que nous avons évoqués auparavant au sujet de cette propriété (Section 3.3.1, page 122). Concernant les représentations globales d'un tag, nous représentons celles-ci avec un classe dédiée `moat : Meaning`, qui réifie la signification elle-même en proposant un lien `moat : meaningURI` vers une URI (la signification proprement dite, le lien étant unique) ainsi qu'un ensemble de liens `foaf : maker` vers les utilisateurs l'ayant défini. Une propriété `moat : hasMeaning` permet ensuite d'établir un lien entre une instance de `moat : Tag` et de `moat : Meaning` afin de représenter ces différentes significations globales comme le montre la figure (Figure 3.18, page 130) et le code RDF associé qui suivent (Listing 3.15, page 131). Ici le tag `apple` est identifié par `http://example.org/tag/apple` puisque nous définissons également une URI pour le tag lui-même, comme proposé par la Tag Ontology sur laquelle notre modèle se base.

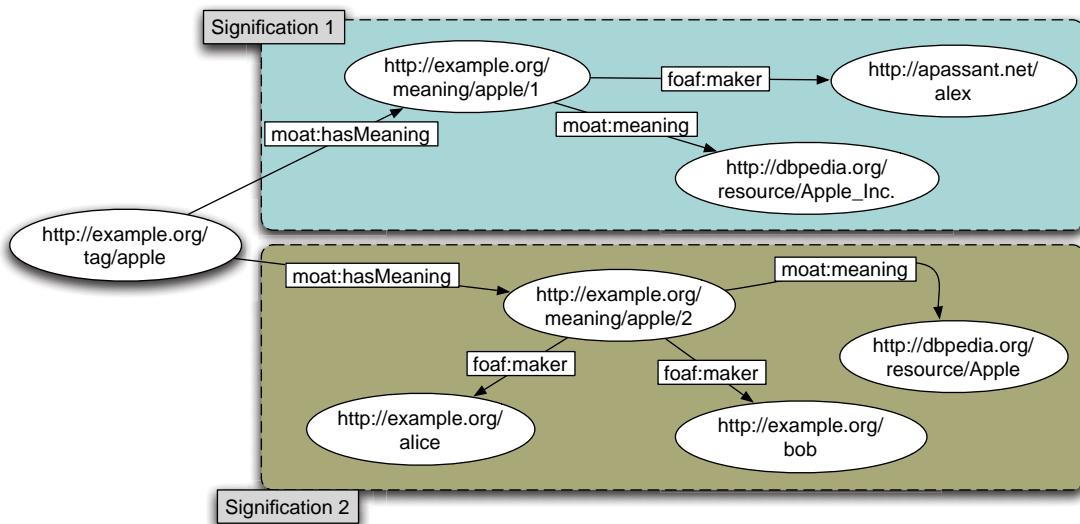


Figure 3.18: Significations globales du tag `apple` avec MOAT

¹⁰⁶Espace de noms `http://moat-project.org/ns#`, préfixe `moat` par la suite.

```

<http://example.org/tag/apple> a moat:Tag ;
  moat:hasMeaning <http://example.org/meaning/apple/1> ;
  moat:hasMeaning <http://example.org/meaning/apple/2> .

<http://example.org/meaning/apple/1> a moat:Meaning ;
  moat:meaningURI <http://dbpedia.org/resource/Apple_Inc.> ;
  foaf:maker <http://apassant.net/alex/>

<http://example.org/meaning/apple/2> a moat:Meaning ;
  moat:meaningURI <http://dbpedia.org/resource/Apple> ;
  foaf:maker <http://example.org/alice> ;
  foaf:maker <http://example.org/bob> .

```

Listing 3.15: Significations globales du tag "apple" avec MOAT

La représentation de la signification locale d'un tag se base quant à elle sur l'utilisation de la classe `tag:RestrictedTagging` de la *Tag Ontology*. Il est en effet nécessaire pour modéliser cette signification locale de considérer les tags pris de manière individuelle et en conséquence de considérer autant d'actions de *tagging* qu'il y a de tags afin d'éviter les problèmes de concordance qui peuvent arriver si l'on représente au sein d'une même action plusieurs tags et plusieurs significations. À partir de cette classe, nous avons introduit une propriété `moat:tagMeaning` qui permet de faire un lien au sein d'une action de *tagging* entre un tag et sa signification dans ce contexte comme l'illustrent le code (Listing 3.16, page 131) et la figure qui suivent (Figure 3.19, page 132).

```

<http://example.org/post/1> a sioc:Post ;
  foaf:maker <http://apassant.net/alex> ;
  dct:title "Nouvel iPhone disponible" ;
  moat:taggedWith <http://dbpedia.org/resource/Apple_Inc.> .

<http://example.org/tagging/1> a tag:RestrictedTagging ;
  tag:associatedTag <http://example.org/tag/apple> ;
  tag:taggedBy <http://apassant.net/alex> ;
  tag:taggedResource <http://example.org/post/1> ;
  moat:tagMeaning <http://dbpedia.org/resource/Apple_Inc.> .

```

Listing 3.16: Signification locale du tag "apple" avec MOAT

Cet exemple laisse de plus apparaître l'utilisation d'une propriété `moat:taggedWith`. Celle-ci permet d'établir un lien direct entre la ressource annotée et le concept représentant la signification du tag, sans pour autant passer par une représentation du modèle quadripartite de l'action de *tagging*. SIOC, SKOC ou encore la *Tag Ontology* proposent des propriétés similaires avec respectivement `sioc:topic`, `skos:subject` (la précédente étant une sous-propriété de celle-ci) ou encore `tag:taggedWithTag`. Cependant, la sémantique de ces propriétés indique qu'elles modélisent explicitement une relation vers le sujet associé au

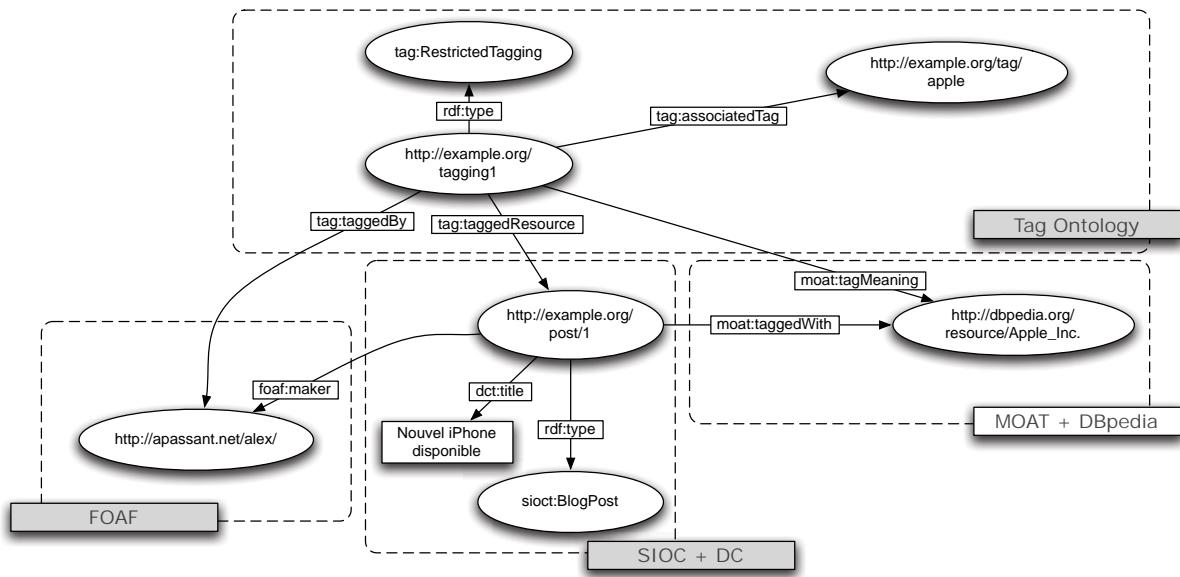


Figure 3.19: Représentation de la signification locale du tag apple avec MOAT et DBpedia

contenu annoté¹⁰⁷. Or, comme nous l'avons explicité dans le premier chapitre de cette thèse, certains tags peuvent être de l'ordre de métadonnées administratives ou structurelles, le tag ne reflétant alors pas un sujet associé au contenu, mais par exemple une information sur la source (e.g. un tag Flickr pour identifier une photo issue de ce site) (Section 1.2.3, page 38). De ce fait, en considérant comme proposé par MOAT que des concepts vont être utilisés en complément de ces tags, il est nécessaire de proposer une propriété qui va permettre d'établir un lien direct entre ressource annotée et concept sans pour autant considérer ce concept comme sujet. La propriété `moat:taggedWith` a ainsi pour objectif de répondre à cette problématique. Notons également qu'une simple règle d'inférence permet de passer d'une relation quadripartite à une relation directe entre ressource et concept utilisant cette propriété, comme le montre le code qui suit (Listing 3.17, page 132).

```
{
  iii a tag:RestrictedTagging ;
    tag:taggedResource uuu ;
    moat:tagMeaning vvv .
} => {
  uuu moat:taggedWith vvv .
}
```

Listing 3.17: Règle d'inférence pour MOAT, représentée en N3

¹⁰⁷ <http://librarytechnz.natlib.govt.nz/2008/09/adding-tags-to-dc-metadata.htm>

La figure suivante représente plus globalement le modèle MOAT et la manière dont interagissent significations locales et globales¹⁰⁸ (Figure 3.20, page 133) . Nous verrons dans le chapitre suivant comment tirer bénéfice de ce modèle, puisqu'en plus de celui-ci, MOAT propose également une architecture collaborative et des outils permettant à une communauté de franchir ce pas entre *tagging* et indexation sémantique sans être directement confronté au modèle et aux annotations (Section 4.3, page 170).

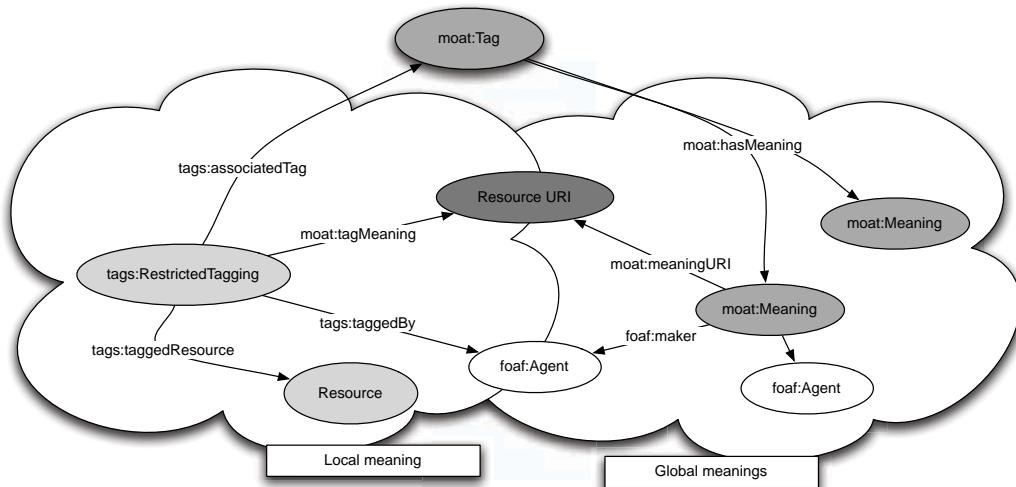


Figure 3.20: Modèle de représentation MOAT

Si nous avons essentiellement défini un modèle et un ensemble de services associés, nous ne nous sommes pas directement intéressés dans nos travaux à l'automatisation du processus. Cependant, ce modèle peut venir en support de telles approches, comme celles présentées auparavant telles que [Specia et Motta, 2007] ou [Van Damme *et al.*, 2007]. C'est par exemple ce que propose [Abel, 2008] en ayant récemment intégré MOAT à l'application GroupMe évoquée plus tôt et en ayant automatisé l'approche d'indexation sémantique. Ainsi, si notre approche se situe dans le domaine des modèles de représentation pour les tags, elle peut être utilisé pour supporter et formaliser les processus de structuration des systèmes à base de tags, permettant par exemple une interopérabilité entre différents algorithmes.

Enfin, une autre spécificité de notre modèle, notamment par rapport à ce que proposent la *Tag Ontology*, SCOT ou le modèle théorique proposé par [Jäschke *et al.*, 2008] (et mis en place dans Bibsonomy) est de ne pas chercher à organiser les tags entre eux pour pallier à leurs limites mais à passer par les concepts associés aux tags pour arriver à cet objectif. Si la possibilité d'organiser hiérarchiquement les tags permet de contextualiser les relations et de conserver une notion de point de vue personnalisée, à la manière de ce que propose

¹⁰⁸La propriété *moat:taggedWith* n'étant pas représentée pour considérer ici uniquement la représentation quadripartite du modèle.

[Zacklad, 2005] avec la notion d'ontologies sémiotiques, notre proposition nous semble plus pertinente pour plusieurs raisons :

- tout d'abord, il nous est apparu en consultant différents cas d'usage de ces principes de structuration de tags que de nombreuses relations ainsi définies sont assez générales comme par exemple, le fait que le tag `apple` soit associé à `iphone` ou `macintosh` ou que `france` soit plus spécifique que `europe`. Or ces relations sont pour la plupart déjà représentées dans des bases de connaissances existantes, notamment issues du projet *Linking Open Data*. Les relations des deux exemples précédents se retrouvent ainsi respectivement dans DBpedia et Geonames. Dans les cas où de telles relations n'existent pas, il nous semble également plus pertinent d'enrichir une base de connaissances existante plutôt que de représenter celles-ci dans un système clos, afin de permettre une réusabilité de telles informations ;
- de plus, alors que les relations taxonomiques classiques ne permettent pas de distinguer les différents liens qui peuvent exister entre tags (par exemple une notion de spécificité géographique ou le lien entre une marque et ses produits), notre approche permet de prendre en compte ces spécificités à partir du moment où les relations existent dans la base de connaissances associée et dans les ontologies sous-jacentes ;
- en conséquence, en ce qui concerne la recherche d'information et la possibilité de découvrir des documents proches, de nombreuses possibilités s'offrent à l'utilisateur. On passe en effet de l'utilisation d'un unique lien relationnel à un parcours de graphe multidimensionnel plus complet. On peut donc décider de visualiser des éléments proches selon un critère ou un autre, par exemple en fonction des produits associés à une marque ou des différentes personnes affiliées à une organisation. Nous déléguons alors la suggestion et l'identification de documents pertinents aux bases de connaissances associées. Se pose malgré tout le problème de la pertinence des différents liens, notamment devant des bases de connaissances contenant des milliers d'assertions, et nous l'aborderons plus tard dans ce mémoire (Section 5.4.3, page 216).

Notons enfin malgré tout que, bien que cela ne soit pas l'approche que nous défendons, notre modèle n'empêche pas la définition de relations simples entre tags, notamment puisque nous réutilisons la *Tag Ontology* et pouvons en conséquence réutiliser les différentes propriétés qu'elle définit à ce sujet.

3.3.4 Positionnement de MOAT par rapport à l'état de l'art

Pour finir ce descriptif de MOAT, nous allons ici étudier le positionnement du modèle proposé par rapport aux différentes ontologies permettant la modélisation des systèmes à base de tags étudiées précédemment (Section 3.3.1, page 122). En termes d'évaluation, nous détaillerons également dans le chapitre suivant différents chiffres relatifs à l'utilisation de MOAT dans notre contexte afin d'évaluer la pertinence de notre approche (Section 4.4, page 182).

À la lecture du tableau qui suit (Tableau 3.3, page 135), on observe que l'approche proposée par MOAT est la seule permettant de prendre en compte la signification des tags. Si certains modèles permettent d'organiser ceux-ci (comme SKOS ou la *Tag Ontology*), ils ne permettent pas d'associer ces tags à des représentations formelles (identifiées par leur

URI) comme nous le proposons. De plus, comme nous l'avons vu, certains de ces modèles permettent d'établir directement des liens entre ressources annotées et représentations formelles mais ceux-ci ne prennent alors pas en compte la notion de tag associé. Ajoutons également le fait que ce lien direct ne peut être utilisé pour des ressources dont le concept annotant n'est pas considéré comme sujet de la ressource. L'approche proposée par MOAT est donc complémentaire avec les modèles existants tout en permettant de prendre en compte ce lien entre tag et signification et non pas uniquement entre ressource et signification du tag.

Ontologie	Format	Supporte la modélisation de					
		Tag	Tagging (simple)	Tagging (tripartite)	Agent	Nuage de tags	Signi - cation
Gruber	N/A						
<i>Tag Ontology</i>	OWL-Full						
SCOT	OWL-Full						
NAO	RDFS						
TagOnt	OWL						
Echarte	OWL-DL						
SKOS Core	OWL-Full						
SIOC	OWL-Lite						
Annotea	RDFS						
Taxonomy	RDFS						
rel-tag	Microformat						
MOAT	OWL-DL						
<i>Tag Ontology + SCOT + MOAT + SIOC</i>	OWL-Full						

Tableau 3.3: Situation de MOAT par rapport à l'état de l'art

Cette complémentarité permet ainsi à MOAT, associée à la *Tag Ontology*, SIOC et SCOT de proposer un ensemble complet d'ontologies dédiées à la représentation des différentes caractéristiques des systèmes à base de tags sur le Web Sémantique : actions de *tagging* (*Tag Ontology*), utilisateurs (SIOC), nuages de tags (SCOT) et significations (MOAT). Par l'intermédiaire de cet écosystème d'ontologies, de tels systèmes peuvent être considérés comme des éléments à part entière du Web Sémantique, toujours dans cette vision d'une complémentarité globale entre Web 2.0 et Web Sémantique.

CONCLUSION

Nous avons présenté dans ce chapitre l'ensemble des ontologies utilisées au sein de notre médiateur sémantique et la manière dont elles interagissent pour former un modèle complet de représentation pour les activités, les documents et les données manipulées au sein de communautés Web 2.0 en entreprise. Nous avons tout d'abord présenté SIOC, modèle

aujourd'hui utilisé dans de nombreux cas d'utilisation relatifs à cette complémentarité entre Web 2.0 et Web Sémantique et qui nous permet dans notre contexte de représenter uniformément les documents créés depuis différents outils, de manière autonome comme nous allons le voir dans le chapitre suivant (Section 4.1, page 138). Nous avons ensuite présenté un ensemble d'ontologies de domaine relativement légères (pour la plupart reposant sur des modèles existants) qui permettent ainsi de modéliser différentes assertions métier au sujet de certains domaines d'expertise abordés par l'entreprise. Enfin, nous avons présenté MOAT, modèle permettant de combiner ontologies, bases de connaissances formelles, tags et folksonomies afin d'offrir un moyen de résoudre les problèmes de ces dernières tout en conservant leur souplesse. Nous avons également vu que certains de ces modèles dépassaient le cadre de l'Entreprise 2.0 et pouvaient être également utilisés sur le Web.

Plus particulièrement, nous retiendrons de ce chapitre la manière donc ces différents modèles se complètent pour offrir une vision complète et modulaire de différentes strates de représentation des connaissances dans un contexte de communautés actives autour de thématiques particulières. Ces différents modèles permettent ainsi de prendre en compte aussi bien les interactions sociales que les contenus créés via ces interactions sociales, tout en articulant ces différents niveaux de représentation via MOAT comme le montre la figure qui suit (Figure 3.21, page 136).

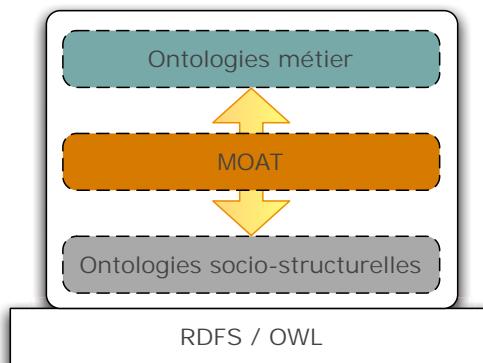


Figure 3.21: Articulation d'ontologies pour l'Entreprise 2.0

Nous allons maintenant nous intéresser aux outils et processus permettant le peuplement de ces différents modèles.

Chapitre 4

Annotations sémantiques et peuplement collaboratif d'ontologies

INTRODUCTION

Dans ce chapitre, nous allons nous intéresser aux différents outils mis en place pour exploiter les modèles précédents (Section 3, page 83) dans un objectif d'annotations sémantique et de peuplement collaboratif d'ontologies, aussi bien dans notre contexte d'écosystème sémantique pour l'Entreprise 2.0 que sur le Web en général.

Tout d'abord, nous présenterons les solutions mises en place pour la production automatique d'annotations socio-structurelles depuis les outils de blogs, wikis et les flux RSS (Section 3.1, page 84). Ces annotations sont naturellement basées sur SIOC et les différents vocabulaires associés (Section 4.1, page 138). Nous détaillerons ici les outils développés dans le contexte de la plate-forme Hermès mais aussi ceux destinés à usage plus large, notamment une API dédiée à la production de données SIOC. Ces différents outils permettent ainsi la production à grande échelle de données représentées avec SIOC, favorisant son acceptation sur le Web comme nous l'avons vu précédemment (Section 3.1.6, page 101).

Nous nous intéresserons ensuite aux méthodes de peuplement d'ontologies métier à l'aide d'outils Web 2.0. Nous argumenterons tout d'abord en quoi les outils habituels nous semblent limités dans un contexte où l'information doit-être constamment à jour puis en quoi les wikis sémantiques nous semblent offrir une réponse adaptée pour permettre un peuplement d'ontologie collaboratif, ouvert et évolutif. Un état de l'art nous permettra de dresser un panorama (non exhaustif) des outils et approches actuels dans ce domaine (Section 4.2.1, page 148). Nous présenterons ensuite un nouveau système de wiki sémantique, UfoWiki (Section 4.2.2, page 154) et détaillerons ses différents objectifs, ses principes ainsi que son architecture logicielle. Nous verrons également la manière dont celui-ci est utilisé dans notre contexte pour peupler les ontologies de domaine présentées auparavant (Section 3.2, page 104). En guise d'évaluation de l'outil, nous comparerons ses caractéristiques avec les systèmes existants et mettrons l'accent sur la manière dont celui-ci a été pris en main dans notre contexte.

Enfin, nous détaillerons le processus et l'architecture logicielle associés à MOAT (Section 3.3, page 120) mis en place pour permettre d'associer *tagging* et indexation sémantique de la manière la plus souple possible (Section 4.3, page 170). Nous verrons également comment celui-ci se couple avec la production automatique d'annotations socio-structurelles définie

auparavant dans ce chapitre. Si nos efforts ne se sont pas concentrés sur une automatisation de l'approche, nous avons mis en place un processus ouvert et collaboratif pour parvenir à cet objectif. Celui-ci combine ainsi principes du Web 2.0 et modélisation des données selon la vision du Web Sémantique. Nous verrons également en quoi ce processus permet de manière plus large d'intégrer des contenus Web 2.0 existants (produits depuis des services populaires comme Delicious ou Flickr) au sein du Web Sémantique avec l'outil LODr (Section 4.3.2, page 178).

4.1 ANNOTATION SÉMANTIQUE DE DOCUMENTS WEB 2.0

4.1.1 Une approche automatisée pour l'annotation socio-structurelle

Comme présenté auparavant, notre premier objectif en termes d'annotation sémantique est de fournir une représentation uniforme des métadonnées socio-structurelles de chaque contenu produit au sein de notre plate-forme, quelque soit l'outil d'origine (Section 2.3.3, page 73). Il peut donc s'agir de flux RSS agrégés depuis l'extérieur ou de billets de blogs et pages wiki rédigés en interne. Ces annotations doivent permettre d'accentuer l'interopérabilité des différents outils en offrant un cadre commun de représentation pour les documents créés au sein de notre écosystème.

Comme l'ont souligné [Karger et Quan, 2004] dans leur vision du *semantic blogging*, l'export d'annotations sémantiques depuis les outils de blog doit se faire sans intervention supplémentaire de l'utilisateur. Si leur argumentaire s'attache essentiellement aux blogs, nous pensons que cela doit-être le cas quelque soit le site ou le service utilisé (blog, wiki, outil de social networking ...), à partir du moment où les données à exporter sont déjà disponibles sous une forme ou une autre au sein du système. Il est en effet inutile de demander aux utilisateurs d'ajouter eux-mêmes ces annotations socio-structurelles (par exemple de définir la valeur de `dct:title` pour un billet de blog auquel un titre est déjà assigné) puisqu'elles seront redondantes avec les données déjà présentes au sein du système. De façon plus précise, ces données peuvent soit avoir été fournies directement par l'utilisateur (titre d'un billet, tags associés à un contenu, connections au sein d'un réseau social ...) soit définies automatiquement par le service lui-même (date de création, URL du document ...). On les distingue ainsi selon les appellations de métadonnées sociales et de métadonnées computationnelles (Figure 4.1, page 139).

La production de ce type d'annotations est donc automatisée à partir d'outils alignant les formats internes (base de données, système de fichiers, APIs...) avec un certain nombre d'ontologies utilisées pour représenter ces données en RDF. Concernant celles-ci, les ontologies présentées au chapitre précédent sont particulièrement adaptées : SIOC, FOAF, DublinCore ... (Section 3.1, page 84) Cette automatisation, que nous allons décrire par la suite, permet d'associer à chaque document son graphe d'annotations correspondant sans intervention supplémentaire de l'utilisateur. Tout document est modélisé de manière autonome en une instance de `sioc:Item` (ou d'une sous-classe) à laquelle différentes propriétés sont rattachées. Notons que nous utilisons `sioc:Item` en raison de notre contexte Web 2.0 mais que nous pouvons simplement utiliser la classe `foaf:Document` si nous souhaitons modéliser des documents plus classiques (rapports de réunion, dossiers d'expertise ...). Les conteneurs

de données sont quant à eux exportés en tant qu'instances de `sioc:Container` (ou d'une sous-classe). Les instances de `sioc:Item` associées sont rattachées au conteneur via la propriété `sioc:has_container`, le conteneur étant lui-même rattaché au site correspondant (`sioc:Site`) via `sioc:has_host`. L'auteur du document est quant à lui modélisé en tant qu'instance de `sioc:User`, associé au document source via `sioc:has_creator`. La figure qui suit exemplifie cette traduction pour un contenu particulier (Figure 4.1, page 139).

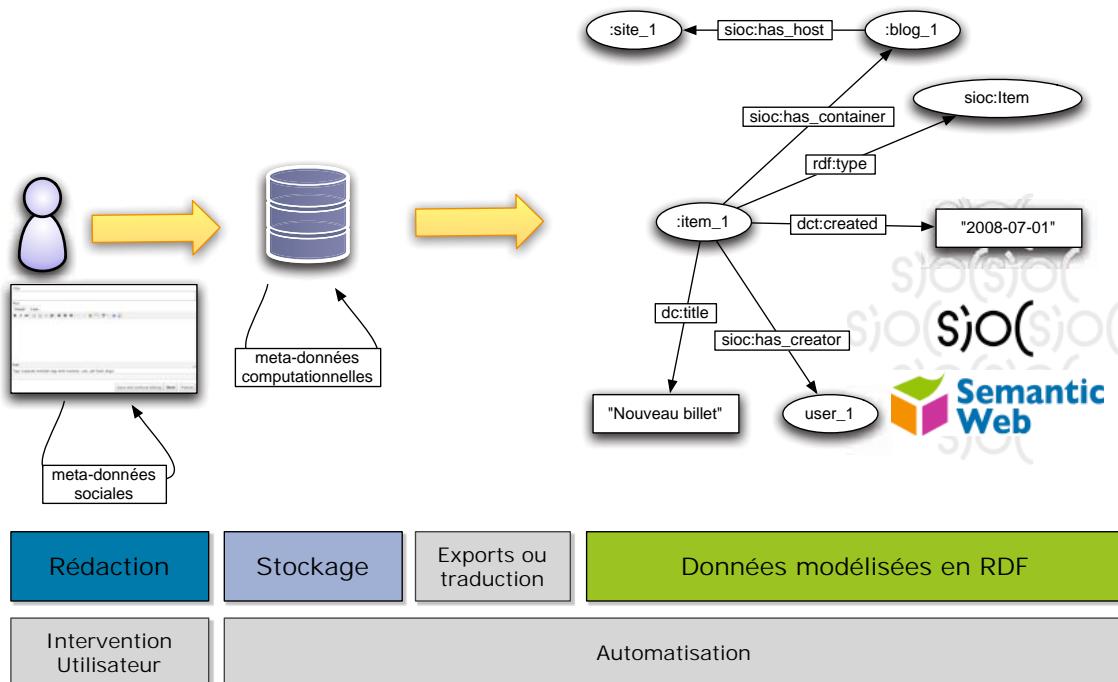


Figure 4.1: Processus générique de production de données RDF depuis des services Web 2.0

Nous allons maintenant détailler différentes méthodes logicielles permettant la production automatique de ces annotations. Celles-ci sont utilisées au sein de notre système mais également de manière plus large sur le Web. Si nous présentons dans cette partie une approche complètement automatisée, nous verrons par la suite que la modélisation des contenus est plus complexe et nécessite généralement une intervention supplémentaire (c'est du moins le choix que nous avons fait) (Section 4.2.1, page 148). C'est également le cas pour le passage du processus classique de *tagging* à l'indexation sémantique (Section 4.3, page 170).

4.1.2 Implémentation au sein de la plate-forme Hermès

Pour rappel, notre système d'information se compose de trois outils à partir desquels nous souhaitons modéliser ces annotations socio-structurelles : un agrégateur de flux RSS, un système de blog et un serveur de wikis (Section 2.1.1, page 50). L'automatisation des exports va donc permettre à partir de ces trois outils d'obtenir un graphe uniifié de données RDF comme nous l'avons présenté dans un précédent chapitre (Figure 2.11, page 75).

Nous allons maintenant présenter les différentes méthodes d'export associées à chaque outil. Cette partie est volontairement technique, les principes de bases ayant été présentés dans la section précédente.

De RSS et Atom vers SIOC

Si RSS offre un premier modèle pour l'interopérabilité entre services Web 2.0, nous avons montré en quoi il nous semblait trop limité et de quelle manière SIOC permettait de pallier à ces limites (Section 3.1.2, page 86). Un premier besoin est donc de traduire les données RSS agrégées au sein de notre système en données RDF représentées avec SIOC. Pour ce faire, nous avons envisagé différentes solutions dans cette optique de production automatisée de données RDF (selon des vocabulaires particuliers) à partir de flux RSS ou Atom¹.

Une première solution est l'utilisation de la clause SPARQL CONSTRUCT (Section 1.1.3, page 25). Celle-ci permet la construction d'un graphe RDF à partir d'un (ou plusieurs) autre(s) graphe(s) RDF et d'un patron de requête donné. On peut ainsi la voir, quoique beaucoup moins riche (pas d'expression conditionnelle par exemple), comme le XSLT [Clark, 1999] du Web Sémantique dans le sens où elle permet la transformation de graphes RDF là où XSLT permet la transformation de documents XML. L'utilisation de cette clause est une première manière d'envisager la transformation d'un flux RSS 1.0 en graphe RDF utilisant SIOC, d'autant plus que l'on peut simplement aligner RSS 1.0 et SIOC, par exemple en considérant `rss :channel` comme une sous-classe de `sioct :Forum`². On trouvera en annexe de ce mémoire la requête associée à cette transformation (Annexe A, page 229). Si l'utilisation de CONSTRUCT permet la traduction de flux RSS 1.0 vers SIOC, elle ne s'applique cependant qu'à cette version de RSS. Ses autres versions ne sont en effet pas basées sur RDF et ne peuvent donc pas être traitées par un processeur de requêtes SPARQL. Il est donc nécessaire de passer par d'autres méthodes pour la conversion de flux RSS *non-1.0* et Atom.

Une seconde possibilité est l'utilisation directe de XSLT pour la traduction de flux XML (RSS et Atom) en données RDF. C'est d'ailleurs de cette manière que [Karger et Quan, 2004] transforment les flux RSS de différentes versions en flux RSS 1.0 ou que GRDDL propose communément d'extraire des données RDF de documents XML ou XHTML [Gandon, 2007]. Si les différentes versions XML de RSS et Atom reposent sur des DTDs ou schémas XML connus permettant l'écriture de feuilles de styles appropriées, la flexibilité de la sérialisation RDF/XML fait que la production d'une feuille de style pour des flux RSS 1.0 est relativement complexe du fait du nombre de cas à prendre en compte. En pratique cependant, la plupart de ces flux suivent un modèle standard ce qui permet d'utiliser une unique feuille de style. Une autre limite de cette approche est l'impossibilité de gérer des flux RSS 1.0 non sérialisés en XML. Cependant, cet aspect relativement théorique est également à nuancer puisqu'il nous est apparu que le nombre de flux RSS 1.0 présents en ligne et non disponibles en RDF/XML est quasi-nul³.

¹Notons qu'en pratique, nous avons systématiquement choisi dans notre agrégateur d'intégrer la version RSS d'un flux lorsque ces deux formats étaient disponibles.

²Notons que nous ne prenons pas ici en compte la notion d'autorité dans la gestion d'une hiérarchie de classes distribuée, problème soulevé par [Hogan *et al.*, 2008].

³Mis à part quelques exemples, nous n'en avons en fait pas trouvé.

Une dernière solution est l'utilisation d'une API permettant de manipuler des données RSS ou Atom. Ce type d'API permet généralement la transformation de flux RSS en objets (au sens Programmation Orientée Objet, POO par la suite) qu'il est possible de manipuler et d'exporter en RDF via des scripts dédiés. Si cette solution est relativement simple à mettre en place, elle reste malgré tout à nouveau limitée à des flux sérialisés uniquement en XML et selon un schéma prédéfini. Les problèmes évoqués plus tôt (principalement théoriques cependant) ne sont donc pas résolus mais nous avons cependant opté pour cette solution notamment par volonté (1) de ne pas nous aventurer dans les transformations XSL et (2) de re-utiliser une partie des développements effectués autour de l'agrégateur RSS originel (Section 2.1.2, page 53). Ainsi, nous avons utilisé l'API MagpieRSS⁴, permettant de manipuler des flux RSS avec PHP. C'est à partir de cette API que nous avons implémenté l'agrégeur RSS utilisé au sein de la plate-forme. Il a donc été possible d'ajouter simplement un processus de traduction vers SIOC en tant que *plug-in* de la plate-forme d'origine, toujours dans cette idée de système de médiation au-dessus d'outils existants. Notre script de conversion est de ce fait assez léger (une trentaine de lignes de code), l'essentiel étant géré par l'agrégeur et l'API en question. Nous verrons par la suite de quelle manière nous avons enrichi cet export avec l'ajout d'annotations destinées à l'indexation sémantique des contenus issus de flux RSS (Section 5.3.2, page 209).

Quoi qu'il en soit, ces trois solutions, chacune avec leurs avantages et leurs limites, permettent de traduire des flux RSS en données RDF basées sur SIOC, comme l'illustre la figure qui suit (Figure 4.2, page 141).

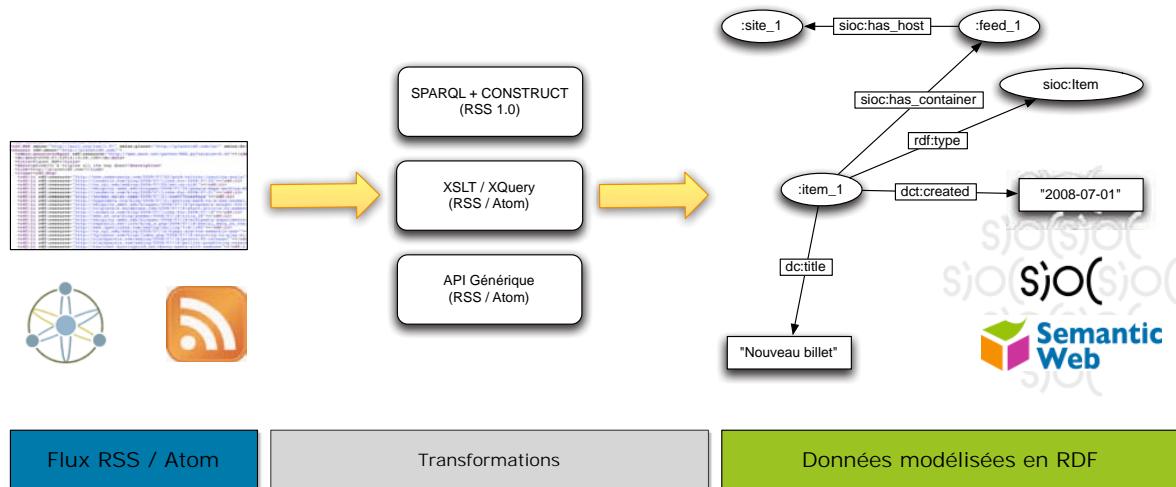


Figure 4.2: Processus de traduction RSS / Atom vers SIOC

Ce processus de traduction nous permet en outre d'identifier un réel besoin en termes d'unification des mondes RDF et XML. Des efforts récents se sont concentrés sur cette pro-

⁴<http://magpierss.sourceforge.net/>

blématique comme par exemple :

- une approche permettant de combiner transformations XSL et SPARQL, proposée par [Berrueta *et al.*, 2008] ;
- XSPARQL [Akhtar *et al.*, 2008], qui propose une extension à la fois de SPARQL et de XQuery pour permettre des requêtes combinant ces deux langages. Cette proposition identifie par ailleurs un ensemble de cas d'utilisation relatifs à ces processus de traduction de données XML en RDF [Passant *et al.*, 2009a] ;
- les extensions spécifiques de certains moteurs SPARQL comme par exemple celles proposées par Corese qui permettent de combiner SPARQL, XSLT et Xquery⁵.

Enfin, citons également SPARQL++ [Polleres *et al.*, 2007] qui vise à proposer des méthodes plus poussées de traduction de graphes RDF pour pallier à certaines limites de CONSTRUCT par rapport à XSLT mais aussi, comme nous l'avons déjà évoqué précédemment, GRDDL qui permet d'extraire un ensemble d'annotations RDF depuis n'importe quel document XML (Section 1.1.1, page 12).

Annotations socio-structurelles avec SIOC depuis les blogs et les wikis

Contrairement à la traduction de données RSS vers SIOC, pour laquelle les outils peuvent bénéficier d'un format source semi-structuré et standardisé (malgré ses différentes versions), la production d'annotations socio-structurelles en RDF depuis des services Web 2.0 génériques est plus complexe. En effet, chaque outil ou service dispose généralement de sa propre structure pour modéliser ses données, qu'il s'agisse d'informations documentaires et structurelles (titre d'un billet, appartenance d'un document à un wiki donné, etc.) ou des comportements sociaux qui s'y rapportent (commentaire, édition d'une page, etc.). Les structures de bases de données sont ainsi distinctes sous Drupal, Wordpress tout comme le sont les APIs (aussi bien en termes de requêtes que de structuration des résultats) sous Flickr ou Twitter. Si l'utilisation de SIOC permet de résoudre cette hétérogénéité en proposant une représentation standardisée de telles informations, elle implique également en premier lieu le développement d'applications spécifiques pour chaque outil et service. On peut certes imaginer utiliser les flux RSS associés à ces services pour représenter ces informations avec SIOC comme nous l'avons étudié précédemment mais cet export restera alors limité aux derniers contenus publiés (Section 3.1.2, page 86).

Dans notre contexte, nous nous sommes plus particulièrement intéressés à la production d'annotations socio-structurelles depuis nos plates-formes de blogs et de wikis, celles-ci étant abondamment utilisées (Section 2.1.4, page 59) et comportant de ce fait un grand nombre de documents qu'il nous semble important de représenter avec SIOC au sein d'un tel écosystème sémantique. Bien que ces deux outils soient basés sur le système Drupal⁶, les structures de bases de données sous-jacentes restent distinctes. Ainsi, une seule table est nécessaire au stockage des blogs et de leurs billets, alors que trois d'entre elles sont utilisées pour les wikis. À celles-ci vient également s'ajouter une table partagée pour la représentation des utilisateurs. Afin de passer de ces structures hétérogènes à un modèle commun d'annotations sémantiques, nous avons participé au développement d'un *plug-in* générique

⁵http://www-sop.inria.fr/edelweiss/software/corese/v2_4_1/manual/new.php

⁶<http://drupal.org>

permettant l'export de données SIOC depuis Drupal⁷. Ce *plug-in* permet ainsi de représenter de manière complètement autonome le graphe d'annotations socio-structurelles associé à chaque document créé via ce système. L'export se fait de plus en temps réel, *i.e.* chaque document créé dispose immédiatement de son graphe RDF associé.

Afin de coller au plus près de nos besoins, ce *plug-in* public a en outre été adapté en fonction de certaines caractéristiques spécifiques à notre plate-forme :

- d'une part, le module wiki étant un module spécifique à notre système d'information, l'export de ses données n'est pas géré par le *plug-in* public. Nous avons donc défini différents alignements entre les structures de bases de données relatives aux wikis et les propriétés et relations définies dans SIOC. Par exemple, la table de jointure entre les wikis et leurs pages permet d'établir des liens `sioc:has_container` entre les instances associées (respectivement de `sioc:WikiArticle` et `sioc:Wiki`)
- d'autre part, le *plug-in* public se contente de créer des instances de `sioc:Item` sans spécifier un type plus précis. Puisque nous souhaitons dans nos requêtes pouvoir distinguer le type de contenu (Section 5.2.1, page 196), nous avons précisé celui-ci en typant les contenus exportés avec le module Types de SIOC (Section 3.1.3, page 92). De ce fait, notre implémentation produit soit des instances de `sioc:BlogPost` soit de `sioc:WikiArticle`, en fonction de l'outil utilisé et du type de document créé.

Contrairement aux flux RSS qui proviennent de l'extérieur et pour lesquels nous ne représentons pas le créateur de chaque élément de flux, nous nous attachons ici à fournir une représentation RDF de celui-ci, à la fois d'un point de vue de son compte en ligne (`sioc:User`) et de la personne physique associée (`foaf:Person`). À chaque utilisateur de la plate-forme sont donc associées deux URIs distinctes et un graphe d'annotations RDF associé, par exemple :

- `http://athena.der.edf.fr/?q=sioc/user/1#user`, identifiant de l'utilisateur en tant qu'entité virtuelle ;
- `http://athena.der.edf.fr/?q=sioc/user/1#person`, identifiant de la personne physique correspondante ;
- `http://athena.der.edf.fr/?q=sioc/user/1` fichier RDF associé listant certaines propriétés associées à ces deux identifiants (nom, e-mail, URL du blog ...).

Ce *plug-in* permet ainsi d'obtenir automatiquement, pour chaque contenu de blog ou de wiki, un graphe d'annotations RDF associé comme le montre la figure qui suit (Figure 4.3, page 144).

De manière plus précise, la figure qui suit représente la traduction d'un billet de blog donné en instance de `sioc:BlogPost` grâce à ce *plug-in* (Figure 4.4, page 144).

4.1.3 API SIOC et passage à l'échelle de l'annotation socio-structurelle de documents Web 2.0

Comme nous l'avons vu dans la section précédente, la production d'annotations socio-structurelles depuis des services Web 2.0 implique le développement de *plug-ins* ou outils spécifiques depuis ces différents services. Pour faciliter ces développements et dans l'op-

⁷<http://drupal.org/project/sioc>

CHAPITRE 4 : ANNOTATIONS SÉMANTIQUES ET PEUPLEMENT COLLABORATIF D'ONTOLOGIES

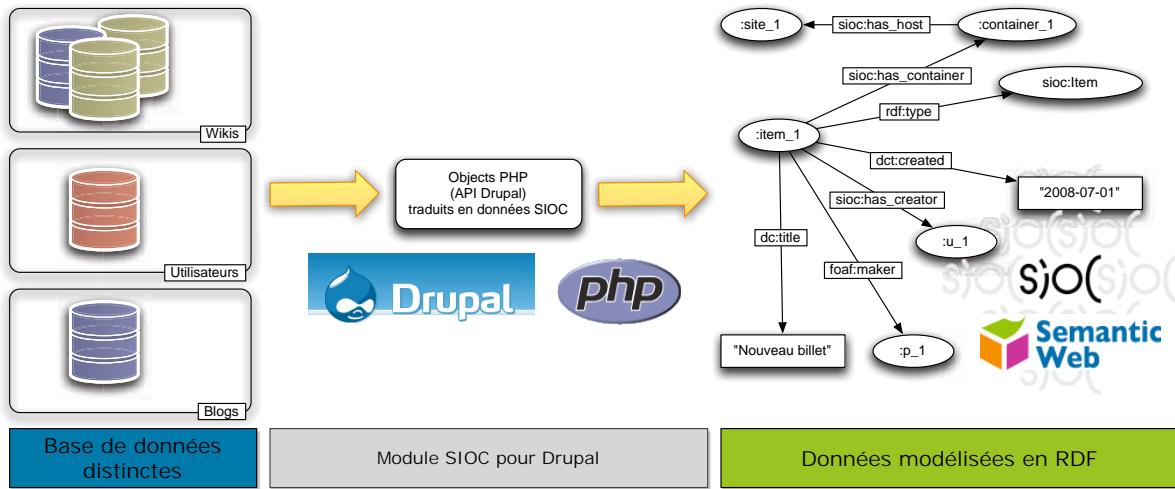


Figure 4.3: Processus de traduction des données de blogs et wikis vers SIOC

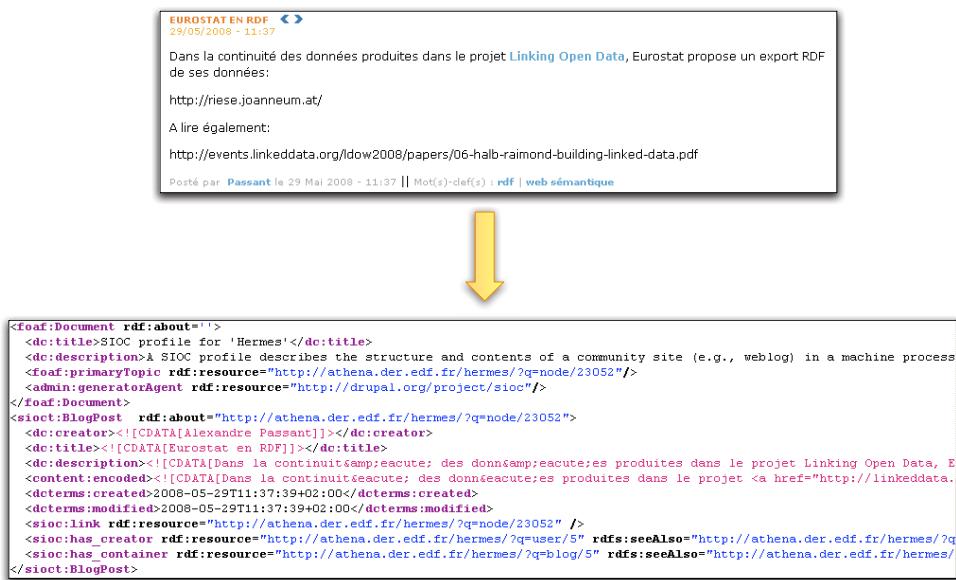


Figure 4.4: Exemple de traduction d'un billet de blog vers SIOC

tique de généraliser la production de telles annotations à l'échelle du Web, nous avons ainsi mis en place une API dédiée à la production de données RDF représentées avec SIOC.

Par nature, la programmation dédiée au Web Sémantique est généralement orientée triplets. Des APIs comme Jena⁸, Redland⁹ ou RAP¹⁰ proposent ainsi par défaut de définir un modèle (ou graphe) RDF auquel on ajoute un certain nombre de triplets. Le code ci-dessous montre par exemple l'utilisation de Jena pour créer une instance de `sioc:Post` associée à diverses propriétés et relations (Listing 4.1, page 145).

```
String postURL = "http://example.org/blogpost";
String siocPost = "http://rdfs.org/sioc/ns#Post";
String dcCreator = "http://purl.org/dc/terms/creator";
String rdfType = "http://www.w3.org/1999/02/22-rdf-syntax-ns
#type";

Model model = ModelFactory.createDefaultModel();
Resource myBlogPost = model.createResource(postURL);
myBlogPost.addProperty(rdfType, siocPost);
myBlogPost.addProperty(dcCreator, "Alexandre Passant");
```

Listing 4.1: Utilisation de Jena pour représenter des données RDF

Même si cette approche est assez intuitive lorsque l'on est habitué aux représentations du Web Sémantique, puisque l'on construit un graphe en instanciant des ressources (définies par des URIs) et en leur assignant diverses propriétés et relations, elle est peu naturelle pour les développeurs adeptes des principes de POO. Tout comme nous pensons que le succès du Web Sémantique passera par des interfaces simples en termes de publication et de visualisation de données RDF, tel que nous le rappelons régulièrement au sein de ce mémoire, nous estimons qu'il en sera de même pour l'adoption de celui-ci par les développeurs. Dans cet objectif, nous avons donc développé une API PHP proposant une interface orientée-objet pour la production de graphes d'annotations RDF basés sur SIOC¹¹. Notre API définit ainsi un ensemble de classes PHP (ainsi que différentes fonctions associées) alignées avec le noyau de SIOC (Section 3.1.3, page 89). Par exemple, la classe (PHP) `SIOCSite` permet de créer une instance (RDF) de `sioc:Site` et ses méthodes permettent d'ajouter les utilisateurs (`sioc:User`) et forums (`sioc:Forum`) associés. Un extrait du code de cette API se trouve en annexe de ce mémoire (Annexe 4.1.3, page 143).

L'API dispose également d'une classe `SIOCExporter` qui permet la génération du graphe RDF associé. Celle-ci dispose de deux méthodes distinctes :

- une première (`createRDF`) permettant simplement de générer le graphe RDF sérialisé en RDF/XML, qui peut alors être sauvegardé au sein d'un fichier ;
- une seconde (`output`) dédiée à l'utilisation de l'API sur le Web et permettant d'afficher le graphe en se chargeant de définir le type de contenu approprié au niveau du serveur

⁸<http://jena.sf.net>

⁹<http://librdf.org>

¹⁰<http://www4.wiwiiss.fu-berlin.de/bizer/rdfapi/>

¹¹<http://wiki.sioc-project.org/index.php/PHPExportAPI>

Web, i.e. application/rdf+xml¹².

Outre les facilités offertes au développeur pour la production de graphes d'annotations, cette API permet de s'assurer que les données produites sont conformes à l'ontologie et aux bonnes pratiques associées (Section 3.1.4, page 94). De plus, en cas d'évolution de SIOC, une simple mise à jour de l'API est nécessaire pour adapter les contenus produits à la nouvelle version de l'ontologie. Cette API permet donc aux développeurs de se soucier ni de l'ontologie en elle-même, ni des principes de modélisation RDF.

Parmi les autres fonctionnalités que propose cette API, signalons la production automatique de liens rdfs:seeAlso entre contenus exportés au sein d'un même site (documents, utilisateurs, conteneurs), suivant ainsi les bonnes pratiques de publication de données interconnectées sur le Web Sémantique [Bizer et al., 2007b]. Par exemple, pour un billet comportant un commentaire, en plus d'exporter des relations entre le billet et (1) son commentaire (sioc:has_reply), (2) son conteneur (sioc:has_container) et (3) son créateur (sioc:has_creator), différents liens rdfs:seeAlso seront produits vers les graphes d'annotations RDF correspondants. Ceci permet à des navigateurs RDF comme Tabulator¹³ [Berners-Lee et al., 2006] de profiter de ces liens pour découvrir de nouvelles données au sujet de ces différentes instances mais surtout aux approches de crawling de découvrir un réseau complet de données à partir d'un unique document [Harth et al., 2006]. En effet, la présence de ces liens au sein des graphes créés par l'API permet à partir de n'importe quel point d'entrée de remonter jusqu'à l'instance de sioc:Site et à partir de là de retrouver l'ensemble des données exportées depuis un site donné (Figure 4.5, page 146).

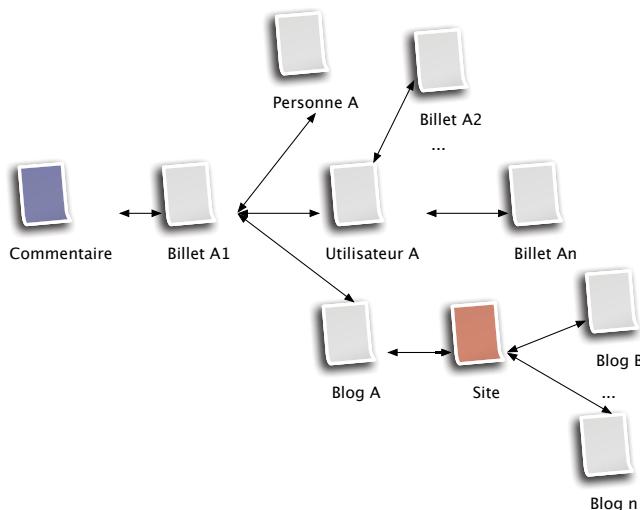


Figure 4.5: Représentation de liens rdfs:seeAlso entre documents RDF avec l'API SIOC

Si l'API peut-être utilisée à partir de données brutes, une utilisation plus judicieuse est de la coupler avec l'API fournie par le service que l'on souhaite exporter. Dans ce contexte,

¹²<http://www.ietf.org/rfc/rfc3870.txt>

¹³<http://www.w3.org/2005/ajar/tab>

on utilise (1) l'API du service pour transformer les données source en objets PHP puis (2) l'API SIOC pour transformer ces objets PHP en données RDF. On profite ainsi d'une double couche d'abstraction qui permet au développeur de se soucier ni des formats internes, ni du modèle RDF souhaité en sortie. C'est cette solution qui a été privilégiée au sein de l'exporteur SIOC vBulletin¹⁴ (outil pour la mise en place de forums de discussions) ou encore pour mettre en place le service d'export de données FOAF et SIOC depuis Flickr que nous avons développé [Passant, 2008b].

Cette API est aujourd'hui utilisée dans différentes applications, qu'il s'agisse d'exporteurs et *plug-ins* pour des services existants (vBulletin, Flickr, MediaWiki ...) ou d'applications spécifiques comme SMOB [Passant *et al.*, 2008]. Cette dernière application bénéficie ainsi de l'API pour proposer un système de microblogging ouvert et décentralisé, et qui plus est représentant l'ensemble des données produites avec SIOC, permettant ainsi leur intégration au sein de la *SIOC-o-sphère*. Notons également que suite à la mise à disposition de cette API, d'autres APIs SIOC ont été proposées par la communauté, notamment en Java¹⁵ et en Perl¹⁶, avec des principes similaires. Ces différentes APIs poursuivant toutes le même but nous permettent ainsi d'envisager une multitude de nouveaux services produisant des données représentées avec SIOC, accentuant encore plus sa présence sur le Web (Section 3.1.6, page 101).

Pour conclure, signalons que l'API que nous proposons ici a été développée pour les besoins précis de SIOC et n'est donc pas aussi flexible que les APIs permettant de généraliser la définition de classes (au sens POO) à partir de tout modèle RDFS ou OWL. À ce sujet, citons ActiveRDF [Oren *et al.*, 2007] (Ruby On Rails), le module schemagen¹⁷ de Jena (Java) ou encore RAP¹⁸ (PHP). Ces solutions, plus génériques, sont cependant plus lourdes et nous avons préféré pour cette API SIOC proposer un module indépendant et léger (un seul fichier), plutôt que de se baser sur une API plus complexe dont la générnicité n'aurait pas été utile dans notre contexte. Dans le cas d'une API orientée lecture, la démarche est différente puisqu'il est nécessaire d'interpréter le graphe RDF, étape qui s'avère plus complexe. Le module d'import SIOC pour WordPress¹⁹ utilise par exemple RAP, tout comme PHOAF²⁰, API que nous avons développée pour permettre de manipuler simplement des fichiers FOAF via des méthodes de POO. Cette dernière est notamment utilisée dans FOAF-Map²¹ [Passant, 2006], un des premiers services de *mash-up* sémantique, permettant de visualiser un réseau social modélisé avec FOAF sur une carte GoogleMap et de naviguer au sein des différents profils ainsi représentés mais aussi d'identifier des personnes partageant des centres d'intérêts similaires (Figure 4.6, page 148).

¹⁴<http://www.vbulletin.com/>

¹⁵<http://mavenrepo.fzi.de/semweb4j.org/site/sioc-api/index.html>

¹⁶<http://search.cpan.org/~geewiz/SIOC-v1.0.0/>

¹⁷<http://jena.sourceforge.net/how-to/schemagen.html>

¹⁸<http://www4.wiwiiss.fu-berlin.de/bizer/rdfapi/>

¹⁹http://wiki.sioc-project.org/w/SIOC_Import_plug-in

²⁰<https://gna.org/projects/phoaf>

²¹<http://foafmap.net>



Figure 4.6: Cartographie de réseaux sociaux avec FOAFMap

4.2 UFOWIKI POUR LE PEUPLEMENT D'ONTOLOGIES MÉTIER

4.2.1 Wikis sémantiques et peuplement d'ontologies : intérêt et état de l'art

Un autre aspect important à prendre en compte dans notre contexte est celui du peuplement d'ontologies, dans cet objectif de représentation interprétable de données métier (Section 3.2, page 104). Bien que ce processus puisse dans certains cas être assisté ou semi-automatisé via l'analyse de corpus de textes [Kiryakov *et al.*, 2004] [Amardeilh *et al.*, 2005], il peut aussi se baser sur une approche manuelle de production d'annotations confiée à une équipe dédiée. Celle-ci est généralement restreinte et peut être composée aussi bien d'experts du domaine que de spécialistes en ingénierie des connaissances. Si cette collaboration permet de s'assurer de la qualité des données produites, à la fois en termes de valeur intellectuelle (via l'expert du domaine) et de qualité sémantique (via les spécialistes en ingénierie des connaissances), elle rend délicat le maintien et l'évolution de bases de connaissances à flux tendu. Ce maintien s'effectuant en effet en vase clos, via une équipe restreinte et prédefinie, il implique l'impossibilité pour des contributeurs externes de faire profiter l'équipe de leur expertise, à partir du moment où ils ne font pas partie du groupe destiné à maintenir ces bases de connaissances. Un autre point à prendre en compte est celui du transfert de connaissances, notamment lorsque le ou les experts du domaine quittent l'entreprise. D'une part, ce processus peut être relativement long selon les domaines et d'autre part, certains corps de métier peuvent ne plus exister, rendant ce transfert encore plus délicat. De plus, signalons que les outils associés (Protégé²² par exemple) sont en général destinés à un public avancé, ne serait-ce qu'en termes d'interface utilisateur.

On peut ici faire un parallèle avec ce que nous avons présenté précédemment dans ce mémoire, à savoir l'avantage des outils Web 2.0 en entreprise (et des pratiques liées) par rapport à des structures informationnelles classiques (équipe restreinte, *workflow* ...) (Section 2.1.1, page 50). Si ceux-ci permettent une évolution ouverte et spontanée de l'information,

²²<http://protege.stanford.edu>

il nous a paru intéressant de réfléchir à des principes similaires pour une ingénierie des connaissances collaborative et ouverte. En conséquence, dans ce contexte d'Entreprise 2.0, nous avons étudié le rapprochement entre ces processus Web 2.0 et les principes de peuplement d'ontologies. C'est au travers des wikis sémantiques et plus particulièrement au sein d'un nouvel outil de ce type, UfoWiki (Section 4.2.2, page 154), que nous avons étudié et mis en place cette convergence.

Tout comme le Web Sémantique est une extension du Web, les wikis sémantiques sont une extension des wikis permettant d'ajouter à ceux-ci des méthodes de représentation formelle des connaissances. Ces représentations peuvent se concentrer selon les outils sur la structure ou sur le contenu des pages et conservent dans tous les cas les principes d'utilisation des wikis (Section 1.2.2, page 35). Alors que nous avons insisté dans la section précédente sur les annotations socio-structurelles (Section 4.1, page 138), nous allons ici considérer principalement les wikis permettant la modélisation du contenu des pages, *i.e.* la formalisation de connaissances métier. Ceux-ci permettent ainsi d'établir un pont entre le Web de documents et le *Web de Données* (Section 1.1.4, page 27), comme l'illustre à nouveau la figure ci-après (Figure 4.7, page 149). Le wiki devient ainsi le support d'un ensemble de données connectées via différents graphes d'annotations, permettant à terme l'enrichissement des fonctionnalités offertes. Nous voyons donc les wikis sémantiques comme des interfaces permettant, du fait de leur philosophie (ouverture, collaboration ...) le peuplement d'ontologies par et pour tous [Passant et Laublet, 2008e].

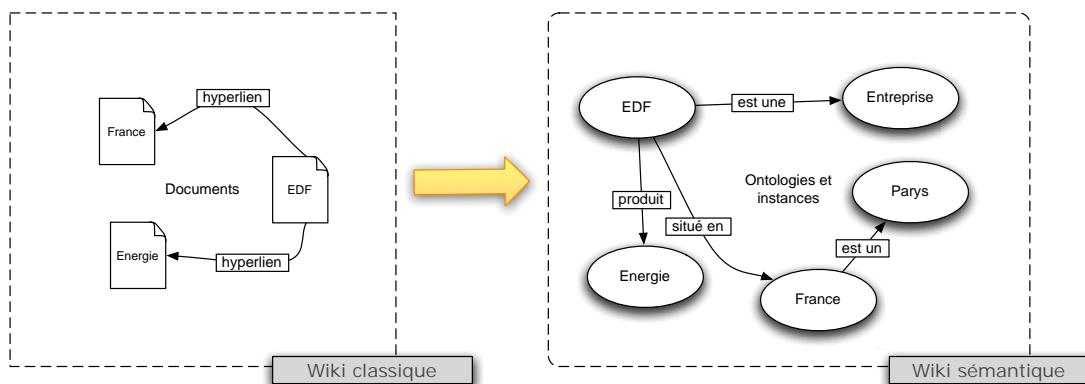


Figure 4.7: Du wiki au Web Sémantique

Depuis le premier *workshop* consacré aux wikis sémantiques [Völkel et Schaffert, 2006], de nombreux prototypes suivant cette approche de gestion de base(s) de connaissance(s) ont vu le jour. Nous allons ici présenter ceux qui nous semblent les plus pertinents par rapport à l'état de l'art du domaine. Celui-ci n'est pas exhaustif et nous invitons le lecteur souhaitant approfondir le sujet à parcourir les actes de la série SemWiki²³ ou à consulter [Buffa *et al.*, 2008] qui dresse également un état de l'art assez complet sur le sujet. Rappelons

²³<http://semwiki.org>

à nouveau que nous nous concentrons ici sur l'utilisation de wikis pour le peuplement d'ontologies de domaine et que nous ne considérons pas des prototypes orientés métadonnées socio-structurelles, comme par exemple SweetWiki [Buffa *et al.*, 2008] que nous évoquerons dans la partie de ce chapitre consacrée au *tagging* (Section 4.3.2, page 175). Nous mettons également de côté les approches permettant l'extraction d'ontologies ou de bases de connaissances à partir de wikis existants, comme par exemple DBpedia [Auer *et al.*, 2007] ou les nombreux travaux autour de l'utilisation de Wikipedia en tant que base de connaissances formelle²⁴[Nakayama, 2008] [Wu et Weld, 2008].

Platypus

Platypus²⁵ [Tazzoli *et al.*, 2004] est certainement le premier prototype de wiki sémantique recensé. Il permet l'annotation sémantique au travers de formulaires *prédicat / objet* associés à chaque page (mais distincts de la zone d'édition principale), un certain nombre d'ontologies prédéfinies (FOAF, DOAP, DublinCore...) permettant le choix du prédicat. Il est également possible d'insérer directement des annotations RDF/XML ou Turtle au sein des documents. Il n'y a malheureusement pas de distinction entre le document et le concept associé, le sujet de chaque assertion étant l'URL de la page wiki, alors que les formulaires peuvent laisser penser le contraire (par exemple en proposant le prédicat `foaf:knows`, dont le domaine est `foaf:Agent`, disjoint de la notion de document, *i.e.* `foaf:Document`). Ce manque de distinction (que nous avons introduit en introduction de ce mémoire (Section 1.1.2, page 16)) peut conduire à des inconsistances, notamment si un utilisateur crée une annotation en considérant la page comme sujet alors qu'un autre annote en considérant le concept.

En termes de valeur ajoutée, les annotations sont utilisées pour enrichir l'interface de visualisation du wiki. Platypus permet ainsi à l'utilisateur de naviguer entre pages via ces annotations en complément de la navigation hypertexte classique. Par exemple, il est possible de passer d'une page à une autre lorsque celles-ci (ou les concepts sous-jacents) sont liées par une propriété quelconque via les annotations.

SemPerWiki

SemPerWiki²⁶ – *Semantic Personal Wiki* [Oren, 2005] – est un wiki personnel qui s'utilise sur le poste de travail, dans la mouvance des outils du *Semantic Desktop*. Il se rapproche plus du bloc-notes personnel avec un mode d'édition wiki que d'un véritable wiki dans la mesure où l'aspect participatif n'entre pas en jeu. Les annotations doivent être saisies directement par l'utilisateur en RDF (syntaxe Turtle) au sein de la page, sans assistance malgré certains préfixes prédéfinis. Ceci destine l'outil principalement à des utilisateurs avancés. Tout comme Platypus, le sujet des annotations est par défaut l'URL de la page, posant les mêmes problèmes que précédemment. Il est cependant possible d'expliciter le sujet de chaque triplet ce qui permet de modéliser au sein d'une page wiki des informations à propos de n'importe quel concept.

Les annotations produites sont également utilisées pour enrichir la navigation. Semper-Wiki propose également un système de requête au sein même des pages, sous la forme de

²⁴A ce sujet, on peut consulter le projet Wikipedia Lab - <http://wikipedia-lab.org/>

²⁵<http://platypuswiki.sourceforge.net/>

²⁶<http://www.eyaloren.org/semperviki.html>

triplets où le concept recherché est remplacé par un ?. Par exemple la requête identifiée par ? `rdf:type foaf:Organisation` listera l'ensemble des organisations recensées dans le wiki.

Semantic MediaWiki

Semantic MediaWiki²⁷ (SMW par la suite) [Krötzsch *et al.*, 2006] est une extension du moteur MediaWiki²⁸, propulsant entre autres Wikipedia. Son mode d'annotation ne se base pas sur l'écriture directe de triplets mais sur une extension de la syntaxe wiki, facilitant la tâche d'appropriation. Par exemple, pour indiquer qu'EDF est situé en France, on saura EDF est implantée en [[se_situe_en::France]], texte qui sera traduit en l'assertion `onto:EDF onto:se_situe_en onto:France` à partir du moment où ce texte est saisi sur une page relative à EDF, le sujet de chaque triplet étant par défaut le concept associé à la page en cours. SMW distingue donc document et concept, en définissant une URI pour chaque concept, différente de l'URL de la dite page²⁹. L'utilisation d'un système d'annotations totalement ouvert, conforme avec la philosophie wiki, permet de considérer SMW comme un wiki sémantique dédié non seulement au peuplement d'ontologies, mais aussi au maintien des modèles associés (voire à leur définition). Cependant, cette ouverture conduit rapidement à des problèmes d'hétérogénéité sémantique. Alors qu'un utilisateur choisira la syntaxe `se_situe_en` pour modéliser une relation de localisation, un second pourra préférer `est_localisé_en`. Les annotations produites seront donc totalement indépendantes, leur intérêt se trouvant restreint puisque sans sémantique commune. Le même problème se pose pour la gestion des classes, celles-ci étant déterminées à partir des catégories assignées aux pages. Notons cependant que SMW offre la possibilité d'aligner certaines relations et catégories avec des ontologies existantes et que les modèles créés à partir du wiki sont exportés en OWL-DL et donc réutilisables dans d'autres applications.

Pour tirer parti des annotations, SMW propose différents modes de navigation avancés. C'est notamment le cas des pages `Property` listant l'ensemble des triplets utilisant une propriété particulière³⁰ ou des pages `Special:Browse` listant l'ensemble des assertions relatives à un concept³¹. Mais surtout, SMW offre un système de requêtes avancées, modélisées avec une syntaxe wiki particulière et permettant l'inclusion de réponses à des requêtes complexes au sein même des pages wiki. Il est par exemple possible de lister l'ensemble des événements recensés au sein d'un wiki, comme le montre le code suivant³² (Listing 4.2, page 152).

IkeWiki

IkeWiki³³ [Schaffert, 2006] se base quant à lui sur des ontologies prédéfinies, permettant de s'assurer de la qualité sémantique des annotations produites. L'utilisateur est assisté au

²⁷<http://semantic-mediawiki.org/>

²⁸<http://mediawiki.org>

²⁹Les premières versions de l'outil ne faisaient cependant pas cette distinction.

³⁰e.g. http://semanticweb.org/wiki/Property:Swoogle_hits

³¹e.g. <http://semanticweb.org/wiki/Special:Browse/SIOC>

³²Code source issu de <http://semanticweb.org/wiki/Events>.

³³<http://ikewiki.salzburgresearch.at/>

```

{{#ask:[[Category:Event]] [[end date:::>{{CURRENTYEAR}}-{{CURRENTMONTH}}-{{CURRENTDAY}}]] |
?title = Name|
?has location city = City|
?has location country = Country|
?Start date|
?End date|
?Category:Conference = C|
?Category:Workshop = W|
format=table|limit=50|sort=end date
}}

```

Listing 4.2: Requête interne au sein de MediaWiki

moment de la pose de liens entre pages : un certain nombre de prédictats lui sont proposés, qui sont ensuite traduits en relations entre les concepts associés à ces pages. De plus, chaque page peut être associée à une classe via un parcours de l'ontologie (ou plutôt de la taxonomie des classes), le concept associé à la page étant alors défini comme instance de la classe en question. IkeWiki utilise également les annotations produites pour l'aide à la navigation, notamment en affichant pour chaque page la hiérarchie de classes associées. L'outil bénéficie également de capacités d'inférence, en gérant les notions de sous-classes et sous-propriétés pour l'aide à la navigation et permet l'utilisation de requêtes SPARQL pour interroger la base de connaissance.

Notons également qu'IkeWiki modélise également un certain nombre d'annotations socio-structurelles à l'aide d'un vocabulaire propre et propose en plus une modélisation des discussions associées aux pages wikis, en utilisant SIOC³⁴. C'est à notre connaissance le seul outil à modéliser ses pages de discussions en RDF. Ceci nous semble particulièrement intéressant dans la mesure où l'on peut ainsi identifier la communauté qui s'établit autour d'un concept donné. S'il s'agit pour l'instant d'un simple export, il y a selon nous un fort intérêt à considérer une approche plus poussée permettant de modéliser le discours argumentatif associé.

OntoWiki

Basé sur Powl³⁵ [Auer, 2005], éditeur d'ontologies en ligne, OntoWiki [Auer *et al.*, 2006] est à la frontière entre le wiki sémantique et l'éditeur classique d'ontologies et de bases de connaissances. En effet, OntoWiki n'utilise pas à strictement parler de pages *wikis* comme dans les outils précédents mais propose un système de vues virtuelles établies au dessus d'une ou plusieurs bases de connaissances. Chaque graphe ou triplet est ainsi représenté via un fragment de page dynamique qui lui sert à la fois d'interface de visualisation et d'édition. Ceci permet d'offrir différents niveaux de représentation et de granularité pour une navigation très souple. Par exemple, il est possible d'obtenir une page listant l'ensemble des instances d'une classe donnée, une seconde relative à l'ensemble des propriétés (et leurs

³⁴<http://tinyurl.com/6n2dg2>

³⁵<http://ontowiki.net/Projects/Powl>

valeurs) d'une instance particulière, ou bien encore une autre indiquant tous les triplets de la base de connaissance utilisant une certaine propriété. L'utilisateur est là aussi assisté lors de l'édition et la création de nouvelles assertions, avec notamment un système d'autocomplétion suggérant les instances possibles pour chaque propriété. Tout comme SMW, OntoWiki permet de faire évoluer le modèle dynamiquement mais utilise une approche plus formelle : chaque nouvelle propriété doit ainsi être définie comme `ObjectProperty` ou `DataTypeproperty`.

Les annotations produites sont utilisées d'une part pour produire les différentes vues et ainsi proposer une navigation directement liée à l'ontologie, mais aussi pour offrir à l'utilisateur un moteur de recherche enrichi de fonctionnalités sémantiques. Ainsi, la recherche plein-texte est couplée aux connaissances acquises, permettant à l'utilisateur de spécifier à quelle classe, instance ou propriété il veut restreindre celle-ci. De plus, un système de vue avancé permet de visualiser les annotations sous différentes formes : vue calendaire pour les données proposant des attributs temporels, géolocalisation pour celles associées à des coordonnées, etc. Une autre originalité d'OntoWiki réside dans ses aspects poussés de collaboration et de participation. Chaque modification – quelque soit sa nature – est tracée selon les principes de réification RDF, permettant d'identifier l'auteur ou la date de création de chaque assertion. Il est en outre possible de commenter et annoter chaque triplet, ceci dans une optique d'élaboration de réseaux sociaux autour de la construction de ressources ontologiques.

Malgré cette composante, OntoWiki reste plus proche – comme nous l'avons déjà mentionné – de l'éditeur d'ontologies en ligne que du wiki sémantique tel que nous le concevons, *i.e.* un outil offrant une certaine modularité entre le wiki plein-texte et l'annotation sémantique et reposant sur des ontologies, prédefinies ou évolutives.

AceWiki

Le système AceWiki³⁶ [Kuhn, 2008] est assez original dans sa démarche, puisque son approche d'annotations repose sur l'utilisation du modèle de langue naturelle contrôlée proposé par ACE – *Attempto Controlled English* [Fuchs *et al.*, 2000]. Ainsi, la saisie de chaque page wiki est assistée (ou contrainte, selon le point de vue) pour produire un contenu directement interprétable par le moteur wiki et par extension traduit en annotations RDF grâce à un processus d'alignement entre ACE et RDF(S)/OWL. Tout comme SMW ou Ontowiki, AceWiki permet de faire évoluer le modèle utilisé. Par exemple, une phrase comme `Country is a Class` induira la création d'une nouvelle classe `Country`, permettant ensuite l'utilisation de `France is a Country`, immédiatement traduit en l'annotation RDF correspondante.

L'aspect qui nous semble le plus intéressant dans AceWiki est l'utilisation de possibilités avancées de raisonnement, via l'intégration du raisonneur Pellet [Sirin *et al.*, 2007]. Toujours en utilisant ACE, les utilisateurs ont la possibilité de définir des contraintes de classes – par exemple `Every country has at least 1 city` – qui seront ensuite modélisées en OWL et utilisées pour valider la consistance du modèle lors de l'ajout de nouvelles annotations. Si des faits non conformes aux contraintes sont ajoutés, ceux-ci seront immédiatement

³⁶<http://attempto.ifi.uzh.ch/acewiki/>

ment notifiés à leur auteur. On peut ici reprocher un manque de traçabilité dans le raisonnement, puisque c'est toujours (dans la version actuelle) le dernier fait ajouté qui est considéré comme faux, sans que l'on puisse visualiser les autres faits qui ont conduit à cette conclusion³⁷. Un système de discussion associé à chaque fait non consistant serait une option intéressante, permettant d'introduire un aspect collaboratif dans la résolution de conflits.

S'il s'agit d'un prototype original, il est selon nous à considérer principalement si l'on souhaite bénéficier de possibilités avancées de raisonnement, l'utilisation d'ACE limitant les cadres d'utilisation de l'outil (ne serait-ce que pour des wikis non anglophones).

4.2.2 Objectifs, principes et architecture d'UfoWiki

En considérant l'état de l'art précédent et l'ensemble des critères que nous avons pris en compte, nous avons décidé d'implémenter UfoWiki³⁸ – *Unifying Forms and Ontologies in a Wiki* [Passant et Laublet, 2008a] [Passant et Laublet, 2008d] – nouvel outil de wiki sémantique. Celui-ci est une extension de la plate-forme développée initialement au sein d'Hermès (Section 2.1.2, page 55). Ainsi, plus qu'un simple outil de wiki, il s'agit d'un serveur de wiki, *i.e.* une application permettant à chaque utilisateur d'instancier un nouveau wiki sémantique pour sa communauté. Se baser sur le service existant nous permet de bénéficier des développements relatifs à la partie wiki classique de l'outil (rétro-liens, historique des versions, etc.) mais surtout de ne pas troubler les utilisateurs en les confrontant à un nouvel outil³⁹. Si l'outil n'est pas public, nous espérons que les idées défendues ici pourront être par la suite implémentées dans des outils comme ceux présentés dans la section précédente.

UfoWiki repose sur les principes suivants :

- *Une représentation des connaissances basée sur des ontologies prédéfinies.* Le fait de reposer sur des ontologies connues nous permet de nous assurer que les annotations produites sont conformes à des modèles préalablement identifiés. Ceci nous semble essentiel dans un contexte d'entreprise afin d'éviter les problèmes d'hétérogénéité sémantique et facilite de plus l'écriture de requêtes relatives aux annotations produites (Section 5.2.1, page 196). Bien que l'on puisse supposer qu'un modèle cohérent émerge de l'utilisation d'un wiki sémantique au modèle ouvert (tel que Semantic MediaWiki), comme c'est le cas pour les systèmes à base de tags, nous ne pouvons nous permettre dans un contexte industriel d'attendre cette masse critique qui permettra (éventuellement) l'émergence d'une sémantique commune. UfoWiki est donc capable, dès sa mise en place, de produire des annotations reposant sur des modèles ontologiques prédéfinis ;
- *Une interface utilisateur simplifiée pour le peuplement d'ontologies.* Si nous souhaitons que les annotations métier soient conformes à un ensemble d'ontologies, notre volonté est également de simplifier leur processus de création, en se basant sur des interfaces ne reposant sur aucun prérequis technique. Ainsi, nous avons fait le choix d'une interface combinant page wiki plein-texte et formulaires destinés à l'annotation. Si l'on peut argumenter que cette restriction (tout comme le choix d'ontologies prédéfinies et figées

³⁷Il s'agit ici d'un choix guidé par une optimisation en termes de performance qui consiste à ne pas recalculer l'ensemble de la base de connaissance à chaque nouveau fait saisi.

³⁸<http://ufowiki.org>

³⁹C'est également une des raisons qui nous a poussé à implémenter UfoWiki plutôt que d'enrichir une plate-forme existante parmi celles présentées auparavant.

pour l'utilisateur final) va à l'encontre des principes de la philosophie wiki, gardons à l'esprit le contexte d'entreprise dans lequel se situe notre approche et l'impératif de qualité sémantique des annotations que nous visons ;

- *Une représentation couplée des annotations socio-structurelles et métier.* Si notre objectif principal avec UfoWiki est le peuplement d'ontologies, nous souhaitons également représenter les différentes métadonnées socio-structurelles s'y rattachant (auteurs, pages, tags ...). Pour ce faire, notre système réutilise des principes définis par ailleurs dans ce chapitre, à savoir la production automatique d'annotations socio-structurelles avec SIOC (Section 4.1, page 138) et la possibilité de *tagging* avancé avec MOAT (Section 4.3, page 170). Afin d'aller plus loin, nous avons étendu le modèle proposé par SIOC pour une modélisation plus fine des liens entre ces deux niveaux de représentation (Section 4.2.3, page 156). Cette extension nous permet de représenter des faits comme "*Le fait qu'EDF est basé en France est issu d'une page wiki créée par Alexandre Passant dans le wiki 'HPédia'*", ce que les wikis sémantiques traditionnels ne sont en général pas en mesure de faire ;
- *Une utilisation immédiate des connaissances produites.* Afin de bénéficier des différentes annotations produites, nous nous sommes ici essentiellement attachés à la mise en place d'un processus de requêtes internes, avec un système inspiré de Semantic MediaWiki. Ces requêtes sont définies par les administrateurs et peuvent être ensuite utilisées au sein de toute page wiki. Elles permettent de plus différents modes de visualisation (Section 5.2.1, page 196), à la manière d'OntoWiki. Concernant l'aide à la recherche et à la navigation, ces fonctions ont été portées au niveau du médiateur (Section 5.4, page 212). De plus, ces macros prennent également en compte des principes d'inférence RDFS reposant sur la subsomption de classes et de propriétés des ontologies utilisées. L'objectif de réutiliser les annotations immédiatement, s'il n'est pas nouveau, permet de motiver les utilisateurs en leur montrant directement l'intérêt d'une telle démarche d'annotation sémantique⁴⁰. C'est également une particularité de notre approche *SemSLATES*, à savoir que l'utilisateur final est à la fois producteur et utilisateur des annotations sémantiques, à la différence par exemple de [Maedche et al., 2003] où la production et l'utilisation d'annotations sont réservées à des communautés distinctes ;
- *Une réutilisation de données externes.* Si certains wikis proposent l'import massif de données RDF (par exemple OntoWiki), notre approche consiste à lier les données du wiki à des données présentes sur le Web Sémantique au moment de l'annotation. Alors que la plupart des wikis, en termes de peuplement d'ontologies, peuvent être vus comme des îlots de données déconnectés puisque définissant leurs propres URIs en vase clos, notre vision permet d'intégrer plus globalement les données produites par UfoWiki au sein du Web Sémantique et réciproquement (Section 4.2.4, page 163). À noter que ces principes permettent également d'envisager UfoWiki comme un producteur de données liées dans cette optique d'un graphe global de données RDF (Section 1.1.4, page 27). D'un point de vue plus pratique, cette intégration est proposée dans un objectif d'enrichissement des fonctionnalités proposées par l'outil, via la mise en place

⁴⁰Ou plutôt d'utilisation du Wiki car le principe d'annotation sémantique leur est entièrement masqué.

de *mash-ups* sémantiques ;

- *Des annotations mutualisées entre les différents wikis.* Enfin, bien qu'UfoWiki repose sur l'idée d'un serveur proposant des wikis indépendants (par rapport aux communautés qui se les approprient, aux sujets abordés ...), les annotations produites sont partagées par l'ensemble d'entre eux au sein d'une base de connaissance commune (Section 4.2.3, page 159). De cette manière, différents wikis peuvent établir des assertions aux sujets des mêmes concepts, dans un objectif d'unification des données produites. Cette mutualisation permet également de réutiliser au sein d'un wiki particulier les informations issues d'un autre wiki, que cela soit pour l'aide à la saisie ou plus généralement pour enrichir les pages à l'aide des macros évoquées précédemment.

4.2.3 Architecture logicielle

Pour arriver à ces différents objectifs, notre système fait intervenir trois composants majeurs : (1) un ensemble d'ontologies, (2) des interfaces d'administration et d'édition et (3) un système de production et de stockage des annotations. Si nous l'avons conçu comme un *plug-in* de l'outil wiki d'origine, nous allons ici le détailler comme un système à part entière, notamment parce que l'outil sur lequel il repose est un outil *ad hoc* développé pour les besoins de la plate-forme Hermès. Gardons malgré tout cette notion d'extension à l'esprit notamment par rapport à notre vision qui propose d'enrichir les outils d'un système d'information existant et déjà pris en main par les utilisateurs plutôt que d'en proposer de nouveaux.

Ontologies

La première partie de l'architecture d'UfoWiki consiste donc en un ensemble d'ontologies venant en support des annotations produites. Puisque nous souhaitons représenter à la fois des annotations socio-structurelles et des annotations métier, deux types d'ontologies sont nécessaires :

- pour la première partie, nous avons naturellement fait le choix de SIOC et de son module Types, pour modéliser la structure d'un wiki et les pages associées avec les classes `sioct:Wiki` et `sioct:WikiArticle`. Le système permettant aussi de taguer les pages, nous reposons sur la *Tag Ontology* et sur MOAT puisqu'UfoWiki intègre des fonctionnalités d'indexation sémantiques à partir de tags (Section 4.3, page 170) ;
- pour la seconde, le wiki reste indépendant des ontologies utilisées, le seul prérequis étant leur modélisation en RDFS/OWL. Dans le cas d'usage qui nous intéresse, nous avons considéré les modèles du chapitre précédent (Section 3.2, page 104).

Toujours en termes d'ontologies, nous avons évoqué dans la section précédente un point qui nous paraît particulièrement novateur, à savoir la modélisation des liens entre annotations socio-structurelles et annotations métier. Pour ce faire, nous avons introduit une propriété `embedsKnowledge`, qui permet de faire le lien entre ces deux ensembles d'assertions. Celle-ci repose sur l'utilisation des graphes nommés [Carroll *et al.*, 2005] et propose ainsi une autre manière d'articuler métadonnées socio-structurelles et données métier en plus de MOAT comme nous l'avons vu en conclusion du précédent chapitre (Section 3.3.4, page 135). Cette propriété permet de lier toute instance de `sioct:Item` à un graphe RDF d'annotations métier (Figure 4.8, page 157). En pratique, nous disposons lors de la création d'une

page sous UfoWiki de deux graphes d'annotations, regroupés au sein de deux documents distincts. Nous lions ainsi l'instance de `sioc:Item` à l'URL du document contenant les annotations métier.

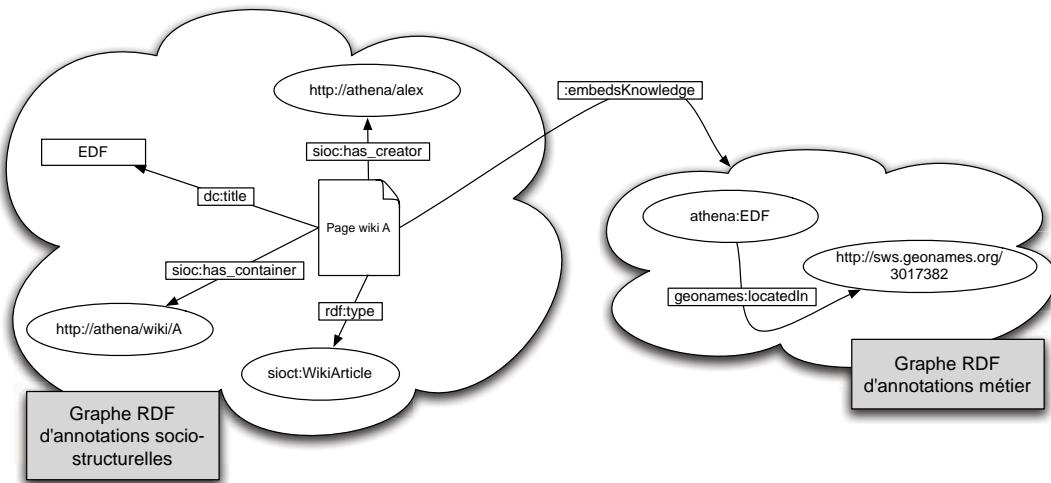


Figure 4.8: Interactions entre annotations documentaires et annotations métier dans UfoWiki

C'est grâce à cette relation `embedsKnowledge` que nous pouvons modéliser finement certaines propriétés associées aux annotations métier. On peut ainsi la considérer comme un moyen de réifier des assertions métier via l'utilisation des annotations socio-structurelles. Comme le montre la figure précédente, cette complémentarité entre les deux graphes d'annotations nous permet d'identifier par exemple qui, quand et depuis quel wiki le fait qu'EDF est basé en France a été établi. Nous verrons dans le chapitre suivant comment nous tirons bénéfice de cette propriété au sein de requêtes SPARQL (Section 5.2.2, page 201). Ce choix de séparer les annotations dans deux documents distincts nous permet également d'envisager un export et une utilisation des annotations selon différents points de vue : annotations socio-structurelles uniquement ou annotations métier, les deux pouvant bien sûr être combinées.

Interfaces

Il convient ici de distinguer (1) les interfaces d'administration dédiées à la gestion des wikis et des ontologies et (2) les interfaces utilisateur dédiées à la visualisation et l'édition de pages. Nous ne détaillerons ici que le premier type et présenterons le second dans la section suivante (Section 4.2.4, page 160).

Pour chaque wiki, une interface d'administration permet de définir, sous forme de *widgets*, les différents patrons d'annotations disponibles. Rappelons que notre outil se base sur une approche d'annotations assistées par formulaires afin de guider l'utilisateur et de s'assurer de leur cohérence avec les ontologies utilisées. Les administrateurs d'un wiki ont ainsi la possibilité de définir :

- des types de pages, chacun associé à une unique classe des ontologies métier utilisées. Il est ainsi possible de définir une page *Personne* et une page *Entreprise*, associées respectivement aux classes `foaf:Person` et `foafplus:Company` (Figure 4.9, page 158). Les classes associées sont définies sous forme préfixées, les préfixes étant alignés par l'administrateur général de la plate-forme avec des modèles existants. Comme nous le verrons, l'utilisateur a ensuite la possibilité de choisir un des types proposés lors de la création d'une nouvelle page, induisant la création d'une instance de la classe correspondante ;

The screenshot shows a user interface for creating a new content type. At the top, there is a navigation bar with the word "Edition". Below it, a sidebar contains two items: "Propriétés" and "Champs", with "Propriétés" being highlighted. The main area has several input fields:

- Content type name:** *
Hpédia - Société
- Description:**
Fiche société pour Hpédia
A one-line description of the content type.
- Associated class URI:**
foafplus:Company
URI of the related ontology class, eg: <http://xmlns.com/foaf/0.1/Person>

Figure 4.9: Association d'un type de page à une classe avec UfoWiki

- des éléments de formulaires qui sont associés aux types de pages précédents via une interface AJAX de glisser-déposer (Figure 4.10, page 159). Ces éléments peuvent être de différents types (zone de texte, case à cocher ...) et permettent la production d'annotations RDF associées aux instances créées via le wiki. Une syntaxe particulière est utilisée par l'administrateur pour définir le lien entre formulaires et annotations, sous la forme `$idA propriété $idB`. Ainsi, `$1 foaf:member $main` permettra d'établir une relation `foaf:member` entre le concept identifié par le premier champ de formulaire (`$1`) et celui identifié par la page en cours (`$main`). Pour faire la distinction entre les propriétés `ObjectProperty` et `DatatypeProperty`, ces *widgets* nécessitent également d'indiquer le type d'objet attendu pour chaque élément dans les cas d'une propriété `ObjectProperty`. Ce type est ensuite utilisé pour l'autocomplétion ou pour la création de nouvelles instances si nécessaire (Section 4.2.4, page 160). D'autre part, ces *widgets* peuvent être mutualisés au sein de plusieurs pages, *i.e.* associés à plusieurs classes. C'est par exemple le cas d'un *widget Localisation* qui peut-être associé à la fois à *Personne* et *Entreprise*.

À la lecture de ce second point, on peut se demander pourquoi cette génération de formulaires n'est pas automatique à partir du moment où chaque page est associée à une classe et où l'on dispose des modèles associées. Cette automatisation est certes possible (en analysant l'ontologie utilisée) mais conduit selon nous à des formulaires beaucoup moins pertinents, en raison de la nature même des ontologies RDFS/OWL et notamment de la modélisation du domaine (au sens `rdfs:domain`) des propriétés. En effet, si l'on souhaite auto-

Figure 4.10: Création de formulaire pour une classe donnée avec UfoWiki

matiser la création de formulaires, il est nécessaire de prendre en compte non seulement les propriétés ayant un domaine correspondant exactement à chaque classe, mais aussi celles ayant un domaine compatible⁴¹. Si cela impose d'une part l'utilisation d'un raisonneur pour identifier ces propriétés, cela peut aussi conduire à une abondance de champs non pertinents. Ainsi puisque nous utilisons FOAF, cette automatisation aurait intégré au formulaire *Personne* un champ *Code ADN* (`foaf:checksum`, domaine non restreint) et au formulaire *Entreprise* un champ *Compte MSN* (`foaf:msnChatID`, domaine défini par `foaf:Agent` dont notre classe `foafplus:Company` hérite). Si nos principes de formulaires explicites – et de la même manière le fait de spécifier le type attendu pour les valeurs de chaque propriété – ferment en quelque sorte l'hypothèse du monde ouvert, cela nous semble indispensable pour proposer des interfaces utilisateur pertinentes⁴².

Production et stockage d'annotations

Enfin, la dernière partie de l'architecture d'UfoWiki est relative à la production et au stockage des annotations RDF. Comme nous l'avons indiqué précédemment, à chaque page wiki sont associés deux graphes d'annotations, distincts mais pour autant interconnectés via une relation `embedsKnowledge` (associée à un lien `rdfs:seeAlso`). Les annotations relatives aux ontologies de domaine sont produites en fonction des formulaires saisis par l'utilisateur, alors que les annotations socio-structurelles sont produites automatiquement de la même manière que nous l'avons vu précédemment (Section 4.1, page 138).

Le stockage de celles-ci se fait de manière unifiée et en temps réel au sein d'un entrepôt de données. Si celui-ci est commun à tous les wikis de la plate-forme, nous l'avons également mutualisé avec les autres outils (Section 5.1, page 186). C'est grâce à cet entrepôt global

⁴¹Nous nous référons ici à la notion de compatibilité des domaines telle que nous l'avons évoqué plus tôt dans ce mémoire (Section 3.2.5, page 117).

⁴²Nous n'avons pas considéré ici la possibilité d'utiliser des ontologies dédiées à la présentation et mise en forme de contenus qui pourraient répondre en partie à la question, comme proposées par [Khushraj et Lassila, 2005].

qui agrège l'ensemble des connaissances produites par les différents wikis que le système d'autocomplétion peut être mis en place, tout comme les différentes possibilités de requêtes offertes par UfoWiki (Section 5.2, page 196). La figure qui suit (Figure 4.11, page 160) représente en outre le système de stockage pour un wiki particulier, ici exemplifié au travers de deux pages wiki, soit quatre graphes d'annotations distincts mais interconnectés. Cette figure met également en avant la possibilité d'établir des annotations au sujet d'une même ressource à l'aide de différentes pages wiki .

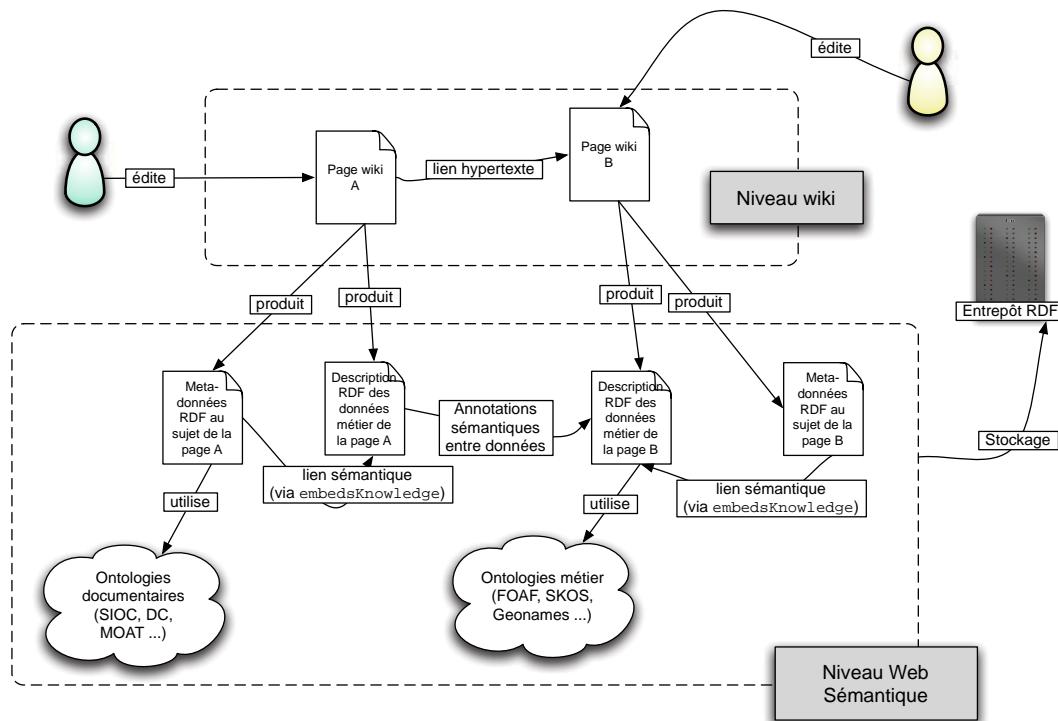


Figure 4.11: Architecture d'un wiki au sein d'UfoWiki

4.2.4 Utilisation d'UfoWiki et peuplement collaboratif d'ontologies

UfoWiki en pratique

Parmi l'ensemble des wikis mis en place au sein d'Hermès, trois d'entre eux ont été enrichis des fonctionnalités de peuplement d'ontologies proposées par UfoWiki⁴³ :

- un wiki destiné à la modélisation des acteurs et de leurs domaines d'activités, nommé HPédia ;
- un wiki destiné à la modélisation et l'organisation taxonomique des différents domaines et métiers ;
- un wiki destiné à la gestion des partenariats.

⁴³Les autres wikis bénéficient cependant des autres caractéristiques d'UfoWiki, notamment les macros.

Pour chacun d'entre eux, différents types de pages et de formulaires ont été créés, associés aux modèles présentés dans le précédent chapitre (Section 3.2, page 104). Ainsi, à chaque création de page depuis l'un de ces wikis, l'utilisateur a la possibilité de choisir le type de page correspondant parmi ceux disponibles pour le wiki en question, chaque type étant associé à une classe particulière. Par exemple, dans HPédia, l'utilisateur peut choisir parmi différents types dont Personnalité (associé à foaf :Person) ou Société (foafplus :Company), les URIs étant masquées à l'utilisateur (Figure 4.12, page 161). Il a également la possibilité de ne pas utiliser de formulaire : dans ce cas, seules les annotations socio-structurelles seront produites.

Vous pouvez choisir de modifier cette page en utilisant un formulaire.

Classique	Un document classique, sans structure (i.e. le type par défaut)
Fiche Hpédia	Modèle de fiche pour Hpédia
Hpédia - Personnalité	Fiche personnalité pour Hpédia
Hpédia - Société	Fiche société pour Hpédia
Hpédia - Organisme de Recherche	Fiche organisme de recherche pour Hpédia
Hpédia - Institution	Fiche institution pour Hpédia
Hpédia - Association	Fiche association pour Hpédia
Hpédia - Autres	Fiche autre pour Hpédia

Vous avez sélectionné le formulaire "Hpédia - Société" (Fiche société pour Hpédia):

Titre: *
nouvelle page

Figure 4.12: Sélection d'un type de contenu avec UfoWiki

Dans le cas où un type particulier de page est sélectionné, l'utilisateur se voit alors proposer une page d'édition composée de :

- un champ d'édition classique (*i.e.* une zone de texte libre), identique à celui proposé dans l'outil initial ;
- un ensemble d'éléments de formulaires correspondants aux *widgets* définis par l'administrateur du wiki pour le type de page concerné.

La figure suivante représente ainsi l'interface d'édition associée au type de page Association au sein du wiki HPédia (Figure 4.13, page 162). On y distingue la zone principale et différents *widgets* (*Localisation*, *Rattachement*, *Domaine et Métier*), dont un premier qui met en valeur les possibilités d'autocomplétion offertes par l'outil. Cette autocomplétion est rendue possible via l'utilisation d'une requête SPARQL en temps réel sur l'ensemble des annotations des différents wikis. Elle prend en compte les caractères saisis par l'utilisateur et la classe associée à ce *widget* afin de déterminer les instances ayant un label (`rdfs:label`) correspondant à la saisie, tout en gérant l'inférence pour proposer également les éventuelles instances des sous-classes associées.

Avant de revenir plus tard sur la macro présentée dans cet exemple (Section 5.2.1, page 196), détaillons tout d'abord ce que nous appelons *instance interne*. Dans la plupart des wi-

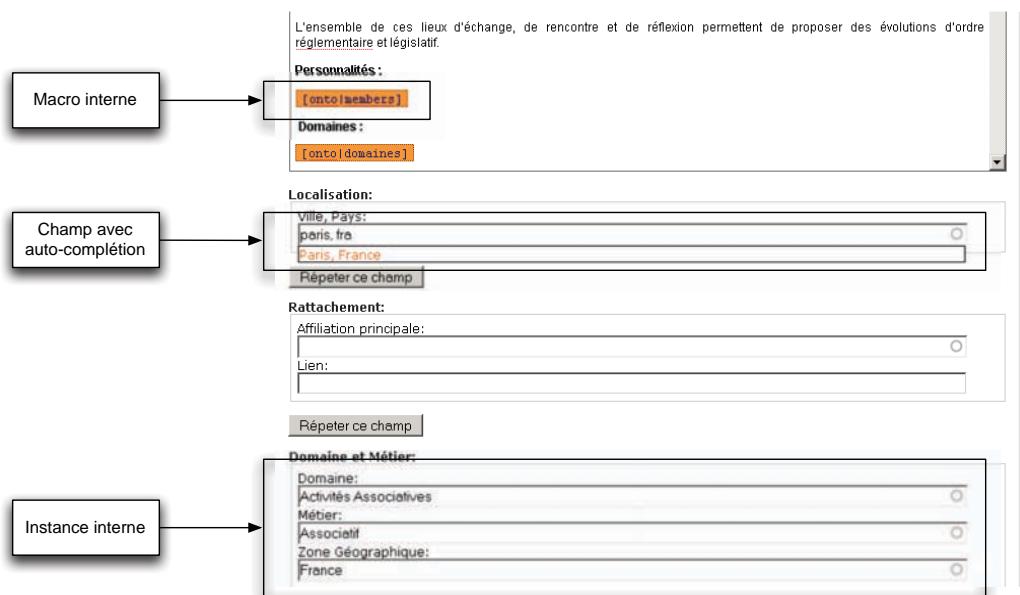


Figure 4.13: Édition d'une page wiki pour la création d'instance via UfoWiki

kis sémantiques, les relations entre instances sont possibles dans la mesure où chaque instance est identifiée par une page donnée. Ce choix s'explique notamment (1) par le lien implicite entre chaque page et une instance associée et (2) par l'utilisation des liens hypertextes pour modéliser les relations entre ces instances. Ceci oblige cependant à disposer d'une page wiki par instance. Si l'on se remémore nos modèles, nous disposons d'une classe `role:Role` qu'il nous semble peu pertinent de représenter de cette manière, notamment car il s'agit d'une simple relation tripartite entre un domaine, un métier et une zone géographique, sans description particulière (Section 3.2.4, page 109). UfoWiki offre ainsi, via un type de *widget* particulier, la possibilité de créer des instances au sein de pages en plus de l'instance principale, comme l'illustre la figure précédente avec cette notion d'instance interne (Figure 4.13, page 162). Notons également que, pour chaque champ dont la valeur est associée à une classe particulière, une nouvelle instance est créée (et typée selon cette classe) si l'en n'existe pas encore au sein de la base de connaissance pour la valeur entrée. Enfin, afin d'associer à chaque page wiki l'instance principale correspondante nous utilisons la propriété `foaf:primaryTopic` au sein du graphe d'annotations socio-structurelles, alors que `sioc:topic` est utilisé pour identifier les autres instances associées à la page en question. Nous verrons dans le chapitre suivant comment cette propriété nous permet de contextualiser les macros au sein d'UfoWiki (Section 5.2.2, page 201). Notons également que pour chaque nouvelle instance créée, UfoWiki va considérer l'URI de cette instance comme signification globale du tag correspondant à son label, et intégrer cette signification au sein du serveur MOAT (Section 4.3, page 170), afin de faciliter le processus d'indexation sémantique à partir de tags.

L'ensemble des annotations RDF produites depuis cet exemple de page wiki, associée à l'organisation *Association des Maires de France*, est disponible en annexe :

- d'une part les annotations socio-structurelles (Section E, page 239) ;
- d'autre part les annotations métier (Section D, page 235).

Comme on peut le voir en analysant ce second document, des URIs particulières sont utilisées pour modéliser les domaines et métiers. En effet, comme nous l'avons évoqué, UfoWiki permet le partage d'annotations produites entre les différents wikis du système, en particulier le partage d'URIs associées aux différentes instances produites. Ainsi, les instances créées au sein du wiki mis en place pour l'organisation des domaines et métiers (et reposant également sur UfoWiki) sont réutilisées lors de la création d'annotations au sein d'HPédia, réutilisation facilité par le système d'autocomplétion. Notons que ce second wiki (relatif aux domaines et métiers) bénéficie également, tout comme HPédia, de possibilités de complétion qui permettent ici d'assister l'utilisateur dans la définition des taxonomies de domaines et métiers (Figure 4.14, page 163). L'utilisation d'un tel wiki permet ainsi une évolution constante de ces taxonomies afin de s'adapter rapidement à l'émergence de nouveaux domaines.

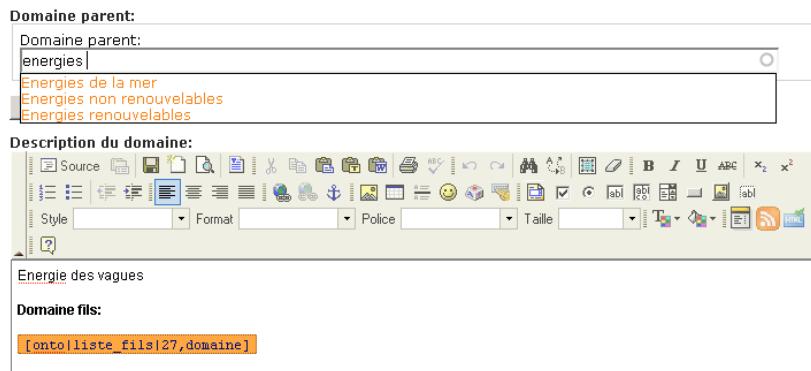


Figure 4.14: Gestion d'une taxonomie de domaines avec UfoWiki

Intégration de ressources externes

Nous avons également évoqué dans les objectifs d'UfoWiki la possibilité d'intégrer des liens vers des concepts déjà existants sur le Web Sémantique et de réutiliser ceux-ci au sein de notre système. Alors que la plupart des wikis sémantiques définissent de nouvelles URIs pour chaque concept qu'ils instantient, il nous semble intéressant de réutiliser dans la mesure du possible des URIs déjà définies pour identifier ces concepts. Ce processus de réutilisation et d'intégration nous paraît pertinent dès lors que l'on veut parvenir à un Web Sémantique où les données sont interconnectées globalement et non pas seulement au sein d'écosystèmes clos.

Supposons en effet que dix wikis sémantiques différents définissent une page au sujet d'une entreprise établie en France. Pour chaque wiki, la relation va être établie en utilisant une URI locale du type `http://mon-wiki.org/resource/France` pour définir l'identi-

fiant associé à la France. Il y aura donc en conséquence, pour un même concept, autant d'URIs qu'il existe de wikis, ce qui est assez paradoxal dans une optique de sémantique commune entre applications. Cette abondance d'URIs conduira en effet à des problèmes d'hétérogénéité sémantique similaires à ce que l'on rencontre par exemple pour les tags (Section 2.2.3, page 63). À l'opposé, utiliser une URI existante pour établir ce lien ou indiquer que le concept instancié est identique à tel autre permet de résoudre ce problème d'hétérogénéité. Par exemple, utiliser dans chaque cas l'URI <http://dbpedia.org/resource/France> permettra d'interconnecter ces wikis entre eux via cette URI unique et référante, partagée entre applications. On établit ainsi de cette manière des passerelles entre des wikis initialement conçus comme des îlots de données indépendants.

En termes d'impacts plus immédiats, ceci offre la possibilité d'enrichir un wiki de connaissances déjà disponibles sur le Web. Dans l'exemple précédent, on peut bénéficier des assertions RDF qui sont rattachées à <http://dbpedia.org/resource/France> pour identifier au sein du wiki quelles sont les entreprises basées en Europe sans que l'on ait eu besoin de déclarer que la France en fait partie, cette relation étant déjà présente dans la description DBpedia associée à l'URI en question.

Pour en revenir à UfoWiki, nous avons donc mis en place un système permettant de réutiliser certaines bases de connaissances externes en son sein. Toujours dans l'optique de ne pas confronter les utilisateurs à ces notions d'URIs, nous avons défini des *widgets* particuliers qui permettent cette intégration de manière simple. Le widget *Localisation* a ainsi été mis en place de manière à interroger automatiquement le service Web Geonames⁴⁴ pour identifier l'URI correspondant à chaque localisation saisie. Il est cependant nécessaire pour l'utilisateur d'entrer explicitement une localisation précise pour éviter les problèmes d'ambiguïté associés à celle-ci (e.g. Paris, France plutôt que simplement Paris), mais cette légère contrainte nous a paru plus simple que de demander à l'utilisateur de lever l'éventuelle ambiguïté lui-même. Notons que cette restriction est due au fait que nous interrogeons le service au moment de la validation et que nous ne possédons pas au sein de notre base de connaissance de l'ensemble des données proposées par Geonames. Dans ce cas, nous aurions pu définir un système d'autocomplétion adaptée afin de résoudre plus simplement ce problème d'ambiguïté, comme le proposent [Hildebrand *et al.*, 2007].

Une fois la ressource identifiée, nous intégrons dans notre médiateur les assertions relatives à celle-ci, ce qui nous permet de :

- bénéficier par la suite du système d'autocomplétion associé à ce *widget*, qui va donc effectuer une requête locale pour identifier les instances de la classe geonames:Feature correspondant à la saisie de l'utilisateur. C'est en ce sens que l'on a pu voir sur l'exemple précédent la suggestion de Paris, France pour la saisie de Paris, Fra (Figure 4.13, page 162);
- profiter de nouveaux services de *mash-ups* sémantiques, permettant notamment de visualiser différentes entreprises d'un secteur d'activité sur une carte Google (Section 5, page 185).

Plus généralement, cette possibilité pour des écosystèmes d'information sémantique d'Entreprise 2.0 de réutiliser des données RDF publiques nous semble avoir un intérêt majeur

⁴⁴<http://www.geonames.org/export/>

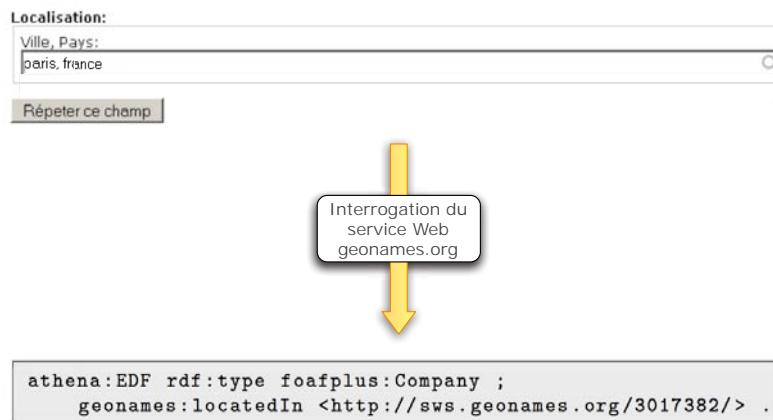


Figure 4.15: Production d'annotations basées sur Geonames avec UfoWiki

pour l'avenir de tels systèmes. En effet, ce processus permet de bénéficier d'un large volume de données publiques, issues notamment de l'initiative *Linking Open Data* (Section 1.1.4, page 27), pour augmenter à moindre coût les capacités de systèmes d'information existants. À la manière des flux RSS qui permettent à un entreprise de bénéficier des connaissances de différents experts sans pour autant être en contact direct avec eux et sans démarche proactive, l'intégration de données publiques permet de bénéficier de connaissances formalisées et réutilisables immédiatement, puisqu'interprétables sans ambiguïté. De plus, en supposant que certaines données d'entreprise soient à terme publiées sur le Web, on bénéficie déjà au sein de ces données de relations vers des ressources existantes, permettant d'amplifier la découverte des informations produites par l'entreprise à partir d'autres sources de données. Au sujet de cette intégration de données externes, on peut également citer des travaux semblables au sein du projet Comma [Cao *et al.*, 2003]. La différence majeure se situe ici selon nous dans la phase d'acquisition et d'intégration de données. Alors que l'approche proposée nécessite différentes méthodes de conversion de données présentes sur le Web en RDF, nous bénéficions dans notre contexte de données déjà disponibles selon ce format mais surtout compatibles avec nos modèles internes, puisque nous avons fait le choix de nous baser sur des modèles abondamment utilisés sur le Web (Section 3.2, page 104). Ceci nous semble ainsi être un point particulièrement pertinent en termes d'adoption des technologies du Web Sémantique en entreprise, et notamment de valeur ajoutée pour l'initiative *Linking Open Data*.

Enfin, pour accentuer cette réutilisation de données publiques, nous avons également mis en place un prototype de *widget* permettant de lier nos instances à celles définies par DBpedia. Toujours dans cette optique d'interfaces de publication simples, l'utilisateur n'est pas confronté à la gestion de l'URI DBpedia, mais utilise simplement un champ qui lui permet d'indiquer la page Wikipedia correspondante, l'URI DBpedia étant identifiée à partir de celle-ci. Cette connexion nous permet à nouveau d'envisager un enrichissement des outils existants, par exemple en affichant l'extrait de l'article Wikipedia associé au concept identifié sur une page donnée.

Comme nous l'avons évoqué en amont, bien qu'UfoWiki ne soit pas un outil disponible publiquement, nous espérons que d'autres wikis viendront à terme bénéficier de cette vision pour être partie intégrante du *Web of Data*.

4.2.5 Evaluation de l'outil et statistiques d'utilisation

Position par rapport à l'état de l'art

Afin d'évaluer notre prototype, comparons tout d'abord ses caractéristiques avec les outils mentionnés auparavant. Notons qu'il s'agit ici d'une comparaison en termes de fonctionnalités, n'ayant pas pu faire d'étude orientée utilisateur afin d'évaluer les avantages et inconvénients des différents outils, aussi bien en termes d'adoption des interfaces de production d'annotations que de qualité sémantique de celles-ci. Ce tableau identifie donc différents aspects qui nous semblent important en termes de wikis sémantiques pour le peuplement d'ontologies. Nous reviendrons plus particulièrement sur les services additionnels offerts par UfoWiki, notamment les possibilités de macros internes et de navigation avancée, dans le chapitre suivant (Section 5.2, page 196).

En termes de production d'annotations, il nous semble important de signaler qu'il n'y a pas selon nous de wiki sémantique idéal. Selon le contexte, la communauté d'utilisateurs et l'usage que l'on souhaite faire des annotations, les différentes méthodes proposées (et les fonctionnalités associées par les outils qui les implémentent) nous semblent toutes avoir des avantages et des inconvénients. Par exemple, ACE se révèle intéressant pour des cas d'utilisation où la consistance des données et les capacités d'inférence passent en premier plan, alors que Semantic MediaWiki peut-être plus pertinent dans un contexte plus souple où l'on souhaite avoir un modèle ouvert et évolutif, en espérant une autorégulation de celui-ci via la communauté. La solution proposée par UfoWiki nous semble être un bon compromis dans un contexte organisationnel où l'on souhaite s'assurer des annotations produites sans pour autant confronter les utilisateurs aux principes de modélisation RDF(S)/OWL. Vis-à-vis des autres caractéristiques d'UfoWiki, ses plus-values par rapport à l'existant nous semblent être :

- les principes d'annotations par formulaire, permettant une représentation simple et assistée d'annotations sémantiques métier ;
- la production simultanée d'annotations socio-structurelles et d'annotations métier, les deux étant de plus combinées. Parmis les wikis considérés, seul IkeWiki offre un modèle complet et pertinent pour ce premier type d'annotations, le modèle SWIVT⁴⁵ de Semantic MediaWiki étant relativement pauvre (seule la notion de page wiki est modélisée) ;
- la complémentarité avec des ressources déjà présentes sur le Web Sémantique, dans un objectif de wikis interconnectés et non plus considérés comme des outils indépendants définissant leurs propres instances en vase clos.

⁴⁵<http://semantic-mediawiki.org/swivt/1.0#>

	Métdonnées socio-structurelles	Peuplement d'ontologies	Services Additionnels
	Ontologies	Annotations	
Platypus	Définies par l'administrateur Annotations RDF / XML	Formulaire triplets	Aide à la navigation
SemPerWiki		Annotations RDF / XML Annotations Turtle	Aide à la navigation Macros internes
Semantic MediaWiki	SWIWT	Générées via le wiki Évolution libre	Aide à la navigation Macros internes
IkeWiki	Modèle IkeWiki SIOC (discussions)	Définies par l'administrateur Prédefinies par l'administrateur Évolution (assistée) via le wiki	Assistance (liens typés) Raisonnement RDFS
Ontowiki		Formulaires	Aide à la navigation Visualisation avancée
AceWiki		Générées via le wiki Évolution (contrôlée) via le wiki	Contraintes (ACE) Raisonnement OWL
UfoWiki	SIOC MOAT (<i>Tagging</i>)	Définies par l'administrateur	Formulaires Raisonnement RDFS Point d'accès SPARQL Visualisation avancée

Tableau 4.1: Positionnement d'UfoWiki par rapport à d'autres wikis sémantiques

Statistiques d'utilisation

Comme nous l'avons précédemment évoqué, trois wikis utilisant UfoWiki ont été mis en place au sein d'Hermès. Afin de mesurer l'acceptation de l'outil, nous avons étudié sur une période de plusieurs mois l'utilisation de l'un d'entre eux, à savoir HPédia, wiki destiné à capitaliser des informations au sujet de différents acteurs académiques et industriels. Celui-ci permet donc le peuplement des ontologies de domaine présentées dans le chapitre précédent, dans le sens où chaque acteur est représenté par des informations générales le concernant (type d'entité, nom, etc.), sa localisation et les différents rôles qui lui sont associés, comme nous avons pu le voir sur une précédente figure (Figure 4.13, page 162). Sur une période de 200 jours sur laquelle porte notre analyse, on peut observer que 173 pages wiki ont été créées, pour un total de 352 instances (Figure 4.16, page 168). Une vingtaine d'utilisateurs ont pris part à cette démarche volontaire de peuplement d'ontologies à travers l'utilisation d'HPédia. Il est important de signaler que la majorité de ceux-ci n'étaient ni formés sur les technologies du Web Sémantique ni particulièrement adeptes d'interfaces logicielles avancées, certains d'entre eux n'ayant jamais utilisé de wiki avant la mise en place de la plate-forme Hermès. Il nous semble en conséquence qu'UfoWiki a pu jouer correctement son rôle d'outil simple dédié au peuplement d'ontologies, d'autant plus que les utilisateurs ayant participé à ce peuplement n'ont jamais été directement confrontés à cette notion d'ontologies et de bases de connaissances au travers de l'outil. On peut également remarquer sur cette figure un pic aux alentours du 75ème jour, correspondant à une période où certaines données présentes dans d'autres bases de connaissances internes à l'entreprise (Lotus Notes) ont été portées au sein d'HPédia.

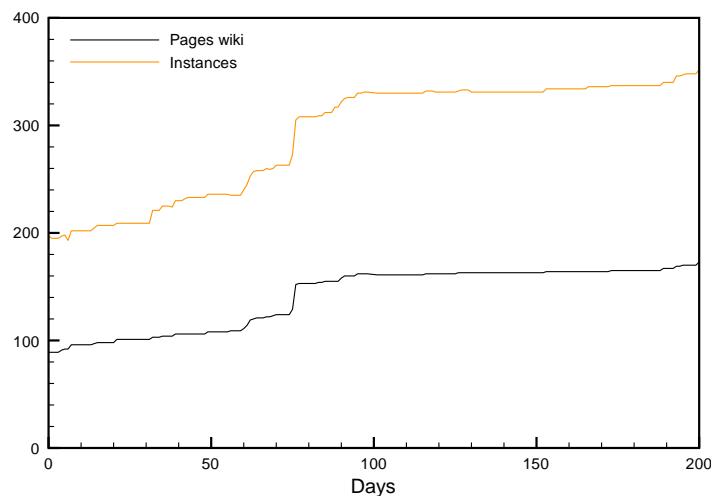


Figure 4.16: Statistiques d'utilisation d'UfoWiki : Pages et instances

Une seconde figure significative du nombre de triplets produits (nous ne considérons ici que les annotations métier) nous montre également qu'une moyenne de 9 triplets RDF ont été produits pour chaque page (Figure 4.17, page 169). Ce nombre est en fait assez

inégal, puisque l'on compte par exemple 29 triplets dans le graphe RDF correspondant à l'*Association des Maires de France* (Annexe D, page 235). Les formulaires sont donc remplis assez différemment selon les acteurs étudiés, certaines pages étant relativement complètes (avec des informations assez poussées sur les différents rôles, les membres associés, etc.) et d'autres plus légères (avec par exemple uniquement la localisation et le domaine d'activité, soit quatre triplets seulement dans ce cas, *i.e.* les deux précédents plus la description de l'acteur en question et son type). Quoi qu'il en soit, cette moyenne nous semble relativement acceptable vu le nombre d'utilisateurs. Si l'on prend par exemple la page Wikipedia consacrée à cette même association, on ne trouve dans son infobox⁴⁶ que 13 assertions, pour 37 personnes ayant participé à son élaboration et 69 éditions⁴⁷.

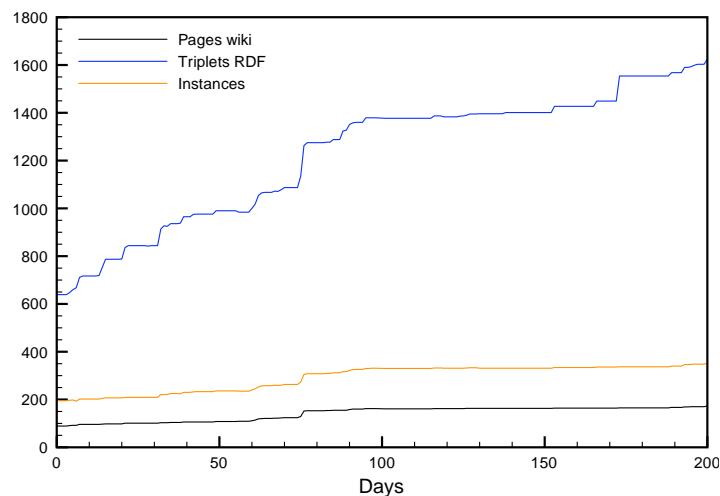


Figure 4.17: Statistiques d'utilisation d'UfoWiki : Pages, instances et triplets

Enfin, malheureusement, nous n'avons pas pu étudier l'évolution de chaque instance et la manière dont la collaboration permettait de faire évaluer celles-ci, le suivi de versions des annotations n'étant pas assuré par UfoWiki, et n'ayant pas conservé les versions précédentes des graphes d'annotations associés à chaque page. Nous tirons cependant bénéfice de cet aspect collaboratif d'édition dans la mesure où, comme nous l'avons dit, les instances sont partagées entre pages d'un même wiki (et de manière plus large entre différents wikis). Ainsi, un utilisateur créant une instance permet à d'autres utilisateurs de définir de nouvelles annotations utilisant cette même instance, par exemple pour définir une personnalité entant que membre d'une organisation donnée. Les principes Web 2.0 d'architecture participative sont donc ici directement appliqués à la définition et l'expansion de graphes d'annotations RDF.

⁴⁶Partie d'une page Wikipedia contenant des informations structurées, qui sert notamment au maintien de DBpedia.

⁴⁷http://vs.aka-online.de/cgi-bin/wppagehiststat.pl?lang=fr.wikipedia&page=Association_des_maires_de_France

4.3 DU TAGGING À L'INDEXATION SÉMANTIQUE

4.3.1 Processus d'indexation sémantique associé à MOAT

Si la production d'annotations socio-structurelles, telle que définie au début de ce chapitre, peut être automatisée, passer d'un processus classique de *tagging* à une indexation sémantique guidée par des ontologies de domaine est plus complexe. Nous avons présenté dans le chapitre précédent MOAT, modèle permettant de lier tags et ontologies de domaine dans cet objectif de transition entre tags et indexation sémantique (Section 3.3, page 120). Celui-ci repose sur la notion de significations associées aux tags, celles-ci étant représentées avec des URIs de concepts du Web Sémantique, en particulier des instances d'ontologies. Afin de mettre ce modèle en pratique et l'intégrer à des systèmes de *tagging*, il est primordial de répondre aux deux questions suivantes :

- tout d'abord, comment rendre ce passage aussi souple que possible pour l'utilisateur final. La simplicité des tags ayant contribué à leur acceptation, il est nécessaire de conserver une approche intuitive pour permettre la réussite de tels systèmes ;
- ensuite, comment mettre en place une architecture de participation au sein de ce processus. Une telle architecture doit permettre le partage des significations au sein d'une communauté, de la même manière que les plates-formes classiques de systèmes à base de tags permettent à tous de bénéficier des apports de chacun (autocomplétion, suggestion ...).

Pour ce faire, nous avons mis en place une architecture logicielle reposant sur le modèle MOAT et basée sur :

- un serveur qui va stocker l'ensemble des tags utilisés au sein d'une communauté donnée ainsi que les significations globales associées à ceux-ci, *i.e.* les URIs des concepts signifiants ;
- différents clients qui vont permettre aux utilisateurs de bénéficier de ces significations lors d'actions de *tagging* pour définir les significations locales de leurs tags. Ces clients interagissent avec le serveur pour permettre l'ajout de nouvelles significations globales au sein de la communauté.

Le processus associé à cette architecture permet ainsi de faire le lien entre *tagging* et indexation sémantique de la manière suivante (Figure 4.18, page 171) :

- l'utilisateur crée un contenu et le tague avec de simples mots-clés ;
- pour chaque tag, le client MOAT va récupérer depuis le serveur auquel l'utilisateur a souscrit la liste, qui peut ne contenir qu'un élément, des significations globales associées à ce tag (*i.e.* les URIs des différents concepts associées) ;
- l'utilisateur va choisir parmi cette liste le concept correspondant à son tag dans ce contexte particulier d'annotation. Si rien ne convient, il a la possibilité de définir une nouvelle signification ;
- une fois le choix validé, le client produit automatiquement l'ensemble des annotations RDF relatives à l'indexation sémantique du contenu annoté.

Ce processus permet donc, via un *workflow* assez léger, de passer du document tagué à un graphe d'annotations RDF relatives à l'indexation sémantique de celui-ci. Nous verrons dans le chapitre suivant comment tirer profit des différentes annotations ainsi créées en

termes de requêtes et de recherche de documents (Section 5.4, page 212).

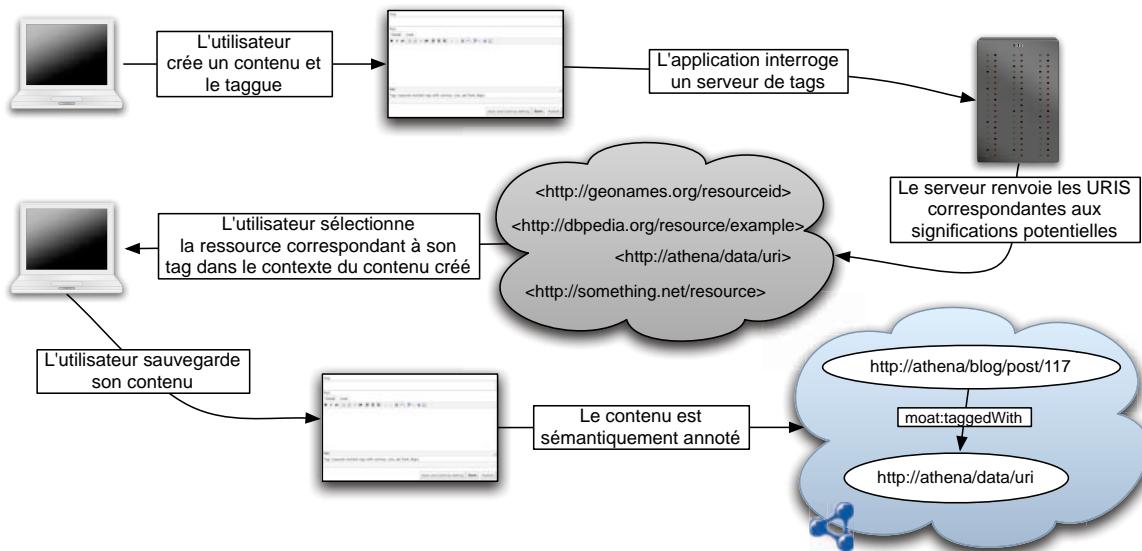


Figure 4.18: Framework utilisateur MOAT

Avant de rentrer dans les détails techniques des outils mis en place pour satisfaire ce processus (Section 4.3.2, page 174), prenons un exemple concret de l'approche. Nous avons volontairement contextualisé celui-ci avec l'utilisation que nous en faisons au sein de notre système, en utilisant ici différentes instances présentes dans notre base de connaissances. Un *workflow* d'utilisation, également identifié par la figure qui suit (Figure 4.19, page 173), peut donc être le suivant :

- un utilisateur va taguer un billet au sujet d'un nouveau type de pompe à chaleur avec le mot-clé pac ;
- le client interroge le serveur associé pour connaître les significations globales associées à ce tag. Celui-ci lui renvoie une liste de deux éléments qui contient les URIs `athena:PolitiqueAgricoleCommune` et `athena:PileACombustible}`;
- ces deux significations ne correspondant pas, l'utilisateur en ajoute une nouvelle⁴⁸ : `athena:PompeAChaleur` ;
- le choix est validé et le contenu est alors annoté et associé à l'URI choisie (selon les relations définies au chapitre précédent (Section 3.3, page 120)). De plus, la nouvelle signification relative au tag pac est stockée au sein du serveur ;
- un second utilisateur rédige plus tard un billet sur un thème similaire et lui associe le même mot-clé ;

⁴⁸Nous détaillerons sous peu de quelle manière se fait cette assignation d'une nouvelle signification pour un tag donné.

- le serveur va alors renvoyer trois URIs, l'utilisateur n'a donc qu'à en valider une pour que son contenu soit correctement indexé et annoté, de la même manière que précédemment. C'est donc cette étape qui permet de gérer l'ambiguïté. Dans le cas d'un billet annoté avec le même tag mais relatif aux piles à combustible, une autre URI aurait été sélectionnée par l'utilisateur ;
- un troisième utilisateur va ensuite annoter un billet avec le mot-clé pompe-a-chaleur ;
- le serveur ne renvoie ici aucune URI, aucune association n'ayant été définie jusque là au sein du serveur de la communauté concernée pour ce tag ;
- l'utilisateur fait donc le choix d'une nouvelle URI pour la signification associée à ce tag, à savoir athena:PompeAChaleur, celle-ci étant ensuite intégrée au serveur, alors que le contenu est par ailleurs annoté après validation.

Ce cas d'utilisation et les annotations associées mettent en avant deux principes qui sont à la base de MOAT : (1) la gestion de l'ambiguïté des tags, puisque l'on a deux documents associés au même tag (pac) mais liés à deux URIs distinctes (athena:PompeAChaleur et athena:PileACombustible) et (2) la gestion de leur hétérogénéité, puisque nous avons ici deux tags distincts (pac et pompe-a-chaleur) qui rattachés localement à la même URI (athena:PompeAChaleur) permettent au final d'avoir deux contenus indexés avec le même concept. En ce qui concerne l'autre problème classique des tags, *i.e.* l'absence de relations, nous gérons celui-ci en considérant les relations au niveau des URIs significantes, et non pas des tags eux-mêmes. Ainsi, dans l'exemple précédent, on pourra suggérer un contenu indexé par l'URI athena:EconomieDEnergie lors de la lecture du billet associé à l'URI athena:PompeAChaleur, puisqu'il existe (via par exemple une relation SKOS créée à partir du wiki destiné aux domaines et métiers) une relation entre ces deux concepts. Nous détaillerons ces possibilités de découverte de contenus et de thématiques proches dans le chapitre suivant (Section 5.4.3, page 216).

Un aspect mis en avant par le scénario précédent et qui nous semble important quant à l'utilisation de MOAT dans ce contexte d'écosystème sémantique pour l'Entreprise 2.0 est l'utilisation d'instances créées par les wikis sémantiques évoqués précédemment (Section 4.2.2, page 154) afin de définir les significations associées aux tags. Par exemple, un premier utilisateur va créer l'*Association des Maires de France* (instance de foaf:Organization) via UfoWiki, un second pouvant ensuite définir ce concept comme signification associée au tag amf. Il s'agit donc d'un enchaînement naturel entre wikis, ontologies, instances et folksonomies pour enrichir les capacités d'annotations proposées par le système initial. Bien entendu, les principes de MOAT ne sont pas limités aux instances produites ou stockées en interne, comme par exemple les différentes instances de geonames:Feature récupérées depuis Geonames. Il est donc possible d'utiliser n'importe quel concept pour représenter les significations des tags, notamment ceux proposés par l'initiative *Linking Open Data*, comme nous le verrons en détaillant une implémentation publique du client associé ainsi que l'outil LODr (Section 4.3.2, page 178).

Un autre intérêt de ce *framework* est selon nous son cadre d'architecture participative. Puisque les liens entre tags et significations sont partagés au sein d'un serveur de tag utilisé par une communauté donnée, un utilisateur assignant une signification donnée à un tag permet à l'ensemble de la communauté de bénéficier de cette association, comme le montre

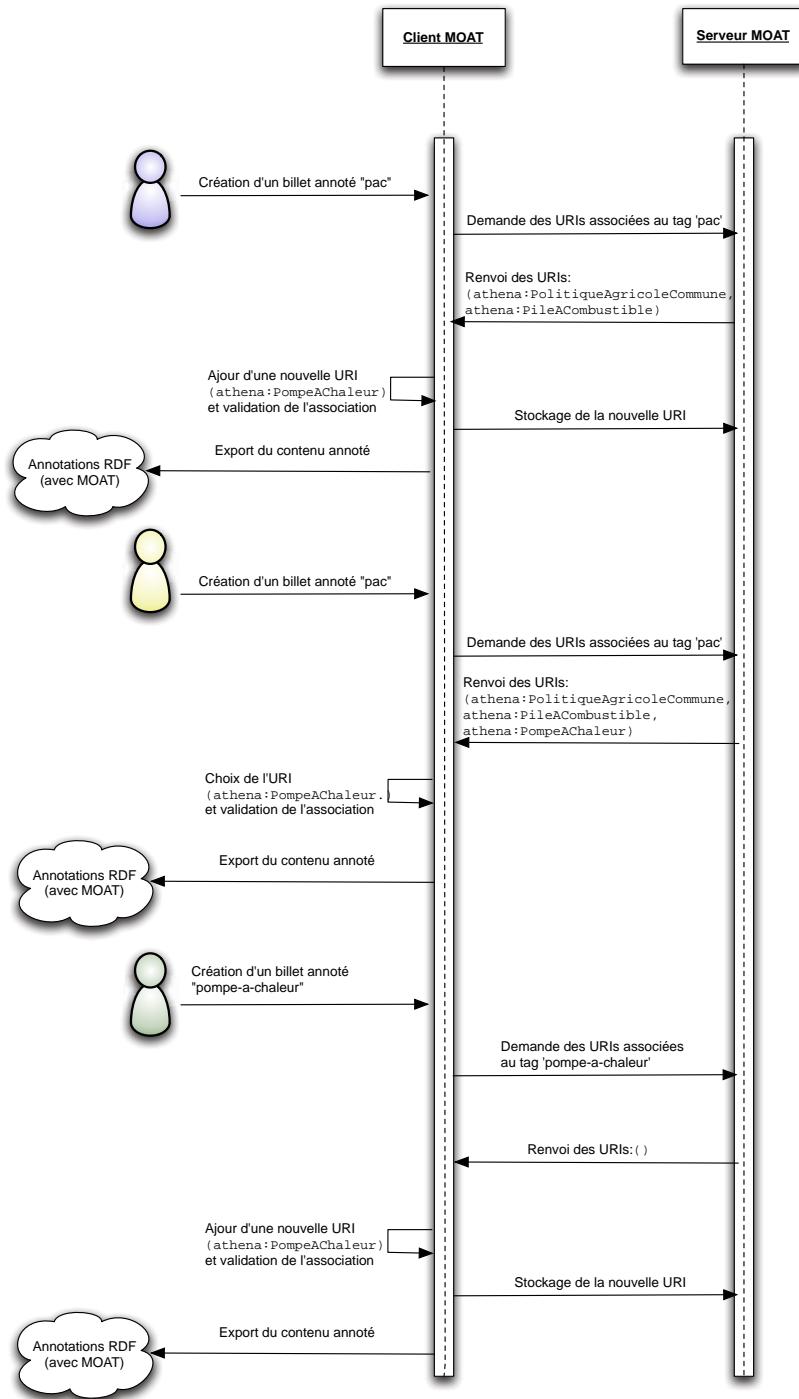


Figure 4.19: Workflow client / serveur et processus MOAT

le scénario précédent avec le tag pac. Le bénéfice de cette architecture de participation est d'autant plus flagrant dans le cas où le processus est combiné avec des instances créées depuis nos wikis sémantiques : les wikis, peuplés par différents utilisateurs, viennent en support de notre folksonomie, également utilisée par différentes personnes. De plus, cette architecture n'est pas figée comme nous l'avons signalé, puisque chaque communauté peut installer son propre serveur, dans la continuité de ce qui est proposé par Annotea (Section 3.3.1, page 125). Les utilisateurs ne sont donc pas liés à un unique serveur central et référent, choix motivé par une optique d'ouverture des données sociales (Section 3.1.5, page 96).

4.3.2 Implémentations logicielles

Client générique pour Drupal

Comme exposé en amont, le rôle d'un client MOAT est (1) d'interagir avec un serveur pour récupérer les significations globales d'un tag au moment de la création d'un contenu, (2) de permettre à l'utilisateur de choisir quelle signification locale il souhaite donner à son tag dans ce contexte et éventuellement d'en définir une nouvelle et (3) de produire les annotations sémantiques associées. Bien entendu, si ces différentes étapes font intervenir des modèles et des mécanismes d'échange reposant sur RDF, il est évident que l'utilisateur ne doit pas directement y être confronté, le client devant être aussi intuitif que possible. Pour ce faire, nous avons implémenté un client MOAT générique pour la plate-forme Drupal⁴⁹.

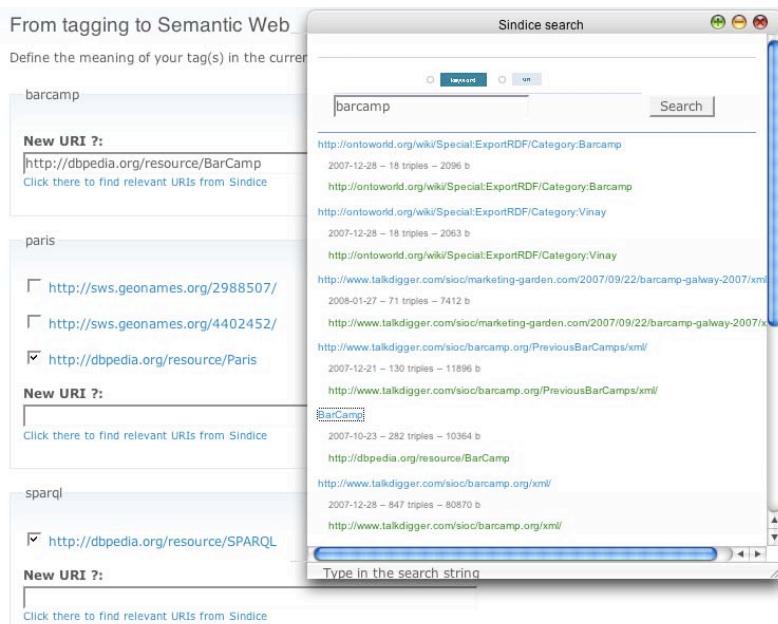


Figure 4.20: Interface utilisateur du module MOAT pour Drupal couplée au *widget* Sindice

Ce premier client se présente donc sous la forme d'un *plug-in* Drupal et propose pour chaque contenu tagué un onglet supplémentaire permettant d'assigner aux différents tags

⁴⁹<http://drupal.org/project/moat>

utilisés leur signification locale. La figure suivante illustre son fonctionnement, ici pour un billet de blog au sujet d'un événement relatif au Web Sémantique ayant eu lieu à Paris auquel ont été associés trois tags (Figure 4.20, page 174). Celle-ci met en avant trois cas possibles, tels qu'ils ont été évoqués précédemment :

- en bas (tag `sparql`), une unique signification a été récupérée depuis le serveur. Il s'agit ici de l'identifiant associé au langage SPARQL via DBpedia. Dans ce cas, la signification a été validée par l'utilisateur (boîte de dialogue cochée) ;
- au dessus, trois URIs ont été récupérées pour le tag `paris`, l'utilisateur ayant fait son choix parmi celles-ci. Afin de faciliter ce choix, les différentes URIs sont proposées en tant qu'hyperliens de manière à ce que l'utilisateur puisse en savoir plus à leur sujet⁵⁰. On remarque également dans cet exemple que si trois URIs sont affichées, en réalité deux d'entre elles sont liées par une propriété `owl:sameAs`, signifiant qu'en dépit de deux URIs distinctes il s'agit de la même instance. Si cette relation n'est pas prise en compte pour le moment dans notre client, nous souhaitons l'intégrer aux prochains développements de manière à limiter les URIs à afficher à celles correspondant à des ressources réellement distinctes évitant ainsi de surcharger l'interface ;
- enfin, pour le tag `barcamp` (premier dans l'interface), aucune URI n'a été associée. Ici, l'utilisateur a la possibilité d'en ajouter une nouvelle dans le champ textuel correspondant. Pour simplifier ce processus, nous avons intégré le *widget* proposé par Sindice⁵¹ [Tummarello et al., 2007]. Cet index du Web Sémantique met en effet à disposition des développeurs un service permettant, pour un terme donné, de lister un ensemble d'URIs correspondantes⁵² (en fonction par exemple du label associé au concept défini par cette URI ou de l'URI elle-même). Cette intégration facilite ainsi le choix de nouvelles URIs pour l'utilisateur.

Ce module MOAT pour Drupal permet également l'export des données ainsi annotées. L'ensemble du contenu exporté utilise donc SIOC, la *Tag Ontology* et MOAT comme nous l'avons présenté dans le chapitre précédent (Figure 3.19, page 132). Ce *plug-in* se base d'ailleurs sur le *plug-in* SIOC que nous avons développé et présenté au début de ce chapitre (Section 4.1.2, page 142). Notons également qu'après notre implémentation pour Drupal, un client MOAT a été développé par OpenLink au sein de la plate-forme OpenLink DataSpaces⁵³ (Section 2.3.4, page 77).

Adaptation du client au sein d'Hermès

Afin de faciliter l'intégration de MOAT au sein de la plate-forme Hermès, nous avons procédé à différentes adaptations par rapport au *plug-in* que nous venons de présenter. Nous pouvons en effet faire deux reproches majeurs à celui-ci notamment dans un objectif d'acceptation grand public :

- proposer des URIs pour choisir la signification d'un tag est vraisemblablement peu intuitif et ce même si elles sont représentées sous forme de liens hypertextes vers leur

⁵⁰ En supposant bien sur qu'elles soient déréférencables et renvoient vers un ensemble d'informations à leur sujet, selon les bonnes pratiques définies par [Berners-Lee, 2006a] (Section 1.1.4, page 27).

⁵¹ <http://sindice.com/>

⁵² <http://sindice.com/developers/widget>

⁵³ <http://vanirsystems.com/danielsblog/2008/02/09/a-few-new-features-in-openlink-data-spaces/>

description. L'objectif de MOAT étant de rendre le processus d'indexation sémantique le plus simple possible et accessible au plus grand nombre, nous sommes ici face à une contradiction en confrontant directement l'utilisateur à la notion d'URI (confrontation encore plus générale avec notre vision d'un système ne déstabilisant pas les utilisateurs) ;

- en admettant qu'il n'existe pas de concept relatif à la signification souhaitée, il est nécessaire de passer par un outil annexe pour créer celui-ci, puis de retourner ensuite au client MOAT pour associer le tag à l'URI de l'instance nouvellement créée.

Pour prendre en compte ces deux problématiques, le client MOAT mis en place dans le cadre de la plate-forme a été adapté de la manière suivante. Tout d'abord, puisque notre approche se base principalement sur l'utilisation de concepts instanciés en interne, notamment via les wikis identifiés précédemment, nous proposons en lieu et place des URIs d'afficher les labels des différentes significations proposées⁵⁴. Ceci rend l'interface beaucoup plus conviviale en termes de validation du concept (classe ou instance) approprié pour un tag comme l'illustre la figure qui suit (Figure 4.21, page 176). Pour aller plus loin, on peut imaginer un lien (voire une pop-up), qui affiche une description plus complète du concept, en utilisant différentes propriétés définies pour celui-ci (par exemple `dct:description`). Notons également que lorsque le tag utilisé est non ambigu et déjà associé à un concept, le lien est pré-validé pour simplifier la démarche d'annotation sémantique, l'utilisateur n'accédant ainsi à cette interface de validation que si le concept associé ne lui convient pas.

Association de mots-clé à l'ontologie

Ici vous pouvez indiquer le sens associé à votre mots-clé en utilisant l'ontologie

"Alstom"

Alstom (Organisation)

Si rien ne convient, vous pouvez [passer à l'interface avancée](#)

"GDF-Suez"

Si rien ne convient, vous pouvez [passer à l'interface avancée](#)



Figure 4.21: Choix d'un concept pour désambiguïser un tag au sein du client MOAT Athéna

Lorsqu'aucune URI ne convient, comme c'est par exemple le cas pour le tag GDF-Suez dans cet exemple, l'utilisateur a la possibilité de passer à une interface avancée qui lui permet de parcourir la taxonomie des classes associée à nos différentes ontologies afin de choisir la classe ou l'instance signifiante associée à ce tag. Ici, ces différentes classes et instances sont à nouveau identifiées par leur label, les URIs associées étant masquées à l'utilisateur (Figure 4.22, page 177).

Si à nouveau aucune URI n'est disponible pour représenter la signification souhaitée, l'interface permet la création d'une nouvelle instance⁵⁵. L'utilisateur a alors la possibilité de

⁵⁴Notons que nous pourrions envisager ce type d'interface pour les données du Web, mais à un coût plus important puisqu'il faudrait déférerencer chaque URI, identifier son label, etc.

⁵⁵On retrouve également ce type d'interface de création d'instance à partir de tags dans SweetWiki.

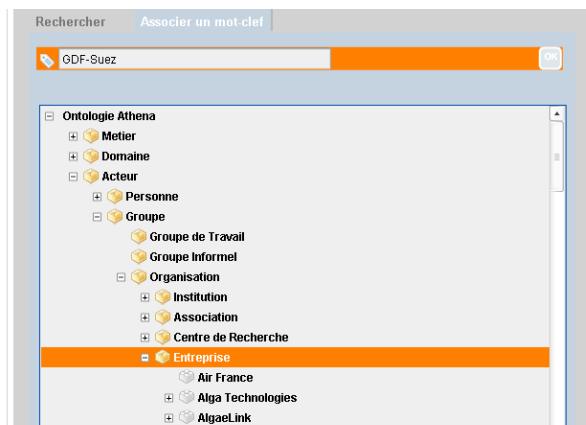


Figure 4.22: Parcours de la taxonomie des classes pour définir une nouvelle signification

sélectionner une classe, le système demandant alors s'il souhaite :

- associer le tag à la classe, *i.e.* considérer la classe comme la signification de ce tag, par exemple pour un tag générique `entreprise` qui serait associé à `foafplus:Company` ;
- associer le tag à une nouvelle instance de la classe en question, *e.g.* dans notre exemple choisir d'associer GDF-Suez à une nouvelle instance de `foafplus:Company`. Dans ce cas, l'instance est automatiquement créée et typée selon la classe choisie et l'utilisateur a la possibilité de définir un label plus parlant que le tag lui-même afin d'identifier la nouvelle instance (Figure 4.23, page 177).

Dans les deux cas, le tag est associé à cette nouvelle URI via MOAT, à la fois localement (pour l'action de *tagging* en cours) et globalement (au sein du serveur).

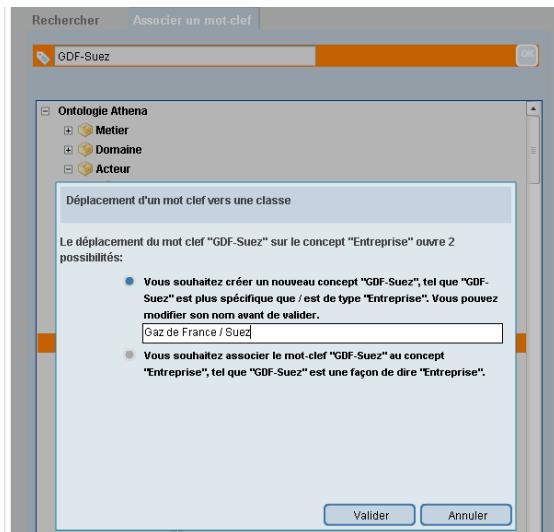


Figure 4.23: Création d'une nouvelle instance et association d'un tag via le client MOAT

Cette interface permet de plus de visualiser l'ensemble des tags associés à un concept. On peut ainsi voir que trois tags différents sont associés à l'instance W3C (Figure 4.24, page 178), l'un d'entre eux ayant été comme nous l'avons dit assigné automatiquement via UfoWiki à partir du label de cette instance.



Figure 4.24: Visualisation des différents tags associés à un concept

LODr : Indexation sémantique pour des contenus Web 2.0 existants

Si les deux interfaces précédentes proposent des possibilités d'annotations sémantiques avec MOAT pour des documents créés spécifiquement au sein des plates-formes associées, il nous a semblé pertinent de proposer un moyen d'utiliser ces mêmes principes pour indexer sémantiquement des données présentes sur le Web. Nous avons ainsi développé l'application LODr [Passant, 2007a] permettant d'annoter, via MOAT, des contenus produits depuis diverses applications Web 2.0 : Flickr, SlideShare, Delicious, etc.

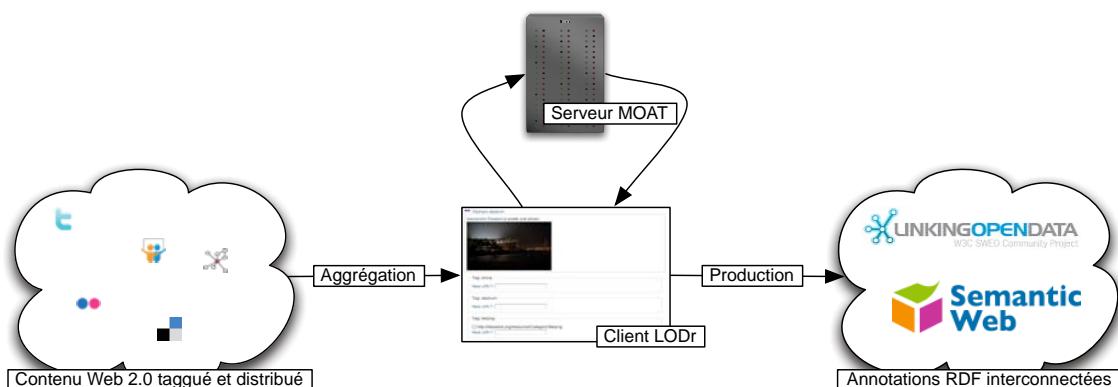


Figure 4.25: Architecture de LODr

Le principe de l'application, qui repose également sur la notion de clients et de serveur de tags, est le suivant (Figure 4.25, page 178) :

- un utilisateur installe l'application LODr⁵⁶ sur son serveur Web, et s'identifie via son URI principale (e.g. <http://apassant.net/alex>). A partir de cette URI, l'appla-

⁵⁶<http://code.google.com/p/moat-project/>

tion va identifier les différents services Web 2.0 auquel cet utilisateur a souscrit, en supposant que cette URI soit déréférencable et que ces informations y soient fournies en utilisant `foaf:holdsAccount` (Section 3.1.4, page 94). Cette première étape permet également d'éviter la notion de *social network fatigue* évoquée précédemment (Section 3.1.5, page 96) et met en avant l'utilisation de FOAF comme point d'accès central à différentes activités en ligne. Notons également qu'à la différence de Faviki, LODr permet à un utilisateur de continuer à utiliser ses applications favorites pour publier et annoter ses contenus ;

- à partir de ces différents profils, l'application va identifier les flux RSS correspondants à chacun d'entre eux⁵⁷. Ces flux sont alors agrégés au sein du client et immédiatement traduits en RDF en utilisant SIOC, FOAF et la *Tag Ontology* via un système d'adaptateurs propre à chaque service. Des adaptateurs sont ainsi disponibles pour Flickr, Delicious, SlideShare, Bibsonomy ou encore Twitter, et il est relativement aisément d'en écrire de nouveaux (une vingtaine de lignes de code). Par ailleurs, nous nous sommes ici aperçus que ces différents services avaient des manières relativement distinctes de modéliser les tags dans leurs flux RSS, certains utilisant une propriété `dc:subject`, d'autres leur propre propriété, etc., renforçant le besoin d'une sémantique commune pour représenter ceux-ci (Section 3.3.1, page 122). Une fois ces données traduites et représentées en RDF au sein de l'application, chaque élément de flux est immédiatement exporté en RDFA au sein de l'interface de visualisation, offrant un premier niveau de sémantique commune pour des outils et silos de données initialement distincts et aux formats hétérogènes ;
- enfin, l'utilisateur a la possibilité d'associer les différents tags utilisés à des concepts existants en suivant les principes de MOAT et via une interface similaire à celle proposée par le client Drupal (Figure 4.26, page 179). Cette interface a cependant l'avantage de pouvoir : (1) se greffer à différents *endpoints* SPARQL pour suggérer des concepts en fonction du tag utilisé et (2) de proposer un label humainement lisible pour le tag à partir du moment où le concept associé à déjà utilisé au sein de l'application, les annotations RDF qui lui sont associées étant alors intégrées au sein du client.

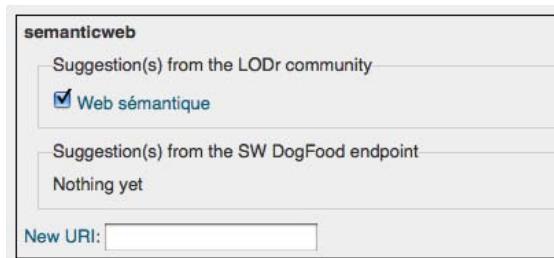


Figure 4.26: Assigmentation d'une URI à un tag particulier avec LODr

Ces trois étapes permettent ainsi de passer de contenus Web 2.0 disjoints et annotés via de simples mots-clés à un ensemble de graphes RDF interconnectés et utilisant des URIs de

⁵⁷Ce processus s'effectuant ensuite de manière régulière.

référence, permettant leur découverte et leur exploitation de manière avancée, comme nous le verrons dans le chapitre qui suit (Section 5.4.3, page 216). Signalons également qu'une fois les contenus annotés de cette manière, l'application permet de visualiser un *nuage de concepts*, en plus du traditionnel nuage de tags, celui-ci étant généré à partir des labels (`rdfs:label`) des différentes instances annotantes et pouvant de ce fait être visualisé en plusieurs langues (Figure 4.27, page 180). Le problème de multilinguisme est ainsi pris en compte en passant des tags au URIs, non seulement pour la pose de tags mais aussi pour leur visualisation.

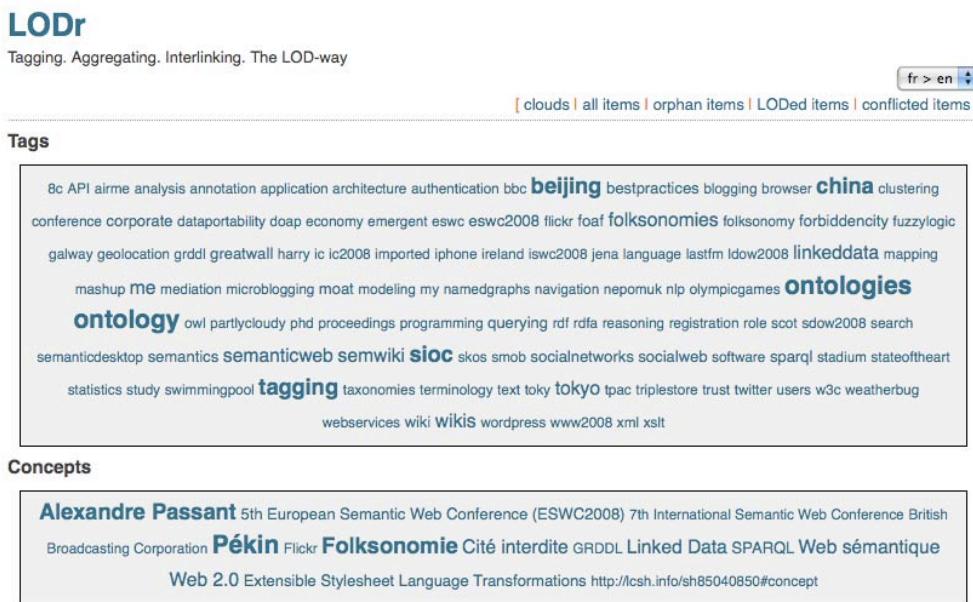


Figure 4.27: Nuage de concepts avec LODr

Serveur de tags et protocoles de communication

La notion de serveur de tags joue un rôle central au sein du processus d'indexation sémantique lié à MOAT. C'est en effet en son sein que vont être stockées les différentes significations globales des tags et c'est par son intermédiaire qu'elles sont délivrées aux clients.

Ces significations, représentées sous forme d'annotations RDF définies avec MOAT (Listing 3.15, page 131), sont stockées au sein du serveur non pas sous forme de fichiers (e.g. un fichier associé à chaque tag) mais dans un entrepôt de données RDF. Notre implémentation d'un serveur MOAT est écrite en PHP et est également disponible librement⁵⁸. Plutôt que de dépendre d'un entrepôt spécifique, le serveur dispose d'une interface qui lui permet de se greffer sur différentes solutions logicielles existantes. Deux entrepôts sont pour le

⁵⁸<http://moat-project.org/server>

moment supportés : ARC2⁵⁹ et 3store⁶⁰. L'adaptateur ARC2 est en réalité un adaptateur générique qui permet de s'adapter à n'importe quel entrepôt supportant les langages SPARQL et SPARUL pour l'ajout de nouvelles données (avec sa clause LOAD) (Section 5.1.3, page 192). L'interaction client-serveur, aussi bien en lecture qu'en mise à jour, repose exclusivement sur le protocole HTTP et sur les principes d'architecture REST – Representational state transfer [Fielding, 2000] – ce qui rend assez simple son intégration au sein de systèmes Web 2.0 existants, puisqu'il n'est pas nécessaire de repenser l'architecture applicative. Enfin, la communication entre clients et serveur se fait par l'échange de graphes d'annotations MOAT, RDF étant ici idéalement utilisé comme format d'échange entre différents composants logiciels.

Concernant la définition des tags et l'interrogation du serveur pour en obtenir la liste des significations globales, nous utilisons les principes définis par [Berners-Lee, 2006a], notamment avec des URIs déréférençables pour chacun d'entre eux. À chaque URI de tag est donc associée sa représentation, *i.e.* la description du tag (label) et ses différentes significations globales. Ceci permet pour chaque tag d'identifier simplement les différentes significations globales qui lui sont rattachées, chaque tag portant de cette manière lui-même la sémantique qui lui est associée. La représentation renvoyée dépend également de l'agent logiciel qui effectue la requête, renvoyant ainsi à partir de l'URI d'un tag soit une description HTML, soit une description RDF. Il est également possible d'obtenir une représentation JSON⁶¹ du tag, toujours dans cette optique de faciliter le travail des développeurs qui n'ont ainsi pas à apprêhender les principes de représentation RDF, tout comme pour l'API SIOC.

Pour chaque tag, nous disposons donc :

- d'une URI déréférençable qui l'identifie, déterminée en fonction de l'adresse du serveur et du label du tag, par exemple <http://tags.moat-project.org/tag/sparql> ;
- d'une URL relative à sa description RDF, *e.g.* <http://tags.moat-project.org/tag/sparql/rdf> ;
- d'une URL relative à sa description HTML, *e.g.* <http://tags.moat-project.org/tag/sparql/html> ;
- d'une URL relative à sa description JSON, *e.g.* <http://tags.moat-project.org/tag/sparql/json>. Pour ce dernier point, nous avons également défini la possibilité d'ajouter un paramètre supplémentaire *light*, permettant de délivrer une description plus légère ne prenant pas en compte la composante sociale (via FOAF) des assignations entre tags et URIs, par exemple <http://tags.moat-project.org/tag/sparql/json/light> ;

La mise à jour du serveur, *i.e.* l'ajout de nouvelles significations globales, s'effectue selon des principes similaires, le client envoyant les nouvelles significations au serveur qui les stocke en son sein.

Un autre aspect qui nous semble important et qui n'est pas pour le moment pris en compte dans les différentes implémentations MOAT (client ou serveur) est la prise en compte du réseau social de l'utilisateur pour affiner la suggestion des tags. Comme nous l'avons vu,

⁵⁹<http://arc.semsol.org>

⁶⁰<http://threestore.sf.net>

⁶¹JSON – JavaScript Object Notation – est en effet un format de représentation populaire dans les applications Web 2.0 permettant la représentation d'objets structurés en JavaScript. – <http://json.org>

le modèle MOAT associe en effet à chaque signification globale l'ensemble des utilisateurs ayant considéré celle-ci (Section 3.3.3, page 128). Ainsi, il est imaginable de renvoyer non pas toutes les significations possibles pour un tag lorsque l'utilisateur interroger un serveur, mais uniquement celles définies par des personnes définies comme proches, par exemple avec la propriété `foaf:knows` ou en utilisant des notions de groupes d'intérêt, pouvant être également représentés avec FOAF ou SIOC. Ceci nous semble particulièrement pertinent dans un contexte d'Entreprise 2.0 : en supposant que différentes communautés utilisent le même serveur, un utilisateur de la communauté solaire se verra suggérer en priorité les significations de personnes de sa communauté. Nous pensons que cette méthode permettra de proposer des suggestions de significations pertinentes, puisque centrée sur une communauté d'intérêt plus restreinte et associée à l'utilisateur en faisant la demande.

4.4 RETOUR SUR L'UTILISATION DE MOAT DANS NOTRE CONTEXTE D'ENTREPRISE 2.0

Pertinence d'une telle approche

Un critère important en termes d'évaluation de MOAT nous semble être la manière dont il permet de résoudre les problématiques évoquées auparavant quant aux systèmes à base de tags (Section 2.2.3, page 63). Nous présenterons plus tard les outils de recherche associés et la manière dont ils bénéficient de notre proposition pour prendre en compte ces différentes problématiques, mais donnons simplement ici quelques statistiques qui témoignent selon nous de l'intérêt d'un tel modèle et de la pertinence de cette notion de concepts (URIs) en support des tags. En analysant notre folksonomie d'origine et les différentes annotations RDF représentées avec MOAT grâce au *workflow* et outils présentés dans ce chapitre, nous avons constaté que 1176 tags avaient été associés à 715 URIs de significations différentes.

Comme nous l'avons déjà évoqué (Section 2.2.3, page 63), nous n'avons constaté que très peu de tags sujets aux problématiques d'ambiguïté dans notre contexte, seul un d'entre eux étant associé à plusieurs URIs⁶². En contrepartie, nous avons constaté un problème d'hétérogénéité beaucoup plus présent, comme le montre le tableau qui suit (Tableau 4.2, page 183). On observe ainsi que si 510 URIs sur les 715 recensées ne sont pas sujettes à des problèmes d'hétérogénéité, puisqu'assignées à un seul tag, 205 le sont. 96 instances ont ainsi été associées à deux tags, 70 à trois d'entre eux et 39 à quatre tags ou plus. Par exemple l'instance définissant la notion de *Supercapacité* est associé aux tags *supercapacité*, *supercondensateur*, *ultracapacité*, *ultracapacitor*, *ultracondensateur*. On retrouve ici aussi bien des variations de synonymie (*supercondensateur*, *ultracondensateur*, etc.) que des variations causées par la nature multilingue des tags (*ultracapacité*, *ultracapacitor*). Nous avons également observé que si cette hétérogénéité est en général le fait de plusieurs utilisateurs annotant avec différents tags, elle peut également émerger à un niveau personnel. Ainsi, dans l'exemple précédent, trois utilisateurs ont permis d'arriver à cette hétérogénéité, l'un d'entre eux utilisant trois tags distincts. Nous avons plus particulièrement constaté cette hétérogénéité personnelle au niveau de tags et d'instances représentatifs de noms de

⁶²Notons que cela ne signifie pas qu'un seul tag est ambigu dans la folksonomie, puisque seuls 1176 tags sur un total de 12257 ont été ici considérés.

personne (nom complet et nom de famille), de zones géographiques (par exemple USA et états-unis et de technologies (synonymie, multilinguisme mais aussi abréviations).

Il nous semble ainsi que notre approche consistant à utiliser des représentations formelles en support de ces tags pour définir leur signification prend tout son sens. Nous verrons plus tard de quelle manière un moteur de recherche dédié aux différents documents annotés dans le cadre de notre plate-forme tire bénéfice de MOAT pour prendre en compte ce problème abondant d'hétérogénéité.

Nombre de tags associés à l'URI	Nombre d'URIs correspondant
1	510
2	96
3	70
4 et plus	39

Tableau 4.2: Distribution des tags au sein de la plate-forme Hermès

Adoption de MOAT sur le Web

En termes d'adoption, bien que le modèle soit utilisé dans certaines applications (en dehors de celles que nous avons développées (Section 4.3, page 170)) comme par exemple Openlink Data Spaces ou GroupMe, nous devons reconnaître que l'impact de MOAT sur le Web est beaucoup moins important que celui de SIOC (Section 3.1.6, page 101). Cependant, l'idée défendue ici – et initiée avant la définition de MOAT et des outils associés [Passant *et al.*, 2006] [Passant, 2007c] – à savoir l'utilisation de connaissances formelles, non ambiguës et interopérables en support des systèmes à base de tags a été récemment intégrée au sein du standard SST – *Simple Semantic Tagging*⁶³ –, auquel nous avons contribué et dont un des objectifs est de simplifier et populariser cette idée de tags sémantiques sur le Web. Mené par différents acteurs du Web 2.0 et du Web Sémantique (Yahoo ! SearchMonkey, Free-Base, Zemanta, Faviki, AdaptiveBlue⁶⁴, DERI Galway) il vise à proposer un modèle certes moins complet que ceux définis précédemment mais en contrepartie beaucoup plus simple à prendre en main, dans un objectif d'adoption à grande échelle de cette pratique au sein d'outils Web 2.0, principalement via des annotations RDFa pour représenter ces liens entre tags et URIs. Le modèle proposé est par ailleurs aligné avec certains vocabulaires présentés ici, dont MOAT⁶⁵⁶⁶

CONCLUSION

Dans ce chapitre, nous avons présenté différents outils permettant la production d'annotations sémantiques à partir d'outils Web 2.0. Nous avons tout d'abord présenté différentes

⁶³<http://semantictagging.org>

⁶⁴<http://www.adaptiveblue.com/>

⁶⁵<http://semantictagging.org/ns>

⁶⁶Note aux relecteurs : En raison de différentes contraintes légales au sein de certaines entités participant à ce standard, celui-ci n'a pas encore été annoncé officiellement mais devrait l'être courant Mars.

applications dédiées à la production automatisée d'annotations socio-structurelles depuis les blogs, wikis et flux RSS via des alignements entre ontologies et structure internes. Nous avons ensuite détaillé UfoWiki, plate-forme de wikis enrichie de fonctionnalités permettant le peuplement d'ontologies via un système assisté d'annotations, en présentant notamment la manière dont cet outil permettait de s'intégrer plus globalement dans la vision d'un *Web of Data*. Nous avons également pu voir la manière dont cet outil a été adopté dans notre contexte afin de permettre un peuplement collaboratif d'ontologies de domaine. Enfin, nous avons présenté les différents processus et outils associés à MOAT, permettant de passer d'un processus simple de *tagging* à la production d'annotations sémantiques dans un but d'indexation de contenus Web 2.0 via des URIs de classes ou d'instances d'ontologies.

Ainsi, il est important de garder en tête le rôle joué par l'utilisateur final dans ces différents outils, qu'il soit acteur pour la production d'annotations (cas des wikis et de MOAT) ou bien qu'il soit pris en compte dans les annotations elles-mêmes (production automatisée d'annotations socio-structurelles). Cette convergence entre Web 2.0 et Web Sémantique est ainsi rendue possible en prenant en compte aussi bien le facteur humain que l'implémentation logicielle, l'utilisateur ayant un rôle primordial à jouer dans la réussite de cette convergence. Nous allons maintenant, dans le chapitre qui suit, nous intéresser à la manière d'exploiter ces différentes annotations.

Chapitre 5

Intégration et utilisation d'annotations sémantiques distribuées

INTRODUCTION

Alors que nous avons détaillé dans le précédent chapitre la production d'annotations sémantiques depuis différents outils, nous allons ici nous concentrer sur leur utilisation. Plus particulièrement, nous allons voir en quoi ces annotations permettent la mise en place de services innovants en termes d'intégration de données, de navigation et d'accès à l'information. Ce chapitre permettra ainsi de voir l'apport concret de la méthodologie *SemSLATES* et de l'utilisation des technologies du Web Sémantique au sein de systèmes d'Entreprise 2.0.

Pour débuter, nous reviendrons sur la nature distribuée des annotations que nous cherchons à utiliser, organisation due à la nature même des processus d'annotations (Section 4, page 137). Nous présenterons différentes méthodes permettant d'exploiter ces données ainsi réparties et argumenterons notre choix de disposer d'un entrepôt de données au cœur de l'architecture de médiation. Nous détaillerons également les critères que nous avons pris en compte pour le choix de celui-ci parmi les solutions existantes sur le marché. Enfin, nous insisterons sur son intégration au sein de l'architecture existante. Plus particulièrement, nous détaillerons les protocoles de communication mis en place qui nous permettent de disposer d'un niveau d'abstraction tel que l'architecture est indépendante de l'entrepôt lui-même, aussi bien en termes d'intégration de données que de services venant exploiter celles-ci.

Puisque c'est ici ce qui nous intéresse, nous allons ensuite détailler différents services exploitant ces annotations. Tout d'abord, nous nous concentrerons sur la manière dont Ufo-Wiki (Section 4.2.2, page 154) tire bénéfice des différentes annotations produites en son sein. Nous verrons plus particulièrement comment ces annotations sont utilisées pour la mise en place de pages dynamiques via un processus de macros sémantiques (Section 5.2.1, page 196). Ces macros, qui peuvent être contextualisées (Section 5.2.2, page 201), permettent notamment de masquer à l'utilisateur la complexité des requêtes SPARQL associées à l'interrogation de ces graphes d'annotations. Nous verrons également comment ces annotations sont utilisées pour la mise en place de processus avancés de visualisation, au travers d'interfaces à facettes et de *mash-ups* sémantiques combinant instances d'ontologies peuplées depuis les wikis et données RDF publiques (Section 5.2.3, page 203).

Ensuite, nous verrons comment ces annotations permettent, en couplant différents niveaux de représentation (SIOC, MOAT et ontologies du domaine), d'interconnecter blogs

et wikis, deux éléments distincts en termes d'outils mais liés par les données auxquelles ils font référence (Section 5.3.1, page 207). Cette interopérabilité nous permet de répondre à la problématique de fragmentation d'informations au sein de systèmes d'Entreprise 2.0, où les informations au sujet de différents objets sont réparties entre plusieurs services. Nous montrerons ensuite (1) en quoi il est possible d'utiliser ces annotations pour l'indexation automatique de flux RSS entrants (Section 5.3.2, page 209) et (2) de quelle manière elles permettent d'augmenter l'expérience utilisateur en termes de navigation des contenus internes via un système de projection des connaissances (Section 5.3.3, page 211).

Enfin, nous détaillerons les principes et la mise en place d'un moteur de recherche sémantique intégré au sein de cette architecture de médiation (Section 5.4, page 212). Nous expliciterons tout d'abord ce que nous entendons par recherche d'information sur le Web Sémantique et détaillerons ensuite le fonctionnement de ce moteur et la manière dont il tire bénéfice des différentes annotations produites et des ontologies associées, tout en masquant à nouveau la complexité du système aux utilisateurs. Pour finir, nous montrerons aussi comment il est possible d'exploiter cette sémantique pour étendre la recherche d'information et suggérer de nouveaux concepts, en se basant sur les relations existantes au sein des graphes d'annotations.

5.1 STOCKAGE DES DONNÉES ET PROTOCOLES ASSOCIÉS

5.1.1 De la nécessité d'un entrepôt de données

Comme nous l'avons vu au travers des chapitres précédents, notre proposition d'écosystème sémantique pour l'Entreprise 2.0 repose sur (Figure 5.1, page 187) :

- différents outils destinés à la production et à l'édition de contenus (blogs, wikis, agrégateur RSS) pour lesquels les actions utilisateur et la composante sociale jouent un rôle important (Section 2.1, page 50) ;
- des graphes d'annotations sémantiques produits à partir de ces outils, relatifs d'une part à la structure et aux interactions sociales qui en découlent et d'autre part au contenu même des documents produits via ces outils (Section 4, page 137) ;
- un ensemble cohérent d'ontologies légères venant en support de ces différentes annotations, où l'on distingue notamment celles dédiées à la représentations des métadonnées socio-structurelles de celles portant sur données métier (Section 3, page 83).

L'ensemble des données RDF(S)/OWL ainsi produites et utilisées forme ainsi un unique graphe de représentation – via des liens directs entre instances ou par l'utilisation d'ontologies communes. Or, celui-ci – en plus d'être relativement complexe du fait des différents niveaux de représentation qu'il prend en compte – est fortement distribué au sens où les différents sous-graphes (*i.e.* les documents RDF) qui le composent sont répartis dans l'écosystème sémantique mis en place. En effet, à chaque document produit correspond un – ou plusieurs pour les wikis – graphe(s) d'annotations, stockés au niveau de l'outil d'origine, les ontologies étant elles stockées au sein d'un serveur central, certaines étant cependant réparties sur le Web (Figure 5.2, page 188).

Comme nous l'avons mentionné auparavant (Section 2.3, page 69), nous avons fait le choix de ne pas interroger à la volée les différents graphes d'annotations mais de stocker

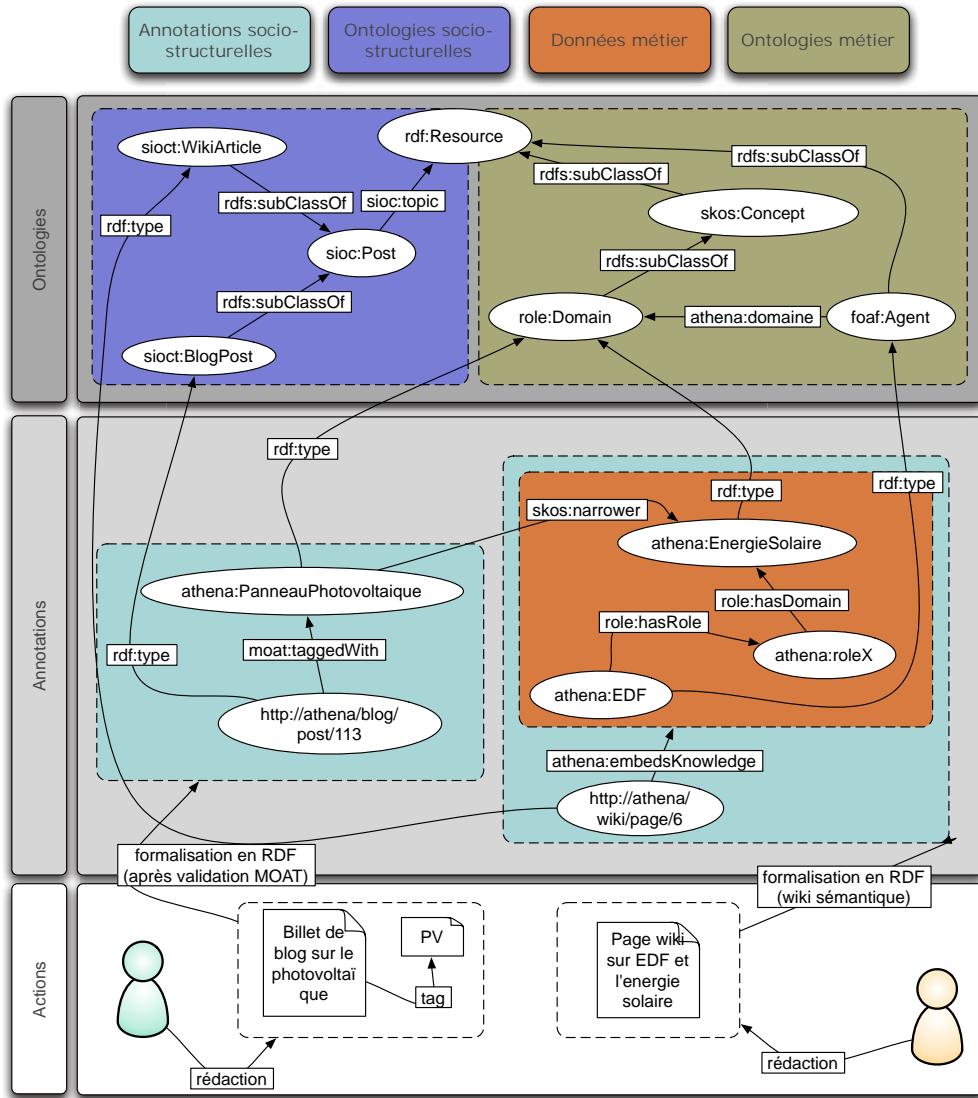


Figure 5.1: Vision globale des actions, annotations et ontologies d'un écosystème sémantique pour l'Entreprise 2.0

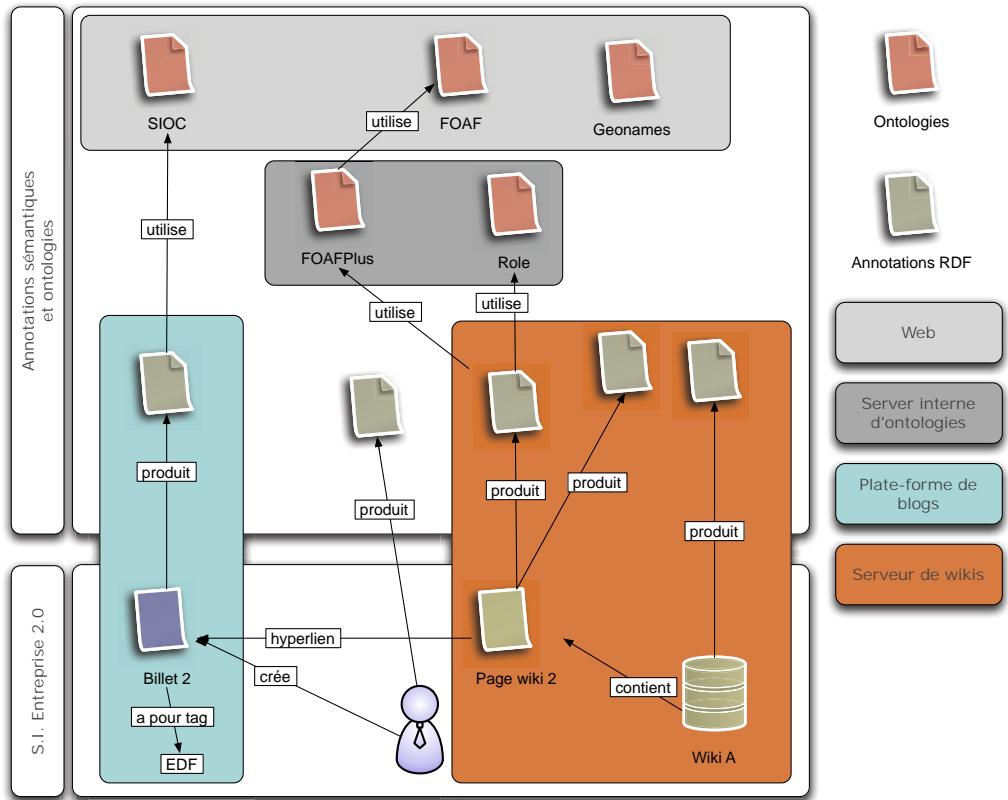


Figure 5.2: Répartition des ontologies et annotations au sein du système

ceux-ci au sein d'un entrepôt centralisé au sein de notre architecture de médiation. Ce choix, essentiellement motivé par des raisons de performances, fait ainsi de notre architecture une approche hybride entre les systèmes de médiation classiques (qui considèrent l'interrogation de données à la source) et les entrepôts de données, à la manière de ce que propose par exemple [Xyleme, 2001]. Nous allons ici présenter les différents arguments qui nous ont conduits à cette décision, en décrivant notamment les alternatives qui se sont offertes à nous et les problèmes qu'elles soulèvent.

Une première solution, si l'on ne souhaite pas disposer d'entrepôt global, est d'interroger directement les données à la source au moment de la requête¹. Il est ici nécessaire de considérer l'ensemble des annotations si l'on veut disposer d'une base de connaissances exhaustive pour y répondre². Ceci pose deux problèmes majeurs :

- il faut tout d'abord accéder à chaque graphe d'annotations et donc connaître son existence et son emplacement sur le réseau. Bien que l'utilisation de liens rdfs:seeAlso

¹Tout au long de ce chapitre, nous ferons référence à la notion de requêtes SPARQL quand nous parlerons de requêtes et d'interrogation de données.

²Exhaustive, et non complète, en raison de la notion de monde ouvert lié au Web et au Web Sémantique.

offre cette possibilité (Section 4.1.3, page 143), cela implique de parcourir chaque document pour en découvrir de nouveaux. Nous ne disposons donc pas de moyen immédiat d'en obtenir une liste complète qui permettrait leur chargement en mémoire ;

- de plus, les annotations étant réparties, le temps de latence lors de l'accès à celles-ci est également à prendre en compte, sans parler des possibilités d'erreur réseau qui sont aussi à considérer. Le tout étant bien entendu lié au nombre de documents présents dans l'écosystème.

Si l'on peut se satisfaire d'une telle solution dans un système ne comptant qu'une dizaine voire centaine de graphes, l'approche est délicate dans un contexte comme le nôtre avec plusieurs dizaines de milliers de graphes. Nos différents adaptateurs ont en effet permis la production de plus de 17000 graphes relatifs aux billets de blog et près de 2000 pour les wikis, auxquels viennent s'ajouter les graphes d'annotations produits à partir des contenus RSS. De plus, la génération dynamique de données – du fait de la nature même des outils – oblige à constamment identifier les nouveaux graphes produits afin de maintenir cette base à jour³.

Malgré tout, il est sensé de penser qu'en fonction de la requête, tous les graphes d'annotations ne seront pas nécessaires pour y répondre. Par exemple, une requête comme "*Quels sont tous les acteurs investis dans l'énergie solaire*" n'aura sans doute pas besoin des annotations associées à un billet annoté par le seul concept de Centrale Nucléaire. On peut ainsi imaginer, pour répondre à une requête donnée, n'utiliser que les documents RDF qui comportent des assertions jugées nécessaires à sa résolution. Paradoxalement, cette solution impose une connaissance *a priori* des graphes à interroger en fonction des critères de recherche, impliquant par exemple un système d'index permettant d'évaluer si un graphe est nécessaire ou non à la résolution d'une requête. Ceci nous amène à penser qu'il est tout aussi simple, quitte à centraliser des informations, de stocker directement les graphes d'annotations dans un entrepôt centralisé, comme nous l'explicitons juste après⁴. De plus, une telle structure d'index est relativement complexe à mettre en œuvre, particulièrement si l'on souhaite prendre en compte les éventuelles unions, intersections, disjonctions et autres axiomes qui peuvent être modélisés dans les ontologies utilisées en support de ces annotations.

Une autre idée, à mi-chemin entre l'agrégation complète de documents et la structure d'index est de laisser le système découvrir lui-même les graphes d'annotations potentiellement nécessaires à la résolution de différentes requêtes. C'est ce que propose en partie l'API Semantic Web Client⁵, qui permet d'effectuer des requêtes sur le Web Sémantique à partir d'un point d'entrée unique. L'API va découvrir de nouveaux graphes d'annotations en suivant les liens `rdfs:seeAlso` et identifier selon certains critères si ces graphes sont potentiellement utiles pour la requête d'origine⁶. Afin d'optimiser cette découverte de graphes, l'API permet également l'utilisation de Sindice, index du Web Sémantique que nous avons

³Devant le dynamisme des cette génération de données, nous avons également écarté les solutions qui consistent à stocker en mémoire un ensemble d'annotations, celles-ci étant en général adaptées à des jeu de données n'évoluant pas.

⁴Bien entendu, l'utilisation d'un entrepôt n'empêche pas l'apport d'un système d'index supplémentaire pour optimiser la résolution de requêtes.

⁵<http://www4.wiwiiss.fu-berlin.de/bizer/ng4j/semwebclient/>

⁶Le site Web du projet décrit l'algorithme en détail.

introduit précédemment⁷ (Section 4.3.2, page 174). À nouveau, on revient ici à l'utilisation d'une structure centralisée pour optimiser les requêtes.

Devant les difficultés et paradoxes soulevés par ces solutions, nous avons décidé la mise en place d'un entrepôt de données centralisé au sein de notre architecture de médiation, stockant les différents graphes produits par les adaptateurs en quasi temps-réel via un protocole de communication que nous décrirons plus loin (Section 5.1.3, page 192). L'utilisation d'un entrepôt couplé à ces protocoles de mise à jour offre l'avantage de permettre l'accès aux différents graphes d'annotations via un unique point d'accès clairement identifié et constamment à jour, les problèmes de découverte et d'accès aux données étant donc résolus. En conséquence, les requêtes sont à tout moment résolues en utilisant l'ensemble des connaissances produites au sein de l'écosystème. Le fait de disposer de cet entrepôt nous permet également d'envisager de meilleures performances en termes de requêtes complexes qui articulent les différents niveaux d'annotations et d'ontologies utilisés dans notre écosystème.

Si l'on peut reprocher à cette solution une certaine redondance en termes de données (les graphes d'annotations étant en effet présents à la fois au niveau des outils d'origine et au sein de l'entrepôt), gardons à l'esprit que nos travaux s'inscrivent dans un contexte industriel où un bon niveau de performances est nécessaire en termes d'accès à l'information. Notons également que la réussite de ce choix architectural repose sur le fait (comme nous allons le voir en présentant les protocoles de communication) que nous avons un certain contrôle sur les outils production de données. Malgré ces observations, cette solution nous semble donc optimale dans cette approche d'architecture de médiation sémantique venant se greffer au dessus d'un système d'Entreprise 2.0 existant.

5.1.2 Besoins et choix de l'entrepôt

La notion d'entrepôt de données RDF est sans doute aussi large que celle de système de gestion de base de données relationnelles au sens où il s'agit d'un concept générique et qu'il existe de nombreuses implémentations logicielles. Ainsi, bien que les outils de ce type partagent le même objectif de stockage et d'interrogation de données RDF, ils diffèrent par les fonctionnalités qu'ils offrent ainsi que par leur manière de gérer ces données – et bien sûr par leurs performances. Bien que les différentes méthodes de stockage utilisées (structures spécifiques, bases de données relationnelles, etc.) puissent jouer sur les performances, comme l'ont montrés différents *benchmarks* ([Lee, 2004] [Bizer et Schultz, 2008]), nous avons basé notre choix non pas sur ce critère mais sur un ensemble de caractéristiques nécessaires à la mise en place de notre architecture. Nous avons ainsi identifié les prérequis suivants :

- la gestion des graphes nommés, ou *named graphs* (Section 1.1.2, page 16). Un entrepôt supportant ceux-ci peut ainsi gérer la provenance de chaque assertion RDF, chose importante à partir du moment où notre médiateur se base sur des graphes complets d'annotations et non pas de simples triplets (*i.e.* nous considérons chaque triplet dans le contexte du graphe auquel il appartient et souhaitons conserver ce contexte). Ce support est également un prérequis à la combinaison $\{G_{m(SI)} \cup G_{d(SI)}\}$ que nous avons introduit précédemment (Section 2.3.2, page 71) et qui permet d'articuler les différents

⁷<http://www4.wiwiss.fu-berlin.de/bizer/ng4j/semwebclient/#sindice>

niveaux de sémantique proposés dans notre approche via l'utilisation de la propriété `embedsKnowledge` introduite avec d'UfoWiki (Section 4.2.2, page 154) ;

- un support du langage de requête SPARQL. Si certains entrepôts proposent leur propre langage, nous avons fait le choix dès le début de baser notre architecture sur SPARQL⁸. En particulier, nos prérequis concernant SPARQL sont un support des clauses `SELECT` et `ASK`, et des patrons `FILTER` et `OPTIONAL` (Section 1.1.3, page 25). C'est en utilisant *a minima* cette sous-grammaire que nous pourrons proposer des services avancés à nos utilisateurs, comme nous le verrons par la suite avec des exemples de requêtes utilisées au sein de nos outils. En complément du point précédent, le moteur SPARQL doit également être en mesure de supporter la clause `GRAPH` qui permet d'intégrer la gestion de graphes RDF lors de requêtes, par exemple pour limiter l'interrogation à un certain nombre de ressources (à nouveau, nous exemplifierons ceci par la suite) ;
- en complément du point précédent, le support du protocole de communication associé à SPARQL [Clark *et al.*, 2008]. Ceci nous permet en effet d'imaginer des services développés de manière indépendante de l'entrepôt et venant interroger celui-ci via son point d'accès HTTP selon un protocole standardisé ;
- la disponibilité d'une interface (au sens API) ou d'un langage de requête permettant l'ajout de données dans l'entrepôt, non pas par triplet, mais toujours par graphe complet d'annotations. Des efforts récents sur ce point se concentrent autour de SPARUL (ou SPARQL Update) [Seaborne *et al.*, 2008] et nous reviendrons plus loin sur son utilisation dans notre contexte ;
- des capacités d'inférence, *a minima* concernant les règles RDFS de subsomption associées à `rdfs:subClassOf` et `rdfs:subPropertyOf` (Section 1.1.2, page 21). Ce support va nous permettre, par exemple, pour une requête demandant de lister les instances de la classe `foaf:Agent`, de récupérer également les instances de `foaf:Person` ou de `foafplus:Company` à partir du moment où ces dernières subsument la première dans la hiérarchie de classes associée. Notons ici à nouveau que les entrepôts qui supportent ces capacités d'inférence n'emploient pas tous les mêmes stratégies, certains créant les triplets inférés lors de l'ajout de données, d'autres générant ceux-ci au moment des requêtes.

Au moment où nous avons initié notre architecture, peu d'outils proposaient l'ensemble de ces fonctionnalités⁹. Sesame¹⁰ offrait un support de SeRQL [Broekstra et Kampman, 2005], alors que son implémentation SPARQL était encore embryonnaire et ne couvrait pas la sous-grammaire que nous souhaitions. Joseki¹¹, entrepôt de données associé à l'API Jena ne supportait pas nativement les graphes nommés. ARC (dans sa première version¹²) ne permettait pas la gestion de l'inférence, alors que RAP rendait celle-ci possible mais en contrignant l'administrateur à définir lui-même les règles, celles-ci ne pouvant paradoxalement pas être

⁸Ce choix se situant dans une stratégie plus globale d'utiliser les technologies du W3C, comme nous avons pu le voir tout au long de ce mémoire.

⁹Nous avons concentré uniquement notre étude sur les solutions gratuites ou libres.

¹⁰<http://openrdf.org>

¹¹<http://joseki.sf.net>

¹²<http://bnode.org/blog/2006/02/20/arc-rdf-store-for-php-ensparql-your-lamp>

dérivées automatiquement des ontologies utilisées¹³.

Notre choix s'est finalement porté vers 3store¹⁴ [Harris et Gibbins, 2003]. Ce système d'entrepôt de données RDF supporte en effet nativement la gestion des graphes nommés (ainsi que leur utilisation avec SPARQL) et les possibilités d'inférence associées à RDFS (subsumption de classes et de propriétés) sont automatiquement assurées en fonction des ontologies intégrées à l'entrepôt. Ainsi, à partir du moment où nos différents vocabulaires sont pris en compte par celui-ci, la requête qui suit identifiera aussi bien des instances de `sioct:BlogPost` que de `sioct:WikiPage` stockées au sein de l'entrepôt, grâce aux relations définies dans le module Types de SIOC (Section 3.1.3, page 92).

```
SELECT ?item
WHERE {
  ?item rdf:type sioct:Post .
}
```

Listing 5.1: Requête SPARQL pour l'interrogation de données SIOC via un moteur supportant les principes d'inférence RDFS

Aujourd'hui, d'autres entrepôts nous semblent intéressants à considérer¹⁵ pour parvenir aux mêmes fins, comme par exemple Virtuoso [Erling et Mikhailov, 2007], AllegroGraph¹⁶, Sesame2, Mulgara¹⁷, ou encore ARC2¹⁸. Bien que ce dernier ne supporte pas l'inférence nativement, il offre la possibilité de définir des règles qui seront déclenchées lors de l'ajout de triplets afin d'y parvenir¹⁹. Virtuoso propose quant à lui une indexation plein-texte des littéraux ainsi que des possibilités de raisonnement basées sur OWL. Citons également Co-rese [Corby *et al.*, 2004], qui intègre des extensions SPARQL particulièrement intéressantes comme la notion de requêtes par chemins, et plus uniquement par patrons de triplets. Au vu de cette liste, notons que les protocoles utilisés dans notre architecture pour permettre aux différentes sources de communiquer avec l'entrepôt (Section 5.1.3, page 192) sont indépendants de l'outil utilisé et permettent ainsi un remplacement simple de celui-ci sans avoir à n'apporter de modification au reste de l'architecture. Enfin, s'il est probable que ces outils auront des performances supérieures à celui que nous utilisons actuellement, gardons en mémoire comme l'ont montré [Bizer et Schultz, 2008], qu'il n'existe pas d'entrepôt idéal et que les performances comparées varient grandement en fonction du jeu de données, de leur nombre et du type de requêtes que l'on souhaite faire.

5.1.3 Protocoles de communication

Comme nous l'avons déjà évoqué, il est nécessaire que les données stockées au sein de l'entrepôt soient constamment à jour par rapport aux données produites afin de proposer

¹³<http://apassant.net/blog/2006/03/08/relationship-vocabulary-phoaf-rap-inference-engine>

¹⁴<http://threestore.sf.net>

¹⁵Toujours parmi les solutions gratuites ou libres.

¹⁶<http://agraph.franz.com/allegrograph/>

¹⁷<http://mulgara.org/>

¹⁸<http://arc.semsol.org>

¹⁹<http://apassant.net/blog/2008/10/01/lightweight-subpropertyof-subclassof-inference-arc2>

des services optimaux aux utilisateurs. La composante industrielle de tels écosystèmes sémantiques pour l'Entreprise 2.0 impose en effet un accès pertinent à l'information, la fraîcheur et la temporalité de celle-ci jouant un rôle important dans cette pertinence. Du fait de la structure dynamique et évolutive des différents outils mis en place (qui découle des interactions sociales auxquelles ils sont liés) et de l'objectif de signalement qu'ils visent à satisfaire (notamment les blogs) il est en effet peu pertinent d'avoir un laps de temps trop important entre leur création et leur stockage, celui-ci étant nécessaire à leur interrogation.

Ainsi, un système classique de découverte de nouveaux contenus via un processus de *crawling* lancé à intervalles réguliers ne permet pas de satisfaire totalement notre objectif, du fait du décalage qui existe nécessairement entre la production de contenus et leur intégration dans les outils de recherche. Ce même problème de fraîcheur et de découverte des données RDF se pose également sur le Web, où la distribution est encore plus large et rend le *crawling* d'autant plus complexe [Harth *et al.*, 2006]. Pour aider à cette découverte, une solution pour les producteurs de données consiste à fournir des informations au sujet de la présence de nouvelles annotations. Ceci peut se faire par exemple avec l'utilisation du protocole Semantic Sitemaps [Cyganiak *et al.*, 2008] qui permet d'indiquer l'emplacement de données RDF au sein d'un site dans un format interprétable par ces *crawlers*. Cependant, il existe toujours un delta entre la production et le stockage des annotations.

Une autre possibilité, cette fois proactive, est la notion de signalement de ressources, processus mis en avant avec l'avènement des blogs et des services comme Technorati ou blo.gs²⁰. Dans ce contexte, les blogueurs peuvent configurer leurs outils afin qu'ils envoient automatiquement un signalement (ou *ping*) à ces services à chaque nouveau document créé. Concernant les données RDF, le service *Ping The Semantic Web*²¹ (PTSW) s'inscrit dans ce contexte de signalement adapté au Web Sémantique. De la même manière qu'exposé précédemment avec les blogs, les services produisant des annotations RDF ont la possibilité de signaler ceux-ci à PTSW, qui constitue ainsi un index librement accessible de documents RDF récemment produits sur le Web. Ce signalement peut en outre être effectué directement par les utilisateurs naviguant sur le Web, via l'utilisation du *plug-in* Firefox Semantic Radar²², qui va notifier PTSW de la présence de documents RDF découverts lorsque l'utilisateur navigue simplement sur le Web. On retrouve ici à nouveau les principes d'architecture de participation appliqués au Web Sémantique puisque par simple navigation volontaire du Web, un index de documents RDF se forme, celui-ci pouvant être utilisé pour la mise en place de différents services. Ces services peuvent par ailleurs être considérés comme le dernier maillon de ce que nous considérons être une chaîne complète de production, découverte et consommation de documents sur le Web Sémantique [Bojārs *et al.*, 2007b] (Figure 5.3, page 194).

Par exemple, nous avons mis en place le service doap:store²³ [Passant, 2007b] qui récupère les descriptions RDF de différents projets *open-source* (modélisées avec le vocabulaire

²⁰<http://blo.gs>

²¹<http://pingthesemanticweb.com>

²²<http://sioc-project.org/firefox>

²³<http://doapstore.org>

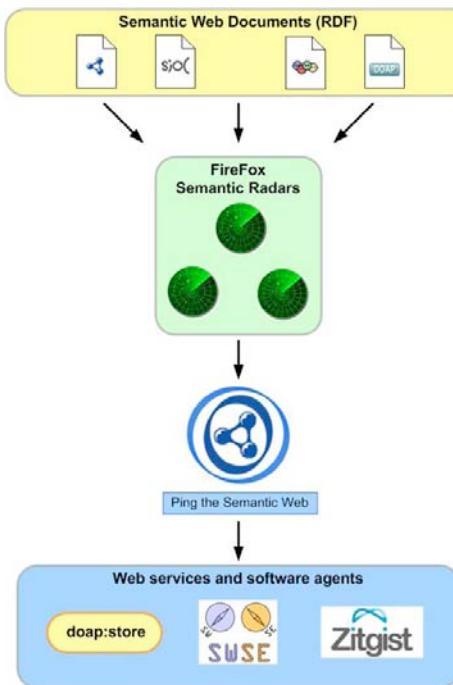


Figure 5.3: Architecture associée à PTSW pour l'indexation et la découverte de documents RDF sur le Web Sémantique [Bojārs *et al.*, 2007b]

DOAP – *Description Of A Project*²⁴) depuis PTSW afin de proposer un annuaire ouvert et distribué de projets logiciels comptant aujourd’hui plus de 9725 projets et 4645 graphes²⁵. Outre le caractère évolutif de l’annuaire grâce à un système régulier d’interrogation de PTSW pour la découverte de nouveaux projets, le principal avantage que nous voyons dans l’utilisation de cette chaîne de traitement est qu’elle résout en partie la problématique de découverte de sources de données pertinentes pour bâtir une application du Web Sémantique, permettant de se concentrer sur l’application elle-même. Ainsi, nous avons pu dans doapstore nous intéresser à la mise en place d’une interface intuitive pour parcourir différentes descriptions RDF de projets logiciels, visualisables sous la forme de simples fiches synthétiques. Le système propose également un nuage de tags extrait des différentes annotations de chaque projet ainsi qu’un moteur de recherche dédié (Figure 5.4, page 195). Si l’ensemble de l’application repose donc sur un ensemble de graphes d’annotations RDF, leur stockage au sein d’un entrepôt de données (utilisant ici OpenLink Virtuoso) et sur l’utilisation de SPARQL pour produire les différentes interfaces de recherche et de navigation, l’approche est complètement transparente pour l’utilisateur.

En reprenant les principes de cette approche de signalement et en les appliquant à l’Entreprise 2.0, nous avons donc proposé et mis en place au sein de notre architecture un sys-

²⁴<http://trac.usefulinc.com/doap>

²⁵<http://doapstore.org/about.php>



Figure 5.4: doap :store : Annuaire et interface de visualisation de projets logiciels modélisés avec DOAP

tème similaire entre les différents outils producteurs d'annotations RDF et le médiateur. La principale différence avec PTSW se situe dans le fait qu'au lieu de constituer une liste des documents RDF à partir de ces signalements, les notifications permettent directement la mise à jour de l'entrepôt avec l'ajout en temps-réel de ces documents au sein de celui-ci. Le signalement (et les actions qui s'en suivent au niveau de l'entrepôt) ne sont en outre pas limités à la création de documents, mais s'adaptent également aux actions de modification et de suppression. Un scénario classique de signalement et d'indexation est ainsi le suivant :

- un utilisateur crée, commente, supprime ou modifie un document, ce qui entraîne la création ou la modification du (ou des pour les wikis) graphe(s) d'annotation(s) associé(s) ;
- l'outil envoie alors un signal au médiateur pour l'informer de la création du ou des graphe(s) ainsi créé(s) ;
- le médiateur reçoit le signalement et indexe le ou les graphe(s) créé(s) au sein de l'entrepôt (en cas de création ou modification) ou bien les supprime (suppression du document). Dans le cas d'un commentaire, le graphe d'annotations du commentaire nouvellement créé est ajouté à l'entrepôt, celui-ci contenant des assertions RDF (`sioc:reply_of`) permettant de faire le lien avec le document d'origine ;

Alors que 3store dispose d'une API spécifique pour l'ajout de graphes d'annotations dans l'entrepôt, nous avons choisi de rendre notre approche plus générique et de ne pas dépendre d'une API propre à l'entrepôt utilisé. Nous nous sommes ainsi basés sur SPARUL, langage de mise à jour de données RDF, et son protocole HTTP associé. Alors que SPARQL permet d'interroger les données RDF d'un entrepôt, SPARUL propose leur mise à jour via des principes similaires. Ainsi, nous avons implémenté une partie de la grammaire SPARUL au sein de 3store, en l'occurrence le support des clauses `LOAD <graph>`, `DROP <graph>`

qui permettent respectivement l'ajout et la suppression d'un graphe d'annotations au sein de l'entrepôt²⁶. Signalons que si SPARUL n'est passé que récemment au statut de *Member Submission* au W3C²⁷, la proposition initiale nous a permis d'utiliser ses principes assez tôt²⁸.

Ainsi, les interactions au sein de notre architecture entre les différents outils et l'entrepôt se font d'une part avec SPARQL pour l'interrogation de données et d'autre part SPARUL pour leur mise à jour et suppression, en utilisant dans les deux cas les protocoles HTTP associés par l'intermédiaire du point d'accès SPARQL/SPARUL de l'entrepôt. De ce fait, n'importe quel entrepôt supportant SPARQL et SPARUL via HTTP peut être utilisé dans notre système²⁹, l'architecture étant ainsi comme nous l'avons évoquée complètement indépendante des outils mais reposant uniquement sur un ensemble de langages et protocoles standardisés. Cette couche d'abstraction nous permet donc au final d'avoir un système complètement indépendant de l'outil utilisé pour le stockage de données comme le montre la figure qui suit (Figure 5.5, page 196) et peut se généraliser à tout écosystème sémantique d'Entreprise 2.0 alors composé :

- d'adaptateurs qui informent le médiateur de la présence de nouveaux graphes d'annotations dans une optique de stockage (SPARUL) ;
- des services externes qui viennent utiliser ces annotations dans un objectif de requêtes, navigation ou visualisation (SPARQL).

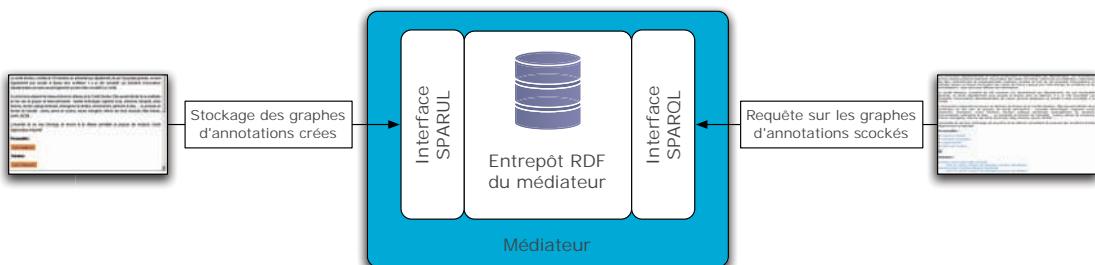


Figure 5.5: Protocoles d'abstraction au-dessus de l'entrepôt de données du médiateur

5.2 ENRICHISSEMENT DES FONCTIONNALITÉS DES WIKIS

5.2.1 Utilisation de macros sémantiques pour l'utilisation d'annotations

Principes des macros sémantiques au sein d'UfoWiki

Dans le chapitre précédent, nous avons présenté les différentes fonctionnalités offertes par UfoWiki en termes de publication de données et de peuplement d'ontologies (Section

²⁶Ces modifications ont récemment été intégrées à 3store. – <http://threestore.svn.sourceforge.net/viewvc/threestore?view=rev&revision=8>

²⁷<http://www.w3.org/Submission/2008/SUBM-SPARQL-Update-20080715/>

²⁸<http://jena.hpl.hp.com/~afs/SPARQL-Update.html>

²⁹Ceux-ci sont de plus en plus nombreux, comme Virtuoso ou ARC2.

4.2.2, page 154). Si celles-ci permettent d'assurer le maintien d'un ensemble de graphes d'annotations RDF, elles n'offrent pas directement la possibilité d'en bénéficier. Or, il est évident que ces annotations ont un rôle à jour en termes d'enrichissement des fonctionnalités offertes par les wikis. Afin d'en tirer profit et ce de la manière la plus transparente qui soit pour les utilisateurs, nous avons réfléchi à la mise en place d'un système de macros sémantiques, permettant d'intégrer dynamiquement au sein des pages les réponses à différentes requêtes SPARQL. Un exemple relativement simple de macro peut être une fonction listant l'ensemble des associations connues au sein d'un wiki, via l'identification des différentes instances de `foafplus:Association` créées. Plus complexe, on peut imaginer une macro qui liste les différents acteurs d'un domaine donné localisés dans une région particulière. Ce système de macros n'est pas une originalité propre à UfoWiki puisque d'autres outils, notamment Semantic MediaWiki, proposent déjà un système similaire duquel nous nous sommes inspirés³⁰. Comme nous allons le voir, les différences se situent principalement dans la manière dont nous combinons plusieurs niveaux d'annotations.

À partir du moment où nous disposons d'un grand nombre de graphes d'annotations RDF, un des avantages de ces macros est de permettre la résolution de requêtes complexes sans que l'utilisateur ne soit confronté ni à la complexité des annotations ni aux patrons SPARQL associés. Nos macros reposent en effet sur l'utilisation d'une syntaxe très simple, *i.e.* `[onto|fonction|param1, ..., paramn]`, où `fonction` correspond à l'identifiant de la macro appelée et `param1, ..., paramn` identifient une liste de paramètres optionnels. Chaque identifiant est associé à une fonction PHP (définie par l'administrateur³¹) qui va exécuter une ou plusieurs requêtes SPARQL sur l'entrepôt (via les protocoles définis auparavant) et formater les résultats obtenus en fragments de documents (X)HTML. C'est ici une des principales différences avec Semantic MediaWiki, nos macros étant définies sous forme de fonctions là où SMW utilise une syntaxe particulière de requêtes au sein des pages wikis (Section 4.2.1, page 151). Si cette souplesse permet à tout utilisateur de définir ses propres requêtes, la syntaxe utilisée se serait sans doute révélée trop complexe dans un contexte d'utilisateurs non-technophiles. Alors que notre approche de macros se concentre sur les wikis, on peut noter la récente proposition de SPARQLScript et des *templates* associés afin de généraliser cette notion de requêtes SPARQL intégrées au sein de pages Web [Nowack, 2008] ainsi que l'extension Firefox Kalpana³² proposée par [Ankolekar et Vrandecic, 2008].

Ces macros sont interprétées au moment du chargement de la page (via un parseur d'expression régulière qui identifie leur éventuelle présence) et les résultats sont immédiatement disponibles à l'affichage (Figure 5.6, page 198). La fraîcheur des données stockées (grâce au système de signalement présenté auparavant) combinée à ce système de macros interprétées offre donc un moyen efficace de profiter en quasi-temps réel d'annotations sémantiques distribuées au sein d'un système d'information. De plus, les instances et annotations étant créées et maintenues de manière collaborative (selon la philosophie wiki), nous tirons profit des principes d'intelligence collective et d'architecture de participation en termes de valeur

³⁰http://semantic-mediawiki.org/wiki/Help:Inline_queries

³¹Ceux-ci sont pour le moment les seuls à pouvoir définir de nouvelles macros puisqu'elles requièrent un accès au code source de l'application, n'ayant pu définir d'interface Web pour assurer leur gestion.

³²<http://www.anupriya-ankolekar.info/kalpana/>

ajoutée de ces macros et d'accès pertinent à l'information.

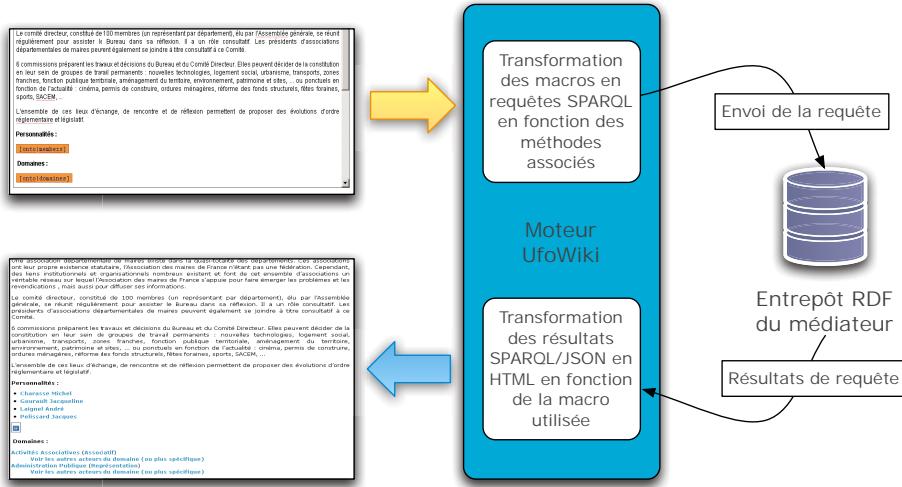


Figure 5.6: Processus d'interprétation des macros au sein d'UfoWiki

Prise en compte du contexte

Comme nous l'avons évoqué dans le chapitre précédent, UfoWiki permet la mise en place de plusieurs wikis, les annotations produites étant ensuite mutualisées au sein d'un entrepôt de données global. Ainsi, il est nécessaire de prendre en compte le wiki à partir duquel les macros sont interprétées pour s'assurer de la qualité des résultats. En effet, une macro identifiant un ensemble d'associations, si elle ne prend pas en compte le wiki depuis lequel elle est exécutée, conduira au même résultat si elle est interprétée depuis le wiki utilisé pour la gestion des partenariats que depuis HPédia, l'utilisateur s'attendant certainement à ce que la liste obtenue corresponde aux entreprises recensées dans ce wiki particulier. Il est donc nécessaire de prendre en compte le contexte de production de la macro pour s'assurer de la pertinence des résultats.

Cette contextualisation est rendue possible en limitant les graphes d'annotations utilisés pour résoudre la requête à ceux produits par ce même wiki. En pratique, nous utilisons l'extension `embedsKnowledge` que nous avons présenté précédemment (Section 4.2.3, page 156) qui trouve ici tout son intérêt en termes d'exploitation d'annotations sémantiques. Combinée au niveau des requêtes SPARQL avec l'utilisation de graphes nommés et de la propriété `has_container` de SIOC, elle nous permet de limiter les requêtes aux graphes produits à partir d'un wiki particulier. La requête SPARQL suivante (Listing 5.2, page 199) exemplifie cette combinaison en identifiant ici les associations créées à partir d'un wiki particulier (que nous associons ici à l'URI `athena:wiki_8`). Celle-ci va identifier les graphes (`?data`) associés aux pages (ici représentées par un nœud anonyme avec `[]`) contenues dans le wiki `athena:wiki_8`, et récupérer au sein de ces graphes (et uniquement de ceux-ci) les instances (`?asso`) de la classe `foafplus:Association`.

```

SELECT ?asso
WHERE {
  GRAPH ?data {
    ?asso rdf:type foafplus:Association .
  } .
  [] :embedsKnowledge ?data ;
    sioc:has_container athena:wiki_8 .
}

```

Listing 5.2: Restriction d'une requête SPARQL aux graphes produits par un wiki donné

Alors que les approches classiques d'utilisation de graphes nommés pour restreindre le contexte d'interrogation de triplets RDF se basent généralement sur l'utilisation d'une simple propriété associée à ces graphes (par exemple `dct:creator` pour en identifier l'auteur) notre proposition va plus loin en permettant d'exploiter un ensemble d'informations supplémentaires au sujet de la page ayant conduit à la production de ce graphe d'annotations métier. Si l'exemple précédent utilise simplement les propriétés associées au contenu de la page en question, on peut imaginer de la même manière utiliser des informations au sujet de son auteur ainsi que d'autres métadonnées documentaires associées, pour par exemple identifier les données produites depuis des pages éditées par un utilisateur particulier sur une période donnée. Plus généralement, cette contextualisation des annotations métier en fonction des annotations socio-structurelles nous semble un point important à prendre en compte si l'on souhaite utiliser avec pertinence ces deux types d'annotations au sein d'applications du Web Sémantique.

Utilisation combinée de données métier de d'annotations socio-structurelles

La plupart des requêtes associées à notre système de macros sont en réalité plus complexes que le précédent exemple. Ce dernier montre en effet une requête qui permet d'identifier l'URI des différentes associations, mais ni les pages wikis associées, ni le titre de ces pages. Pour ce faire, nous tirons à nouveau profit de cette articulation que nous avons proposé entre annotations socio-structurelles et données métier. Nous étendons tout d'abord la requête avec l'utilisation de `foaf:primaryTopic` qui nous permet d'identifier la page principale associée à chaque instance. Cette page identifiée, nous pouvons récupérer différentes informations comme son titre, son URL, son auteur, etc. Ainsi, une macro listant l'ensemble des associations connues et les pages associées sera définie par la fonction PHP et la requête SPARQL qui suivent (Listing 5.3, page 200) et sera simplement appelée par l'utilisateur avec [onto | associations]³³.

Le résultat d'une telle macro est visible sur la figure qui-suiv, chaque lien de la liste à puces renvoyant vers la page wiki en question, la liste permettant également de créer une nouvelle page associée à la classe en question (Figure 5.7, page 201). Une telle macro va donc permettre d'identifier simplement ces associations sans obliger l'utilisateur à parcour-

³³La fonction `sparql_query` présente dans cet exemple fait partie d'une API mise en place en interne pour interagir simplement avec l'entrepôt depuis des applications PHP.

```

function associations() {}
$query = "
SELECT ?page ?title
WHERE {
  GRAPH ?data {
    ?asso rdf:type foafplus:Association .
  } .
  ?page :embedsKnowledge ?data ;
    foaf:primaryTopic ?asso;
    dc:title ?title ;
    rdf:type sioct:WikiArticle ;
    sioc:has_container athena:wiki_8 .
}
";
$res = sparql_query($query);
foreach ($res as $r) {
  $page = $r[ page ][ value ];
  $title = $r[ title ][ value ];
  $n = "<li><a href= \"$page \"$title</a></li>";
}
return "<ul>$li</ul>";
}

```

Listing 5.3: Fonction PHP et requête SPARQL associées à une macro UfoWiki

rir les 173 pages d'HPédia, le bénéfice de l'utilisation des annotations et de manière plus large le passage de documents à des données formalisées étant alors non négligeable. De plus, il est important de noter que les résultats sont immédiatement mis à jour. Ainsi, dès qu'un utilisateur va créer une page conduisant à la création d'une nouvelle instance de `foafplus:Association`, celle-ci sera listée via la macro en question.

Au final, si ces requêtes peuvent s'avérer complexes, notamment puisqu'elles couplent plusieurs niveaux d'annotations et d'ontologies tout en articulant ceux-ci via l'utilisation de relations entre graphes en non plus seulement entre triplets, cette complexité est masquée à l'utilisateur final. Celui-ci ne se soucie lors de l'édition d'une page que de l'appel de la macro – via une syntaxe relativement simplifiée – et bénéficie immédiatement d'un rendu de celle-ci. Signalons également que si les exemples précédents font état de requêtes prenant en compte uniquement le type d'instances à récupérer, les macros peuvent se concentrer non pas sur le type mais sur différentes propriétés, comme par exemple les domaines d'activités ou de la localisation des acteurs, représentés au travers de la notion de rôle (Section 3.2.4, page 109), les deux pouvant être combinés. Par exemple, une macro comme `[onto|acteurs|domaine,localité]` va identifier des acteurs selon le domaine et la localisation associés à leur(s) rôle(s) tout en prenant également en compte les principes d'inférence RDFS afin d'identifier simultanément toutes les instances des sous-classes de `foaf:Agent` correspondant à ces critères.

Les fiches associations :

- AFCCRE - Association française du conseil des communes et régions d'Europe
- AFG - Association Française du Gaz
- AGEFIPH
- Alliance pour la planète
- AMF - Association des Maires de France
- CABA - Continental Automated Buildings Association
- France Nature Environnement
- Maires de grandes villes
- Planète éolienne
- Sauvons la recherche
- SIGEIF - Syndicat intercommunal pour le gaz et l'électricité en Ile de France
- Sortir du nucléaire



- » Type: pages wiki internes
- » Ajouter une page wiki pré-typée

Figure 5.7: Résultat d'une macro sémantique listant l'ensemble des associations recensées au sein d'un wiki

5.2.2 Contextualisation des macros pour augmenter le potentiel de veille

Alors que les processus présentés auparavant permettent de définir des macros prenant en compte le wiki à partir duquel elles sont initiées, il nous a semblé intéressant d'aller plus loin en offrant la possibilité de contextualiser les macros de manière plus fine, *i.e.* non plus au niveau du wiki mais de la page en question, ou plutôt de l'instance associée. Ceci permet ainsi d'identifier simplement des informations concernant celle-ci comme par exemple, pour une organisation, l'ensemble des acteurs d'un même secteur ou ses différents membres, sans référence explicite à l'instance puisque celle-ci est automatiquement identifiée par la macro elle-même.

Pour ce faire, une première étape consiste en l'identification de l'instance associée à chaque page wiki et nous tirons ici à nouveau profit de l'utilisation de `foaf:primaryTopic` introduite dans UfoWiki. Une fois cette propriété identifiée, il est aisément d'adapter les différentes requêtes et de produire les macros associées. Par exemple, la requête qui suit (Listing 5.4, page 202) associée à une macro `[onto|members]` et exécutée depuis une page relative à une organisation donnée permet d'identifier ses différents membres. La variable `$self` est ici définie pour identifier l'organisation en cours et est remplacée par l'URI associée au moment de l'exécution de la requête. Le résultat d'une telle macro est en outre visible sur la figure suivante (Figure 5.8, page 202). On peut ainsi considérer ces macros contextualisées et permettant d'afficher au sein d'une page des informations sur les concepts en relation avec le concept en cours comme des *rétroliens sémantiques*. Alors que les rétroliens classiques identifient simplement les pages ayant des liens vers la page en cours, ces macros permettent de lister (et de catégoriser selon différentes propriétés) les concepts en relation avec le concept associé à la page en cours, les affichant à un endroit approprié sur la page wiki. Cette requête met également en avant l'intérêt d'utiliser des URIs communes entre différentes pages wikis, processus facilité par UfoWiki avec l'autocomplétion associée aux annotations. En effet, cette requête utilise un patron `<$self> foaf:member ?uri`, où `$self` représente l'URI

de l'association en question. Alors que ces différents triplets sont produits à partir de différentes pages wikis, l'utilisation d'URIs communes permet d'identifier à partir de chaque graphe d'annotations qu'il s'agit bien de la même organisation (Figure 5.9, page 203).

```
select distinct ?page ?name
where {
  graph ?g {
    <$self> foaf:member ?uri .
    ?uri rdfs:label ?name .
  }
  ?page :embedsKnowledge ?g ;
    foaf:primaryTopic ?uri ;
    sioc:has_container athena:wiki_8 .
} ORDER BY ASC(?name)
```

Listing 5.4: Requête SPARQL avec contextualisation des macros

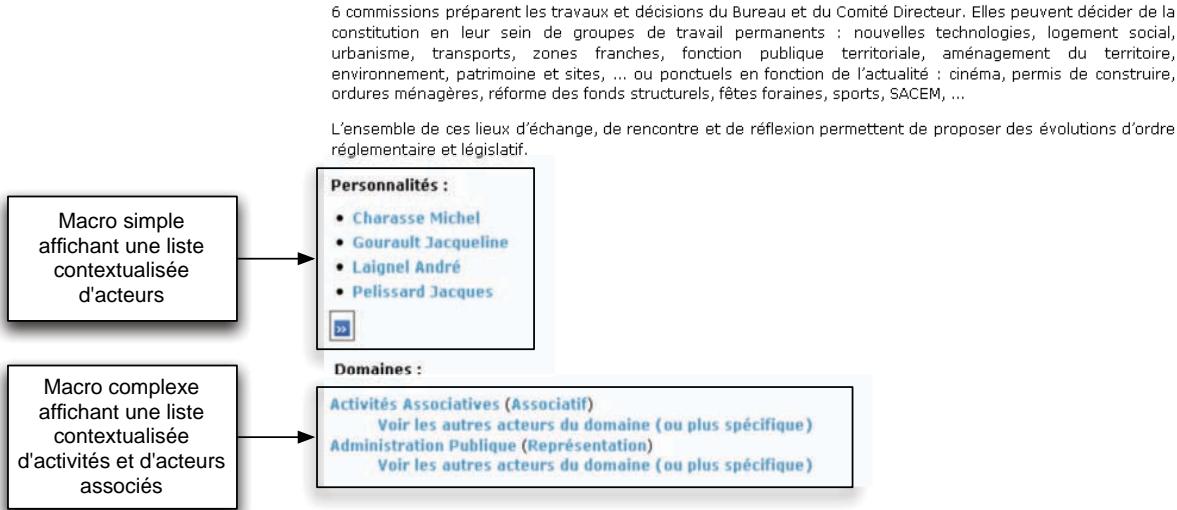


Figure 5.8: Résultat d'une macro contextualisée

Nous pouvons également voir sur la copie d'écran associée à cette macro (Figure 5.8, page 202) le résultat d'une seconde macro plus complexe, qui va lister les différentes activités d'un acteur (domaines et métiers associés à son rôle) mais également pour chaque domaine d'activité l'ensemble des acteurs associés, ainsi que ceux évoluant dans des domaines plus spécifiques. Nous tirons ici bénéfice du choix de SKOS pour représenter les rôles et les domaines, avec la possibilité d'identifier simplement pour un domaine donné l'ensemble de ses sous-domaines grâce à la transitivité de la propriété `skos:broaderTransitive`³⁴,

³⁴<http://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html#broaderTransitive>

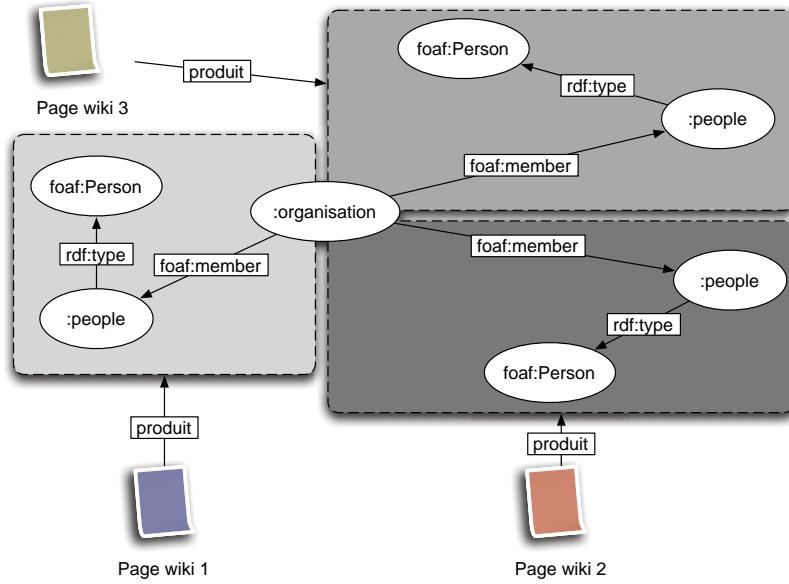


Figure 5.9: URIs partagées entre graphes d'annotations

super-propriété de `skos:broader` utilisée dans notre contexte³⁵. Nous bénéficions à nouveau dans ce cas d'utilisation d'annotations issues de différents wikis : un premier wiki pour la gestion de ces taxonomies de domaines et un second pour établir les relations entre domaines et acteurs de la même manière que précédemment, l'utilisation d'URIs communes permettant de faire le lien entre les différents graphes RDF. Un problème auquel nous avons cependant été confronté et qui reste ouvert est le besoin d'identifier jusqu'où il est nécessaire de considérer un sous-domaine comme pertinent par rapport au domaine initial. Nous reviendrons sur cette problématique en fin de chapitre (Section 5.4.3, page 216).

5.2.3 Interfaces avancées de visualisation et *mash-ups* sémantiques

Navigation à facettes

Si l'on considère le Web Sémantique comme un graphe de relations typées entre nœuds, il est possible d'accéder à chaque nœud selon plusieurs critères, *i.e.* selon les propriétés qui lui sont assignées et les valeurs correspondantes. Par exemple, en considérant les ontologies utilisées dans notre contexte pour définir la notion d'acteur, chaque acteur peut-être considéré selon son type (`rdf:type`), sa localisation (`geonames:locatedIn`), ses rôles (`role:hasRole` et objets associés), ses membres (`foaf:member`), etc. Or, les macros présentées précédemment ne permettent pas de prendre en compte toute la richesse de ce graphe de manière simple et extensible, *i.e.* de visualiser les instances d'ontologies de domaine dynamiquement selon plusieurs points de vue. Ces macros sont en effet généralement conçues

³⁵Nous avons en effet intégré au sein de notre entrepôt les possibilités de raisonnement associés à la transitivité de cette propriété définie comme instance de `owl:TransitiveProperty`.

pour visualiser une unique propriété (le nom de chaque instance, via `rdfs:label`) et requièrent comme nous l'avons montré des requêtes plus complexes pour afficher d'autres propriétés, par exemple les domaines d'activité.

Afin de prendre en compte cette richesse en termes de navigation, nous avons appliqué les principes de navigation à facettes à nos graphes d'annotations [Yee *et al.*, 2003]. Ce procédé, qui permet de proposer différents points de vue pour aborder un objet donné, dans notre cas une instance d'ontologie de domaine, nous semble le plus adapté pour visualiser ces données multidimensionnelles représentées en RDF. Nous avons ainsi défini différentes facettes à prendre en compte pour visualiser chaque acteur (instance de `foaf:Agent` créée depuis UfoWiki) à partir des différentes ontologies utilisées pour représenter celui-ci. L'alignement entre ontologies et facettes a ici été effectué manuellement à partir du moment où nous avons une connaissance précise des modèles utilisés. Dans un contexte où les données reposent sur des modèles plus hétérogènes (par exemple contrôlés par les utilisateurs), la détection automatique de facettes telle que proposée par [Oren *et al.*, 2006] peut alors se révéler nécessaire. Comme nous pouvons le voir avec le schéma qui suit (Figure 5.10, page 204), nous ne nous limitons pas à des facettes qui sont liées à des propriétés directement associées à chaque instance (par exemple `rdf:type`) mais explorons certains objets associés, ici les rôles pour identifier des facettes pertinentes.

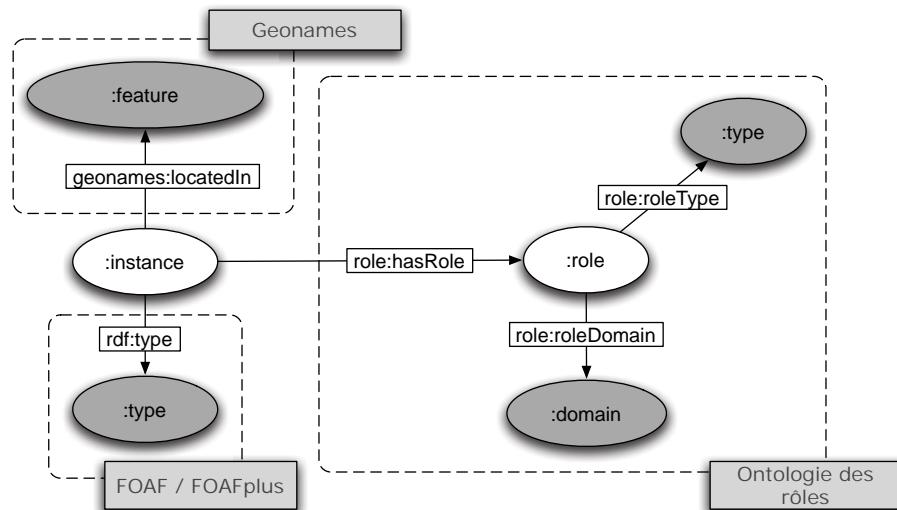


Figure 5.10: Sélection de facettes à partir de différentes ontologies (les facettes sélectionnées sont en gris foncé)

Une fois ces facettes définies, nous utilisons Exhibit [Huynh *et al.*, 2007] pour offrir une visualisation dynamique de ces différents acteurs. Les valeurs proposées pour chaque facettes ainsi que la liste des différents acteurs sont en outre calculées au moment de l'affichage de la page (toujours via SPARQL) ce qui permet d'avoir aussi bien une liste d'acteurs que des facettes de navigation constamment à jour (Figure 5.11, page 205).

BLOCS • CARTE

173 Items

Trier par : libellés, puis par... • Grouper selon le tri

- AFCCRE - Association française du conseil des communes et régions d'Europe (lien)**
libellé: AFCCRE - Association française du conseil des communes et régions d'Europe
type: Association
URI: <http://athena.der.ed...3%A9gions%20d'Europe>
domaine:
- AFG - Association Française du Gaz (lien)**
libellé: AFG - Association Française du Gaz
type: Association
URI: <http://athena.der.ed...C3%A7aise%20du%20Gaz>
domaine: Gaz
- AGEFIPH (lien)**
libellé: AGEFIPH
type: Association
URI: <http://athena.der.ed...=hpedia/item#AGEFIPH>
domaine:
loc: Bagneux, France
lat{lng}: 48.8,2.3
- Alga Technologies (lien)**
libellé: Alga Technologies

Type

- ASSOCIATION
- Centre de Recherche
- Entreprise
- Institution
- Organisation

Domain

- Activités Associatives
- Activités financières
- Administration

Localisation

- (missing this field)
- Paris
- Aix en Provence, France

Figure 5.11: Visualisation à facettes d'un wiki avec Exhibit

Si les facettes proposées ici sont relatives à des données métier, il est possible d'utiliser ces mêmes principes pour visualiser un ensemble de documents (toujours représentés en RDF) en prenant en compte leurs différentes métadonnées socio-structurelles. Bien que nous n'ayons pas mis en place une telle interface au sein de notre plate-forme d'entreprise, nous avons expérimenté cette approche à travers l'application de microblogging SMOB présentée auparavant. Ici, les facettes sont alignées avec différentes propriétés associées à chaque instance de `sioct:Item`. L'auteur (`foaf:maker`), les sujets associés (`sioct:topic`) et la date de création (`dct:created`) de chaque instance de `sioct:Item` sont ainsi prises en compte pour définir les facettes comme le montre la figure qui-suiv (Figure 5.12, page 205).

Semantic MicroBlog - demo timeline

BLOCS • LIGNE DE TEMPS • CARTE

24 MicroBlogPost

Trier par : date, puis par... • Grouper selon le tri

- Tuukka Hastrup**
testing new client
2008-03-14T18:09:21+00:00
- Alexandre Passant**
test
2008-03-14T03:26:19-07:00
- Alexandre Passant**
one more test with new config file
2008-03-14T03:22:34-07:00

Date

- 2008-02-05
- 2008-02-06
- 2008-03-14

Name

- Alexandre Passant
- Tuukka Hastrup

Figure 5.12: Interface à facettes pour visualiser des données SIOC avec SMOB

Mash-ups sémantiques

Toujours dans cette optique de visualisation avancée de données RDF, nous avons mis en place un système de *mash-ups* sémantiques au sein de notre système. Comme nous l'avons détaillé précédemment, UfoWiki intègre automatiquement en son sein des informations RDF proposées par Geonames lorsqu'un acteur est identifié comme associé à une zone géographique particulière. De ce fait, pour chaque acteur lié (via `geonames:locatedIn`) à une instance de `geonames:Feature`, nous disposons de différentes informations relatives à cette zone, en particulier ses coordonnées géographiques. De ce fait, nous avons pu mettre en place, toujours en utilisant Exhibit, un système de géolocalisation permettant de visualiser les acteurs représentés au sein d'HPédia (Figure 5.13, page 206). L'utilisation couplée de ce système de géolocalisation et de navigation à facettes proposée par l'interface permet de plus de contextualiser cette cartographie selon différents critères. Il est ainsi possible d'étudier la situation géographique des acteurs d'un domaine donnée, par exemple localiser l'ensemble des entreprises actives dans le domaine des Energies Marines.

Comme nous l'avons déjà mentionné, cet aspect de réutilisation des savoirs externes en entreprise nous semble particulièrement intéressant, encore plus dans ce contexte de *mash-ups* construit à partir d'outils relativement simples et à forte composante sociale comme les wikis. À cet égard, nous pensons que les applications du Web Sémantique peuvent être jugées non seulement sur leur valeur à utiliser et proposer des données RDF(S)/OWL, mais également sur leur capacités à tirer profit d'autres données représentées selon les mêmes modèles. On peut alors considérer la notion de *mash-up* comme faisant partie intégrante du Web Sémantique, à partir du moment où différentes applications produisent des données interconnectées, permettant ensuite de s'abstraire de ces applications source pour les consommer via d'autres outils.

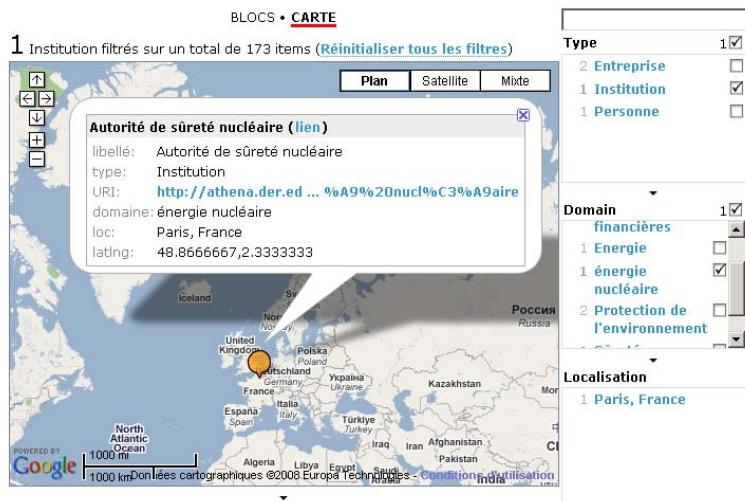


Figure 5.13: Géolocalisation d'un ensemble d'acteurs avec Exhibit et Geonames

Cette notion de *mash-up* sémantique est également prise en compte dans les macros que

nous avons présentées auparavant. Nous avons ainsi défini une macro contextualisée qui va récupérer pour une organisation donnée l'ensemble des membres associés pour à nouveau visualiser ceux-ci sur une carte avec le même principe que précédemment (Figure 5.14, page 207). Il est ainsi possible de visualiser simplement le réseau (filiales, etc.) d'un acteur donné, toujours à partir d'informations issues de différentes pages wiki.



Figure 5.14: Géolocalisation au sein d'une macro contextualisée

Dans l'ensemble, l'utilisation de ce type d'interfaces (aussi bien en termes de navigation à facettes que de *mash-ups* sémantiques) nous permet de retrouver une hypothèse que nous avons défendue dans le premier chapitre de ce mémoire, à savoir l'utilisation d'interfaces à la mode Web 2.0 pour visualiser des données modélisées selon les principes du Web Sémantique (Section 1.3.2, page 43). De plus, ces interfaces intuitives permettent de masquer la complexité du système sous-jacent (agrégation de graphes, intégration de données externes, inférence RDFS) en proposant une navigation relativement simple et originale. Enfin, si ces interfaces sont essentiellement textuelles, cartographie mise à part, il nous semble intéressant de manière plus large de considérer des approches graphiques pour la visualisation de graphes [Herman *et al.*, 2000] dans le contexte du Web Sémantique.

5.3 INTEROPÉRABILITÉ ENTRE APPLICATIONS VIA LES ANNOTATIONS

5.3.1 Intégration des contenus des blogs au sein des wikis

Bien que les exemples présentés auparavant se limitent à l'utilisation de données produites au sein des wikis, nous avons vu dans le chapitre précédent que la plate-forme de blogs permettait également la production d'annotations sémantiques. Il nous semble de ce fait important de prendre celles-ci en compte. Si les annotations produites par les wikis combinent métadonnées socio-structurelles et données métier, les annotations issues de blogs se réfèrent quant à elles uniquement à l'aspect socio-structurel. Or, comme nous l'avons vu, MOAT nous permet de représenter au sein de ces annotations les liens qui existent entre

documents (ici les billets de blog) et des instances d'ontologies de domaine, plus particulièrement dans notre contexte les instances créées depuis les wikis. Ainsi, si les deux outils restent distincts en termes de pratiques et d'usages, ils permettent tous deux la production d'annotations RDF qui d'une manière ou d'une autre font référence à des instances d'ontologies de domaine :

- d'une part, la propriété `foaf:primaryTopic` est utilisée au sein des wikis pour identifier le concept principal (instance d'ontologie de domaine) associé à une page wiki ;
- d'autre part, le lien entre document et instance est représenté au niveau des blogs via l'utilisation de la *Tag Ontology* et de MOAT et en particulier d'une instance de `tag:RestrictedTagging` couplée à la propriété `moat:tagMeaning`.

Ces instances étant identifiées par une même URI, qui fait alors office de jointure, les différents graphes d'annotations sont interconnectés au travers de celles-ci, permettant ainsi de faire le rapprochement entre les deux outils (Figure 5.15, page 208)³⁶. Ce lien entre outils via les annotations sémantiques offre ainsi une interopérabilité accrue entre applications.

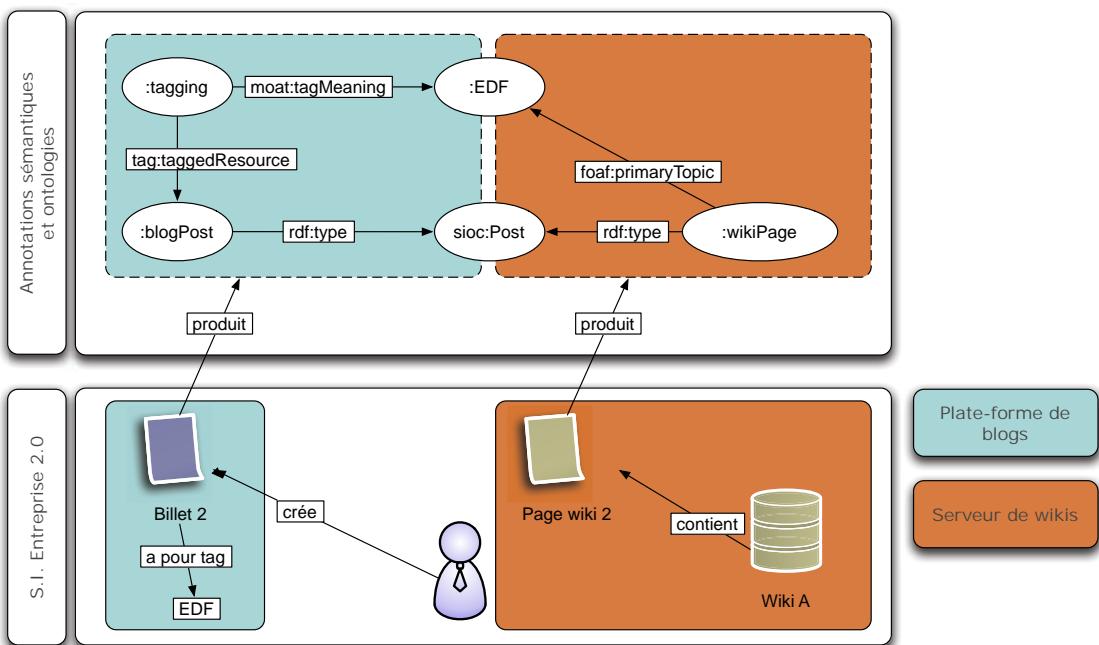


Figure 5.15: Interopérabilité entre applications via l'utilisation d'annotations sémantiques

Afin de tirer profit de cette interopérabilité, nous avons mis en place une macro listant les derniers billets de blogs se rapportant au concept associé à une page wiki donnée. Ceci permet à nouveau d'augmenter l'expérience utilisateur en proposant la mise en commun contextualisée d'informations produites depuis différents outils. D'un point de vue pratique, ce processus repose également sur une simple macro [onto|billet] traduite en la

³⁶On remarque également sur cette figure que le lien entre les deux outils est rendu possible avec SIOC.

requête SPARQL qui suit (comme précédemment, `$self` est utilisé pour représenter l’instance en cours) (Listing 5.5, page 209). Le rendu est proposé sous forme de liste à puces avec des liens vers les billets d’origine. Il est ici important de noter que les différents billets de blog peuvent avoir été tagués originellement avec différents tags, l’utilisation de MOAT nous permettant de considérer l’instance associée et non plus le simple mot-clé pour ainsi établir ce lien nécessaire entre ressources.

```
SELECT ?post ?title
WHERE {
  ?post a sioct:BlogPost ;
    dc:title ?title .
  [] a tag:RestrictedTagging ;
    moat:tagMeaning <$self> .
  tag:taggedResource ?post .
}
```

Listing 5.5: Requête SPARQL pour identifier des billets annotés avec un concept particulier

Nous avons évoqué plus tôt dans cette thèse la notion d’*object-centered sociality* (Section 1.2.3, page 41) en insistant notamment sur un problème particulier, à savoir que celle-ci était généralement fragmentée, les contenus au sujet d’objets similaires étant répartis entre différentes applications (blogs, wikis, forums, réseaux sociaux) (Section 2.2.1, page 62). Comme nous venons de l’exemplifier, l’utilisation d’URIs communes pour référencer les sujets abordés via ces différents outils permet d’interconnecter ceux-ci autour des objets qu’ils évoquent, et de parvenir à cette notion de forums virtuels mise en avant par SIOC. Plus particulièrement, c’est l’utilisation de MOAT couplée à SIOC qui nous permet de considérer ces forums comme des espaces de discussions liés à des sujets communs, réduisant ainsi cette fragmentation. Bien entendu, cette intégration repose également sur la mise à disposition d’URIs communes pour identifier ces objets, et c’est en ce sens que le project *Linking Open Data* nous paraît plus que pertinent puisqu’offrant un nombre important d’URIs de référence (notamment via DBpedia) qui peuvent être utilisées dans ce contexte.

5.3.2 Indexation de flux RSS guidée par les annotations

Alors que les possibilités offertes par les outils précédents permettent de visualiser des informations existantes, il nous semble également souhaitable d’exploiter ces informations pour en produire de nouvelles. Ici, nous ne faisons pas référence à des principes d’inférences reposant sur des axiomes définis au sein de nos ontologies mais à la manière d’utiliser les différentes annotations produites dans un but d’indexation de flux RSS.

Comme nous l’avons explicité plus tôt dans ce mémoire, un très grand nombre de sources de données RSS sont mises à disposition de l’utilisateur au sein de la plate-forme d’agrégation. Pour éviter la surcharge d’information, les utilisateurs ont la possibilité de souscrire précisément à certains flux via une interface dédiée (Section 2.1.2, page 53). En contrepartie, cette pré-sélection conduit parfois à passer à côté de certaines informations importantes, dans la mesure où elles peuvent provenir de flux auquel l’utilisateur n’est volontairement pas abonné (Section 2.2, page 62). Ainsi, plutôt que de considérer les éléments de flux RSS

en fonction de leur source (*i.e.* le flux auquel ils appartiennent) il nous semble pertinent de considérer ces flux en fonction des thématiques qu'ils abordent. On envisage ainsi des flux RSS virtuels organisés par thèmes, ces thèmes étant représentés via des instances d'ontologies de la même manière que nous proposons l'indexation sémantique de billets de blogs.

Afin de passer de ces éléments organisés par source à des éléments organisés par thématiques, nous avons mis en place un prototype d'indexation automatique de flux RSS en fonction des différentes instances qui ont été créées via les wikis (limitées aux sous-classes de `foaf:Agent`) . Le processus d'indexation est assez sommaire et repose sur l'utilisation des liens entre tags et instances définis avec MOAT ainsi que sur les labels associés à ces instances via les wikis afin de construire une table de correspondance entre chaînes de caractères et URIs de concepts, comme le tableau suivant l'exemplifie (Tableau 5.1, page 210). Cette correspondance est ensuite utilisée pour annoter les flux RSS à l'aide d'expressions régulières. Si les critères d'indexation sont satisfaits, une relation `sioc:topic` est créée entre l'élément de flux, représenté avec SIOC comme nous l'avons vu auparavant (Section 4.1.2, page 140), et le concept associé.

URI	Terme associé
<code>athena:Areva</code>	areva
<code>athena:EDF</code>	EDF
<code>athena:EDF</code>	E.D.F
<code>athena:EDF</code>	électricité de france

Tableau 5.1: Associations entre URIs et termes contrôlées par les utilisateurs

Bien entendu, il s'agit ici d'un processus d'indexation très sommaire notamment en termes de rapport signal-bruit, loin d'avoir des résultats aussi pertinents qu'un système comme KIM [Kiryakov *et al.*, 2004]. Si nous pouvons via MOAT repérer plusieurs patrons d'indexation associés à la même instance (par exemple EDF et `électricité de france` pour l'entreprise EDF), ce qui permet à nouveau d'identifier des sujets communs malgré des termes distincts, la gestion de l'ambiguité n'est pas assurée. Les termes associés à plusieurs URIs ne sont ainsi pas pris en compte dans le processus d'indexation. De plus, la lemmatisation n'est pas assurée, ce qui n'est ici pas un problème particulier puisque nous indexons essentiellement des entités nommées mais qui le serait si l'on souhaitait étendre cette indexation aux domaines d'activité par exemple. Pour permettre une indexation plus précise, il est de plus nécessaire de prendre en compte d'autres éléments pour définir plus finement nos schémas d'indexation et les processus associés. Des techniques comme l'exploration contextuelle [Desclés, 1997] sont sans doute une solution pertinente à cette problématique et c'est par exemple sur ces principes que repose le moteur d'indexation EXCOM [Djioua *et al.*, 2006]. Signalons ici également les récents services d'indexation sémantique proposées par Reuters OpenCalais³⁷ ou Zemanta³⁸ qui permettent l'indexation de données non-structurées par des concepts définis au sein du projet *Linking Open Data*. Celles-ci

³⁷<http://www.opencalais.com/>

³⁸<http://www.zemanta.com/api/>

peuvent s'avérer particulièrement utiles dans ce contexte d'écosystème sémantique pour l'Entreprise 2.0, à partir du moment où des données publiques sont utilisées en interne comme nous le faisons avec Geonames.

Malgré la simplicité de notre approche, plus que l'indexation en soi, c'est le processus de création de l'index qui nous semble pertinent et novateur. Cet index est en effet défini non pas en vase clos (avec les écueils que cela peut entraîner, tout comme pour la création d'ontologies (Section 4, page 137)) mais est directement dérivé des comportements des utilisateurs sur la plate-forme : création de pages wiki et d'instances correspondantes, association de tags à ces instances, etc. L'aspect interaction homme-machine, non pas en termes d'interface, mais en termes d'utilisation de données produites par l'utilisateur pour enrichir les applications, prend ici tout son sens comme le rappelle [Gandon et Giboin, 2008] : *dans ces nouvelles approches un point important est que l'utilisateur n'est plus simplement le commanditaire d'un service pour lequel il fournit des entrées et attend des sorties, mais devient une ressource computationnelle de l'architecture logicielle.*

Enfin, puisque nous avons évoqué le bénéfice de l'approche pour l'utilisateur, signalons qu'une macro contextualisée [onto|news] permet de récupérer les dernières nouvelles au sujet d'un acteur donné pour les afficher au sein d'une page wiki, renforçant à nouveau la découverte d'informations pertinentes autour d'un *objet* particulier.

5.3.3 Projection de connaissances pour l'aide à la veille technologique

En complément de cette indexation de flux entrants et afin de proposer une manière supplémentaire d'augmenter l'expérience utilisateur en termes de navigation, nous avons mis en place un système de projection des connaissances. Cette idée, proposée par exemple dans Magpie [Domingue et Dzbor, 2004] permet à l'utilisateur d'identifier au sein de pages Web différentes entités connues par le système dans l'objectif d'accéder à un ensemble d'informations au sujet de celles-ci. Le processus mis en place pour l'identification d'instances au sein des pages est similaire à celui de l'indexation de flux RSS détaillé en amont (et donc sujet aux mêmes critiques) et pour chaque instance identifiée la page est enrichie d'un lien vers les résultats du moteur de recherche associés à cette instance (nous décrirons ce moteur dans la section suivante (Section 5.4, page 212)). La copie d'écran suivante montre ici l'identification du concept CEA au sein d'un billet de blog (Figure 5.16, page 211).

L'[European Conference on Research Infrastructure 2008 \(ou ECRI\)](#), co-organisée par la Commission européenne, se tiendra du 9 au 11 décembre au palais des Congrès de Versailles. Environ 500 acteurs européens concernés par la problématique des infrastructures de recherche notamment dans les secteurs des nanosciences, de la sécurité, de la nanoélectronique, des sciences de la vie et de l'environnement seront réunis à cette occasion. Les deux premiers jours seront consacrés à des exposés, forums et discussions. Le dernier jour sera réservé à la visite d'une dizaine de grandes infrastructures de recherche franciliennes : le Synchrotron Soleil, Supratech de l'IN2P3 et Laserlab de LULI, à l'École Polytechnique, ou le NeuroSpin du  à Saclay. [Conseil Régional IDF- 08/12/2008 Programme de l'ECRI](#)

Identification d'une instance connue

Figure 5.16: Projection de connaissances sur des contenus internes

Outre la possibilité de découvrir des informations supplémentaires au sujet de ces instances, un autre aspect intéressant de ce processus en termes de veille est le signalement d'entités non référencées dans le système, ou plutôt leur non-signalement. Ceci permet par exemple d'avoir une idée des acteurs émergents dans un domaine, au sens où ils n'ont pas été identifiées par la projection puisque non référencées dans la base de connaissance du médiateur.

5.4 RECHERCHE SÉMANTIQUE POUR L'ENTREPRISE 2.0

5.4.1 Recherche d'information et Web Sémantique

Avant de décrire en détail les principes du moteur de recherche implémenté au sein de notre écosystème, définissons ce que nous entendons par recherche d'information sur le Web Sémantique. Traditionnellement, les moteurs de recherche tels que nous les utilisons aujourd'hui se basent sur une recherche documentaire, *i.e.* proposent des documents répondant à un terme de recherche saisi par l'utilisateur. Ceux-ci utilisent généralement des structures d'index inversés (type TF-IDF) permettant d'identifier les documents contenant le terme recherché par l'utilisateur [Salton et McGill, 1986], couplés avec des stratégies plus ou moins fines d'optimisation comme le PageRank de Google [Brin et Page, 1998]. Deux choses nous semblent importantes dans la manière dont ces moteurs fonctionnent :

- la recherche se fait par terme et donc en prenant en compte uniquement une notion syntaxique. Les problèmes d'ambiguïté et d'hétérogénéité sont donc relativement fréquents et peuvent nuire à la qualité des résultats, tout comme pour les systèmes de recherche d'information par tags (Section 2.2.3, page 63);
- le résultat obtenu est un ensemble de documents (textuels ou multimédia) qu'il est nécessaire de parcourir pour avoir une vue synthétique du concept recherché, imposant un travail de recherche supplémentaire à l'utilisateur.

Ainsi, de la même manière que la publication de contenus sur le Web Sémantique consiste à considérer non plus uniquement les documents mais un ensemble de données auxquelles ils font référence, nous pensons que la recherche d'information doit elle aussi prendre en compte ces particularités. En exploitant annotations sémantiques et ontologies, on peut ainsi passer d'un paradigme de recherche centrée sur les documents à une recherche centralisée autour des objets représentés dans ces documents [Guha *et al.*, 2003]. Chaque objet peut en outre être considéré selon différents angles, comme nous l'avons vu en présentant l'utilisation d'interfaces à facettes pour la navigation. Ainsi, un moteur de recherche basé sur les technologies du Web Sémantique doit selon nous être capable :

- d'une part de permettre une recherche par concept et non plus par simple chaîne de caractère. On franchit ici le pas entre le terme de recherche et le concept associé (par exemple une instance d'ontologie) en passant de la syntaxe à la sémantique ;
- d'autre part de délivrer des résultats qui donnent à l'utilisateur une vue synthétique de ce concept, non plus simplement en termes de documents y faisant référence, mais en termes de propriétés et de relations avec d'autres concepts.

Des moteurs comme Yahoo ! SearchMonkey³⁹ [Mika, 2008] ou SWSE⁴⁰ [Harth *et al.*, 2007] prennent ainsi en compte ces aspects pour proposer des résultats de recherche synthétisant des informations provenant de différentes sources de données structurées (RDF au sens large pour SWSE, RDFa et microformats pour SearchMonkey). Notons que si SWSE est capable de prendre en compte n'importe quel modèle utilisé pour décrire ces objets, SearchMonkey se limite à l'interprétation restreinte de certains vocabulaires, parmi lesquels SIOC, comme nous l'avons évoqué auparavant (Section 3.1.6, page 101). Si ces deux moteurs se situent dans une optique de recherche centrée autour de concepts, optique qui nous semble la plus pertinente en termes de recherche d'information sur le Web Sémantique, d'autres outils s'orientent vers une recherche documentaire plus traditionnelle, à la différence près que les documents indexés sont des documents structurés. C'est le cas de certains moteurs que nous avons déjà évoqués dans cette thèse, à savoir Sindice [Tummarello *et al.*, 2007], Watson [d'Aquin *et al.*, 2008] ou Swoogle [Ding *et al.*, 2004], plutôt dédiés à la réalisation d'application utilisant des données structurées qu'à une navigation humaine.

5.4.2 Mise en place d'un moteur de recherche exploitant ontologies et annotations

Nous avons mis en pratique ces principes de recherche sémantique au sein d'Hermès en proposant un moteur de recherche associé à notre architecture de médiation et venant tirer profit des différentes ontologies et annotations présentes dans notre écosystème. Celui-ci permet de visualiser, pour un concept donné, un ensemble cohérent et synthétique d'informations à son sujet, avec des pointeurs vers les différents documents source ayant permis cette synthèse. Notre approche est ainsi une approche mixte entre les moteurs de recherche traditionnels qui délivrent des liens vers un ensemble de documents et les moteurs sémantiques comme SWSE qui délivrent des informations au sujet d'objets particuliers. Ce moteur respecte en outre les deux phases que nous avons mises en avant, à savoir (1) l'identification d'un concept particulier à partir d'un terme de recherche et (2) la mise à disposition d'une synthèse informationnelle au sujet de ce concept.

La première étape consiste ainsi à passer du terme de recherche (*e.g.* solaire) au concept associé (ici l'instance identifiée par l'URI `athena:EnergieSolaire`). Pour ce faire, notre stratégie se base sur l'utilisation des connaissances produites au sein de la plate-forme, tout comme nous l'avons fait pour l'indexation de flux RSS ou la projection de connaissances (Section 5.3, page 207). Pour un terme de recherche t , le moteur va ainsi identifier le concept C qui satisfait au moins un des critères suivants :

- le label (`rdfs:label`) du concept C est égal ou contient le terme t ;
- un tag associé à ce concept C (via MOAT et la notion de signification globale) est égal ou contient le terme t , *i.e.* il existe un tag égal ou contenant t et dont la signification globale est associée à C ;

Une fois le concept identifié (via son URI), la recherche va porter sur celui-ci et non plus sur le terme d'origine, le moteur se situant alors au niveau sémantique et non plus à un simple niveau syntaxique. Si plusieurs concepts sont identifiés, l'utilisateur se voit proposer

³⁹<http://developer.yahoo.com/searchmonkey/>

⁴⁰<http://swse.deri.org>

la liste correspondante afin de sélectionner lui-même le concept recherché et résoudre ainsi les problèmes d'ambiguïté (Figure 5.17, page 214).



Figure 5.17: Choix d'un concept à partir d'un terme de recherche

La seconde étape consiste ensuite en l'identification d'informations pertinentes au sujet de ce concept. Comme nous l'avons exposé au début de cette section, il nous semble important de ne pas uniquement proposer une liste de documents mais d'offrir un synthèse informationnelle à propos des différents attributs et propriétés de ce concept. Plus exactement, nous souhaitons proposer un juste milieu entre ces deux approches, en contextualisant les documents proposés en fonction des propriétés qui les lient (directement ou via les sujets abordés) au concept principal. Ainsi, notre système prend en compte l'ensemble des annotations RDF présentes dans la base de connaissances et faisant référence à ce concept pour proposer à l'utilisateur une page de résultats listant :

- l'ensemble des tags associés au concept, dans un but informatif permettant à l'utilisateur de prendre connaissance des différents mots-clés qui lui sont associés. Cette première étape repose sur l'utilisation de MOAT ;
- la page de référence associée au concept en question, en l'occurrence la page wiki principale issue du wiki HPédia dans le cas des acteurs. Nous reposons ici à la fois sur SIOC et embedsKnowledge (pour identifier qu'il s'agit bien d'une instance de `sioc:WikiArticle` appartenant au conteneur souhaité) et FOAF pour identifier qu'il s'agit de la page principale (avec `foaf:primaryTopic`) ;
- les pages faisant référence à des concepts en relation avec ce concept, toujours identifiées depuis HPédia. Pour une organisation, il peut ainsi s'agir des pages identifiant ses différents membres. La requête utilisée est présentée ci-après et combine ainsi SIOC et annotations métier (Listing 5.6, page 215) ;
- enfin, les différentes pages wiki, billets de blog et flux RSS annotés avec l'URI du concept en question, via l'utilisation de SIOC (`sioc:topic`) et MOAT pour les billets de blog. La recherche se faisant ici par concept, et non plus par mot-clé, cela nous permet de prendre en compte les problèmes initiaux d'hétérogénéité sémantique. En effet, les différents contenus annotés par le concept en question peuvent avoir originellement été tagués avec des mots-clés distincts. Notons également que le moteur fait ici la distinction entre les différents types de documents grâce à l'utilisation du module Types de SIOC au niveau des annotations sémantiques.

```

SELECT ?page ?title
WHERE {
  GRAPH ?data {
    { ?uri ?p <$self> } UNION { <$self> ?p ?uri }
  } .
  ?page :embedsKnowledge ?data ;
    foaf:primaryTopic ?uri;
    dct:title ?title ;
    rdf:type sioc:WikiArticle ;
    sioc:has_container athena:wiki_8 .
}
  
```

Listing 5.6: Identification de pages associées à un concept proche



Figure 5.18: Rendu du moteur de recherche sémantique au sein d'Hermès

À nouveau, l'application mise en place repose entièrement sur un ensemble de requêtes SPARQL utilisant différents graphes d'annotations et ontologies associées, sans confronter l'utilisateur à ces processus de parcours de graphes. De plus, un autre aspect mis en avant par notre interface est la possibilité de créer de nouveaux contenus à partir de celle-ci, notamment lorsqu'il n'en existe pas à ce sujet pour le wiki principal HPédia. Le moteur est ainsi utilisé dans une démarche d'incitation à la production de contenu permettant d'enrichir les connaissances globales au sein du système. Tout utilisateur venant consommer de l'information est donc invité à son tour à devenir acteur, suivant les principes classiques de collaboration sur le Web 2.0, couplés à nouveau à des principes de structuration de données liés au Web Sémantique.

Nous avons de plus défini différents points d'accès permettant d'arriver à ces pages de résultat. Si le premier est naturellement une zone de recherche plein-texte, nous avons vu

dans la section précédente que la projection des connaissances permettait également d'arriver sur la page de résultats pour un concept donné. Une autre manière d'accéder à ces résultats est également proposée dès lors qu'un billet de blog ou une page wiki est associée à un concept via MOAT. Dans ce cas, en plus d'indiquer simplement les tags associés à ce document, le système liste l'ensemble des concepts associés avec pour chacun d'entre eux un lien vers la page associée au sein du moteur de recherche (Figure 5.19, page 216). En termes d'utilisation, des analyses de fichiers de logs sur une période d'un mois nous ont indiqué une trentaine de visiteurs différents ayant accédé à ce moteur.

QUAND LA VILLE VEUT SE CHAUFFER À L'EAU

03/02/2009 - 07:59

De grands travaux de forage vont débuter dans le XIX ème arrondissement de Paris, sur le terrain du futur quartier Claude-Bernard. Pas question d'aller puiser du pétrole, mais... de l'eau chaude ! La Compagnie parisienne de chauffage urbain (CPCU) a décidé d'aller chercher dans le sous-sol l'eau chaude du « dogger », une couche géologique située à 1 700 m environ de profondeur.

En savoir plus :

- [Source](#)

Posté par Yrrien le 3 Février 2009 - 07:59 || Mot(s)-clé(s) : [énergie géothermique](#) | [géothermie](#) | [NRGY](#) | [Paris](#)
|| Concepts: [voir](#)

- [Géothermie](#)
- [Paris, France](#)

Concepts identifiés via MOAT
et lien vers le moteur de recherche

Figure 5.19: Accès au moteur de recherche via les concepts identifiés avec MOAT

5.4.3 Suggestion de concepts et de contenus proches

Implémentation au sein du moteur de recherche interne

En recensant les différents problèmes posés par les systèmes d'annotations à base de tags, nous avons mentionné l'absence d'organisation de ceux-ci comme étant un défaut majeur (Section 2.2.3, page 63). Nous avons en effet montré que cette absence rendait complexe la découverte de contenus proches (au sens des thématiques abordées), particulièrement dans un contexte où les niveaux d'expertise des utilisateurs étaient relativement hétérogènes. En contrepartie, nous avons évoqué la manière dont les processus proposés par MOAT permettaient de répondre à cette problématique. En effet, en passant de termes syntaxiques à des concepts clairement identifiés et interconnectés sur le Web Sémantique pour annoter les documents, il est possible de naviguer dans le graphe d'annotations centré autour de ce concept pour identifier des concepts proches et en conséquence les contenus associés. La figure qui suit montre par exemple comment deux billets de blog peuvent être connectés à partir du moment où l'un a été associé à l'URI `geonames:2988507` et le second à `geonames:3017382`, ces deux URIs étant liées par une relation `geonames:locatedIn` fournie par Geonames (Figure 5.20, page 217). Il est également possible à partir de relations de ce type de lier directement les billets avec la propriété `sioc:related_to` en utilisant la règle d'inférence qui suit (Listing 5.7, page 217).

Comme pour la mise en place d'interfaces à facettes, il est important de considérer que nous sommes en présence de modèles de graphes, au sens où plusieurs types de relations

peuvent exister entre concepts, permettant d'envisager différentes manières de suggérer des concepts (et des contenus) proches. En effet, il nous semble pertinent de proposer des suggestions différentes selon le type d'objet étudié (personne, domaine, zone géographique, etc.) puisque les propriétés qu'il possède sont généralement différentes. C'est l'un des avantages majeurs d'une structure ontologique riche par rapport à une simple taxonomie et c'est une des raisons qui nous a motivé à mettre MOAT en place, notamment par rapport à des approches plus classiques d'organisations taxonomiques de tags où une unique relation hiérarchique est proposée.

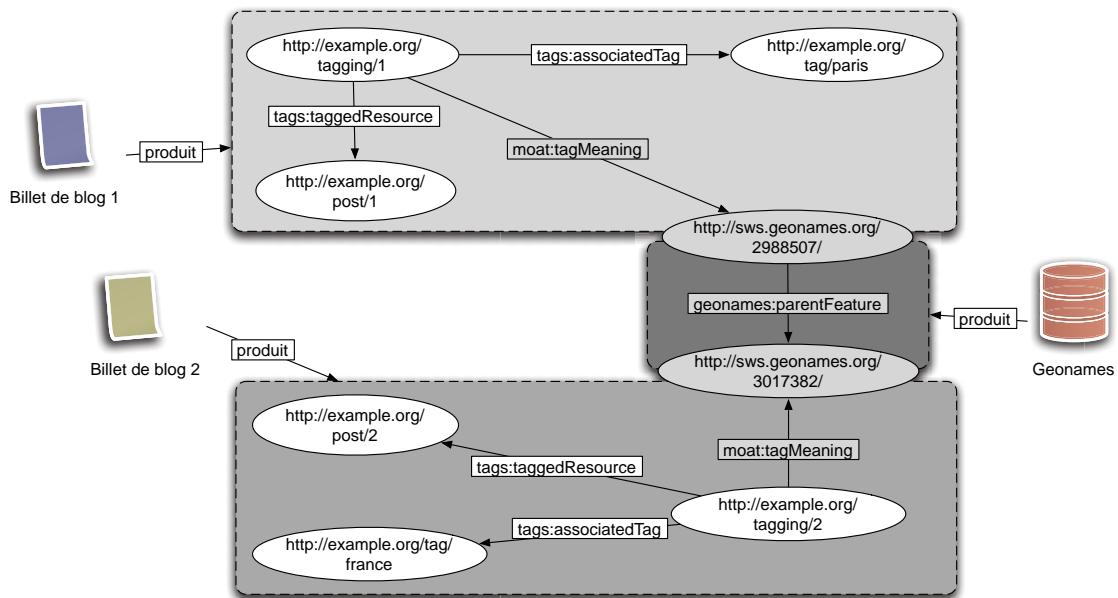


Figure 5.20: Identification de contenus proches via des relations entre concepts associés

```
{
    xxx a sioc:Post .
    [] moat:tagMeaning iii ;
        tag:taggedResource xxx .
    yyy a sioc:Post .
    [] moat:tagMeaning jjj ;
        tag:taggedResource yyy .
    iii rrr jjj .
} => {
    xxx sioc:related_to yyy .
}
```

Listing 5.7: Règle d'inférence pour identifier deux contenus proches en utilisant MOAT, SIOC et des relations entre URIs

Ainsi, nous avons mis en place au sein de notre outil un système de suggestion de concepts proches en définissant pour différentes classes de nos ontologies des règles d'inférence, comme par exemple :

- une première règle, s'appliquant à toute instance de `role:Domain` et permettant d'identifier comme liés⁴¹ des domaines considérés comme plus spécifiques dans la hiérarchies de domaines. Cette règle fait appel à la propriété `skos:broaderTransitive` de manière à considérer tous les concepts plus spécifiques (Listing 5.8, page 218). Comme on peut le voir dans l'exemple qui suit, appliquée ici au concept d'énergie solaire, des concepts relativement pointus tels que `cellule silicium monocristallin` sont suggérés (Figure 5.21, page 218) afin de prendre en compte le problème des différents niveaux d'expertise évoqué plus tôt dans ce mémoire. Nous sommes ici confrontés, en utilisant ces principes de transitivité, au même problème que dans les macros présentées auparavant : à partir du moment où cette inférence est mise en place, il n'y a plus possibilité d'identifier la distance originelle qui sépare les concepts, à moins de parcourir l'ensemble des relations non-inférées `skos:broader`. Les extensions SPARQL proposant des requêtes par chemin nous paraissent ainsi particulièrement utiles dans ce contexte, pour par exemple limiter la suggestion à des concepts situés à un maximum de N relations `skos:broader` par rapport au concept d'origine ;

```
{
  xxx a role:Domain .
  yyy a role:Domain .
  xxx skos:broaderTransitive yyy .
} => {
  xxx :related yyy .
}
```

Listing 5.8: Règle d'inférence basée sur SKOS pour l'identification de concepts proches

Concepts plus spécifiques que "énergie solaire"

- `cellule silicium monocristallin`
- `cellule à colorant`
- `cellule à concentration`
- `cellule CIGS`
- `cellule GaAs`
- `cellule multijonction III - V`
- `cellule organique`
- `cellule photovoltaïque`

Figure 5.21: Identification des domaines plus spécifiques qu'`énergie solaire`

- une seconde, s'appliquant à toute instance de `foaf:Agent` et permettant d'identifier d'autres agents partageant un domaine d'activité en commun avec cette instance (Listing 5.9, page 219). La figure qui suit représente l'application de cette règle pour l'instance associée à Gaz de France (Figure 5.22, page 219).

⁴¹On utilise ici une propriété `:related` pour représenter ce lien.

```
{
  xxx a foaf:Agent ;
    role:hasRole [
      role:hasDomain ddd .
    ] .
  yyy a foaf:Agent ;
    role:hasRole [
      role:hasDomain ddd .
    ] .
} => {
  xxx :related yyy .
}
```

Listing 5.9: Règle d'inférence pour l'identification de concepts proches à partir de relations entre domaines

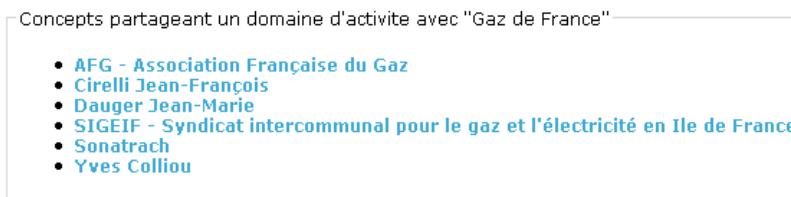


Figure 5.22: Identification d'acteurs proches de Gaz de France selon une règle prédéfinie

On peut voir ici que la complexité des règles varie selon l'usage mais surtout que celles-ci permettent à nouveau de prendre en compte les caractéristiques de chaque objet pour suggérer des concepts proches, en combinant ontologies et annotations. Bien entendu, plusieurs règles peuvent être définies pour une même classe, soit puisque définies explicitement, soit en appliquant les principes d'inférence RDFS. Par exemple, la seconde règle s'appliquera à toute instance de `foafplus:Company`, sous-classe de `foaf:Agent`. D'un point de vue pratique, ces règles d'inférence sont définies au sein de notre système en tant que requêtes SPARQL (modélisées à partir des règles d'inférence N3 détaillées précédemment), et sont appliquées dès lors qu'une instance de la classe donnée est identifiée. Chaque concept proche ainsi identifié est également proposé sous forme de lien hypertexte vers la page associée au sein du moteur de recherche, afin d'accéder aux documents correspondants.

Applications de principes similaires sur le Web

Alors que les règles définies précédemment pour la suggestion de concepts proches sont appliquées essentiellement sur des données internes à l'entreprise, il nous a semblé pertinent de voir de quelle manière cette idée pouvait s'appliquer sur le Web, notamment en prenant en compte le nombre croissant de données RDF disponibles via le projet *Linking Open Data*. Nous avons ainsi mis en place deux expérimentations basées sur ces principes.

Tout d'abord, nous avons implémenté ce principe de suggestion au sein de LODr (Sec-

tion 4.3.2, page 178) en mettant en place deux règles relativement simples :

- une première identifiant tous les concepts en relation directe avec le concept en cours, *i.e.* identifiés comme sujet ou objet d'une relation avec ce concept ;
- une seconde identifiant tous les concepts pour lesquels une propriété donné est partagée avec le concept en cours (*i.e.* à la même valeur).

On peut voir dans l'exemple qui suit que XSLT (dbpedia:XSL_Transformations) est suggéré lors de la visualisation de contenus annotés SPARQL (dbpedia:SPARQL) puisque ces deux concepts partagent la même valeur pour la propriété `skos:subject` au sein de DBpedia, en l'occurrence `dbpedia:Category:World_Wide_Web_Consortium_standards` (Figure 5.23, page 220). L'outil propose en plus une définition du concept visualisé (correspondant à la valeur de la propriété `dc:description`) ainsi qu'une liste d'éléments annotés via ce concept, en utilisant également différentes facettes, notamment pour identifier la source associée à chaque élément. Un aspect intéressant est ainsi la possibilité de visualiser au sein d'une même interface des contenus issus de systèmes distincts (Flickr, SlideShare, etc.) mais au final représentés avec les mêmes modèles (SIOC, MOAT, etc.) et interconnectés via l'utilisation d'URIs communes pour représenter leurs thématiques.

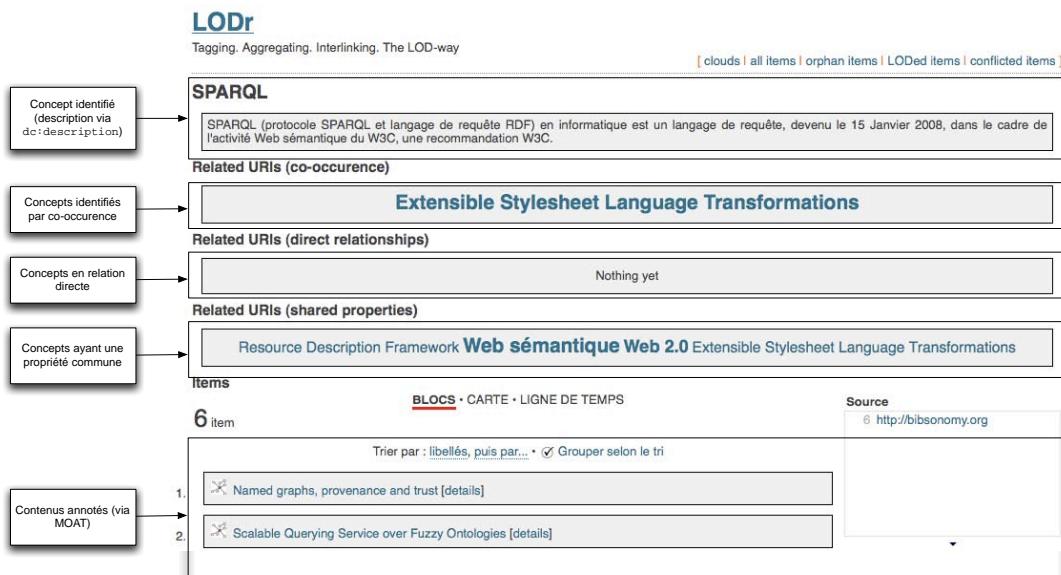


Figure 5.23: Suggestion de concepts proches au sein de LODr

Nous avons également mis en place un processus similaire au sein d'un système de recommandations musicales basées sur DBpedia [Passant et Raimond, 2008]. À partir d'un artiste sélectionné, le système propose différentes listes organisées par critères de similarité (même genre, même label, etc.) en analysant les propriétés associées à chaque instance. Ces critères sont ici définis par avance et il nous est ici apparu en analysant 400 instances d'artistes et de groupes représentés au sein de DBpedia que sur les vingt propriétés les plus couramment associées à ceux-ci, un certain nombre n'étaient pas pertinentes dans cette op-

tique de recommandation (par exemple la propriété dbpedia:wikiPageUsesTemplate) (Annexe C, page 233). Si nous n'avons pas exploré plus loin cette problématique, cela nous semble un challenge important à prendre en compte dans la mesure où un grand nombre de données structurées et interconnectées sont maintenant disponibles sur le Web. Des notions de distance sémantique appliquées à celles-ci pourraient sans doute apporter plus de pertinence à de tels moteurs de recommandation [Rada *et al.*, 1989].

About 'Beastie Boys'



The Beastie Boys are an American hip hop group from New York City consisting of Michael "Mike D" Diamond, Adam "MCA" Yauch, and Adam "Ad-Rock" Horovitz. Since around the time of the Hello Nasty album, the DJ for the group has been Michael "Mix Master Mike" Schwartz, who was featured in the song "Three MCs and One DJ". They started out as a hardcore band in 1979, and adopted their signature look of yellow shirts and black pants with Red Fight and Flava Flav. They switched to hip hop with the release of their debut solo album Licensed to Ill (1986), which enjoyed international critical acclaim and commercial success. The group is well-known for their eclecticism, popular and flippant attitude toward interviews and interviewers, obscure cultural references and kitschy lyrics, and performing in outlandish matching suits. They are one of the longest-lived hip hop acts and continue to enjoy commercial and critical success in 2008, more than 20 years after the release of their debut album. On September 27, 2007 they were nominated for induction into the Rock and Roll Hall of Fame.

- Browse 'Beastie Boys' on last.fm

Interested in artists :

having a similar topic ?

- Capitol Records artists (97 bands/artists including Alyn,Bob Seger,Bonepony, ...)
- Beastie Boys (12 bands/artists including Alfredo Ortiz,Amyer Smith,Awesomie, I Fuckin' Shot That!, ...)
- Grammy Award winners (1973 bands/artists including "Weird Al" Yankovic,112 (band),A Flock of Seagulls, ...)
- New York hardcore punk bands/artists including +-(band),Amen,1000 Islands,1500 Mockingbird Lane, ...)
- White hip-hop artists (76 bands/artists including 2 Live Crew,3rd Bass,TL & Esoteric, ...)
- Def Jam Recordings artists (63 bands/artists including 112 (band),Ashanti (singer),Beverly Sigel, ...)
- Rapcore groups (37 bands/artists including Attican Underground,Back-On,Black Market Hero, ...)
- Songwriting teams (37 bands/artists including Absolute (production team),Ashford & Simpson,Ateleje tragi, ...)
- Jewish hip hop groups (6 bands/artists including 2 Live Jews,Blood of Abraham,Hadag Nahash, ...)
- American hip hop groups (442 bands/artists including 10,000 Cadillacs,116 Clique,13 & God, ...)
- New York hardcore punk groups (43 bands/artists including 108 (band),Agnostic Front,Alova for Enemies, ...)
- Musical groups established in 1979 (78 bands/artists including 45 Grave,A II Z,Amsterdam Baroque Orchestra & Choir, ...)

playing a similar genre ?

- Funk (1161 bands/artists including (Not Just) Kneee Deep,100 Days,100 Nights,12" Collection and More, ...)
- Rock music (12656 bands/artists including "Weird Al" Yankovic,05 EP,(Reach Up for The) Sunrise, ...)
- Hip hop music (4102 bands/artists including "Weird Al" Yankovic,\$100 Bill 'Y'all,(Always Be My) Sunshine, ...)
- Jazz (3331 bands/artists including 58 Miles Featuring Stella by Starlight,'Nuff Said!,Round About Midnight, ...)

from the same label ?

- Def Jam Recordings (233 bands/artists including (You Gotta) Fight for Your Right (To Party!),10 (LL Cool J album),4, 3, 2, 1, ...)
- Grand Royal (44 bands/artists including 456132015,Adam Horovitz,Adam Yauch, ...)

Figure 5.24: Système de recommandations musicales basées sur DBpedia

CONCLUSION

Dans ce chapitre, nous avons présenté différents services et outils permettant de tirer profit d'annotations sémantiques dans un contexte d'Entreprise 2.0. Nous avons, dans un premier temps, argumenté de la nécessité d'un entrepôt de données dans ce contexte et défini un ensemble de protocoles permettant son intégration au sein d'un système dynamique de production d'annotations. Nous avons ensuite présenté différentes approches permettant de bénéficier de ces annotations parmi lesquelles un système de macros sémantiques intégrées au sein d'UfoWiki, l'utilisation d'interfaces à facettes pour la visualisation d'instances d'ontologies de domaine et le mise en place de *mash-ups* sémantiques. Nous avons ensuite détaillé le fonctionnement d'un moteur de recherche sémantique associé à cette architecture ainsi que l'utilisation de règles permettant l'extension de requêtes via le parcours des graphes d'annotations associés aux données métier.

Pour chaque outil, nous avons fait en sorte que ces interfaces soient les plus intuitives possible pour l'utilisateur, pour qui la mécanique sous-jacente (*i.e.* l'utilisation de technologies du Web Sémantique) importe peu. Pour reprendre les propos de David Karger évoquant

les interfaces de navigation pour le Web Sémantique lors d'un panel à SWUI2006⁴², "*whatever is in the cake, what people see is the candle!*". À cet égard, un point qu'il est selon nous important de retenir de ce chapitre est qu'à partir du moment où l'on dispose sur le Web Sémantique de données accessibles et interopérables, il est possible d'imaginer une multitude d'interfaces de navigation et de recherche associées à ces données, la valeur de celles-ci étant alors inestimable dans ce contexte.

⁴²3^{ème} workshop Semantic Web User Interaction – <http://swui.semanticweb.org/swui06/>

Conclusion générale

RETOUR SUR LES IMPACTS DE LA THÈSE

Réponses aux problématiques initiales

En introduction de ce mémoire, nous avons résumé la problématique scientifique motivant nos travaux de la manière suivante : *Comment combiner Web Sémantique et Web 2.0 afin de tirer profit d'interactions sociales issues d'outils du Web 2.0 pour la représentation et l'exploitation de connaissances formalisées selon les principes du Web Sémantique ?* Ainsi, nous avons montré tout au long de cette thèse de quelle manière nous envisagions cette complémentarité à la fois en termes de modèles (Section 3, page 83) et d'applications pour la production (Section 5, page 185) puis pour l'exploitation (Section 4, page 137) de telles connaissances. Avant de résigner globalement nos travaux et d'y apporter un regard critique, revenons sur les trois axes de recherche majeurs définis au début de ce mémoire.

La modélisation des métadonnées socio-structurelles associées aux outils Web 2.0

Nos travaux se sont ici concentrés sur deux modèles principaux, SIOC et MOAT, permettant de prendre en compte pour le premier la modélisation des activités des communautés en ligne (et des documents ainsi créés) et pour le second des aspects particuliers des tags et des actions de *tagging* en faisant notamment le lien avec des ontologies de domaine venant en support des folksonomies. Nous avons ici fait en sorte que ces modèles soient suffisamment génériques pour pouvoir s'intégrer au sein de différents types de communautés, ces deux ontologies étant en outre publiées sur le Web. De plus, afin de faciliter les processus d'annotation sémantique associés à ces modèles, de nombreux outils ont été mis en place, aussi bien au sein de notre écosystème d'Entreprise 2.0 que sur le Web.

La représentation de connaissances termino-ontologiques et le peuplement d'ontologies de domaine à partir d'outils Web 2.0

Ici, nous nous sommes principalement intéressés à l'utilisation de wikis sémantiques pour le peuplement d'ontologies, avec la mise en place d'UfoWiki, système combinant principes d'édition wiki et représentation des connaissances selon les technologies du Web Sémantique. Un point important dans cette approche est le rôle actif de l'utilisateur final, dans une approche collaborative et ouverte de constitution de bases de connaissances termino-ontologiques qui masque à l'utilisateur la complexité des technologies associées. Pour mener à bien cette étape, nous avons également mis en place différentes ontologies de domaine,

processus qui nous a permis d'identifier ce qui nous semble être un ensemble de bonnes pratiques en termes de représentation des connaissances pour l'Entreprise 2.0, en étendant notamment des modèles couramment acceptés sur le Web.

L'exploitation de graphes d'annotations sémantiques pour l'interopérabilité, la mise en commun et la recherche d'informations

En conséquence des deux points précédents, nous avons identifié différentes manières d'exploiter des graphes d'annotations sémantiques, qu'il s'agisse de graphes représentant des métadonnées socio-structurelles ou associés à des données métier. Nous avons ainsi mis en place différentes interfaces permettant d'exploiter ces annotations, de simples macros sémantiques à des interfaces à facettes plus complexes permettant à l'utilisateur de s'approprier la nature multidimensionnelle des objets manipulés pour les visualiser selon différents points de vue. Nous avons ici également vu comment l'utilisation d'URIs communes entre applications, facilitée via MOAT, permettait une interopérabilité accrue entre outils distincts et facilitait également la recherche d'information associée. Enfin, nous avons vu de quelle manière différentes sources de données pouvaient être combinées au sein de *mash-ups* sémantiques articulant données internes et externes.

Vision globale de notre recherche

Plus généralement, nos travaux et le contexte d'entreprise dans lequel nous nous situons nous ont permis de définir la méthodologie *SemSLATES*, vision où les technologies du Web Sémantique viennent en support d'écosystèmes d'Entreprise 2.0 pour répondre à leurs limites via une architecture de médiation entre différents composants logiciels. Ainsi, alors que de nombreuses entreprises migrent à l'heure actuelle vers des solutions d'Entreprise 2.0 où comme le veulent les principes d'*écologie de l'information*, l'utilisateur a un rôle aussi – voire plus – important que les applications elles-mêmes, il nous a paru pertinent d'aller plus loin dans cette vision et de montrer en quoi ces solutions pouvaient tirer profit de technologies du Web Sémantique.

De plus, bien que cette thèse s'intitule *Technologies du Web Sémantique pour l'Entreprise 2.0*, nous avons fait en sorte que l'ensemble de nos recherches puisse être appliqué de manière plus large sur le Web. Pour exemple, SIOC est aujourd'hui utilisé dans de nombreuses applications du Web Sémantique à composante sociale, dépassant ainsi le cadre d'utilisation d'entreprise que nous avons étudié dans cette thèse. L'impact de MOAT est quant à lui plus restreint mais la vision qu'il défend est aujourd'hui mise en valeur par d'autres initiatives du même type dans lesquelles il s'intègre. D'autre part, certaines de nos réflexions et réalisations logicielles s'intègrent de manière plus large dans cette vision de convergence entre Web Sémantique et Web 2.0.

Si, comme nous avons pu le voir dans ce manuscrit, nous ne sommes pas les seuls à défendre ces théories de convergence, il nous semble intéressant d'avoir montré selon différents axes que les représentations formelles proposées par les technologies du Web Sémantique (via RDF(S)/OWL et SPARQL) ne s'opposaient pas, et bien au contraire, à la souplesse des services Web 2.0 et aux notions de participations sociales qui en découlent. Plus particulièrement, un point qui nous semble pertinent dans notre approche est la prise en compte de ces notions de participations sociales selon deux axes complémentaires :

- d'une part en représentant à l'aide de modèles formels les interactions sociales qui peuvent exister au sein de différentes communautés Web 2.0 ;
- d'autre part en permettant l'émergence de bases de connaissances ouvertes et évolutives, dirigées par les utilisateurs finals.

Nous nous inscrivons ainsi dans une vision du Web Sémantique (et du Web de manière plus générale) où l'utilisateur est au centre d'un système global d'information que l'on peut voir comme une chaîne *humain-machine-humain* et où la composante sociale est aussi importante que la machine elle-même. Ainsi, nous pouvons reprendre une de nos précédentes figures et l'adapter comme suit pour définir cette vision d'un Web où les interactions sociales permettent la production d'un ensemble de données interoperables et interconnectées pour le bénéfice de l'utilisateur final (Figure 5.25, page 225). En conséquence, gardons à l'esprit que la réussite d'une telle complémentarité entre Web Sémantique et Web 2.0 repose sur des critères sociaux de participation et d'échange et que l'aspect social est à considérer autant que les formalismes de représentation de données.

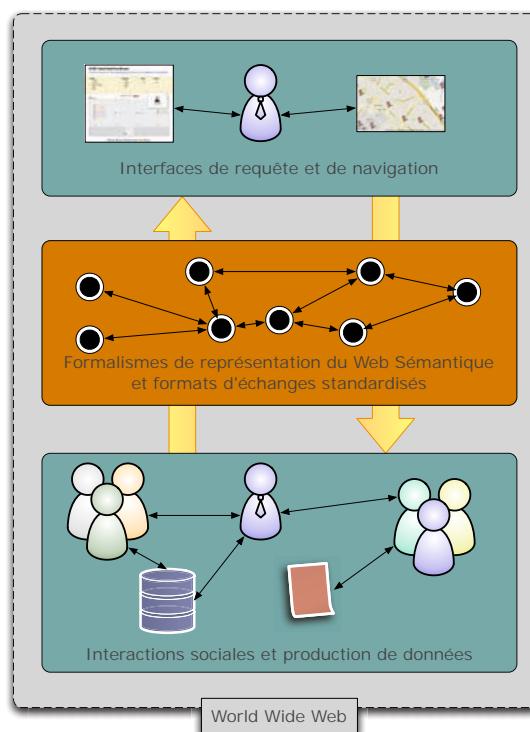


Figure 5.25: Vision du Web axée sur une convergence *humain-machine-humain*

Regard critique sur nos travaux

Il nous semble également important dans cette conclusion de porter un regard critique sur nos travaux, revenant sur certains aspects qui auraient pu être améliorés. Tout d'abord, il aurait sans doute été intéressant de proposer d'autres services et interfaces exploitant les

données RDF produites dans notre contexte d’expérimentation en entreprise. La création des macros sémantiques est en effet pour le moment limitée aux administrateurs et le processus de navigation par facettes, s’il permet de visualiser une partie des connaissances produites selon différents points de vue, ne prend pas en compte toute la richesse et la complexité des graphes d’annotations. Des interfaces graphiques avancées auraient sans doute été pertinentes mais nous pouvons penser que celles-ci pourront se greffer par la suite à l’écosystème mis en place, à partir du moment où les *données* sont disponibles et représentées selon des vocabulaires et formalismes connus. Notons également que nous avons du faire face à certaines limitations techniques qui ont retardé certains développements et contraint certains autres à rester au statut de prototype.

De manière plus globale, on peut nous reprocher d’avoir axé nos recherches sur la définition de modèles et de processus nécessitant une intervention utilisateur plutôt que sur la mise en place des traitements automatiques, notamment pour nos travaux autour des liens entre tags et ontologies avec MOAT. Nous avons cependant vu que ces deux approches ne s’opposaient pas et pouvaient ainsi être combinées. De plus, il nous semble que la définition de vocabulaires de référence est nécessaire pour mener à bien la vision du Web Sémantique et c’est ce en quoi nous avons essayé de contribuer avec MOAT.

Enfin en termes d’évaluation, il aurait sans doute été pertinent de confronter notre système de wikis sémantiques à d’autres outils du même type, à la fois en termes de prise en main et de qualité des annotations produites.

PERSPECTIVES ET RÉFLEXIONS

Perspectives de recherche

À l’issu de cette thèse, différentes perspectives de recherche venant dans la continuité des travaux présentés dans ce mémoire s’offrent à nous. Nous souhaitons ainsi axer une partie de nos travaux futurs autour des problématiques suivantes :

- l’extension de la méthodologie *SemSLATES* afin de prendre en compte d’autres sources de données dynamiques dans cette perspective d’intégration d’informations sociales en entreprise. Il peut ici s’agir de données provenant aussi bien du poste de travail (dans la lignée du *Semantic Desktop*) que de flux d’information issus de services de microblogging, terminaux mobiles et autres senseurs favorisant l’ubiquité numérique ;
- la protection des données personnelles et l’évaluation du degré de confiance des sources d’information sur le Web 2.0, pour lesquelles les technologies du Web Sémantique nous semblent offrir un cadre approprié. Comme nous l’avons évoqué, l’ouverture des données sociales ne nous semble pas aller à l’encontre de ces principes mais nous permet au contraire d’envisager des possibilités avancées de contrôle des informations personnelles, en couplant représentation unifiée de données structurées, politiques d’accès et langages de règles ;
- la mise en place de méthodes avancées permettant l’exploitation de données RDF de plus en plus nombreuses sur le Web, notamment via le projet *Linking Open Data*. Plus particulièrement, il nous semble intéressant de réfléchir à la manière dont celles-ci peuvent être utilisées avec pertinence en termes de navigation, recommandation,

réutilisation et découverte d'information, toujours en prenant en compte leur caractère multi-dimensionnel. Il nous semble également intéressant d'y intégrer à nouveau un aspect social pour identifier des communautés d'intérêt ou des réseaux d'expertise s'établissant autour de ces données.

Réflexions autour du Web (Sémantique)

Nous aimerais conclure ce mémoire en tentant de répondre à une question qui nous a été posée plusieurs fois pendant cette thèse, à savoir "*Où est la killer-app du Web Sémantique ?*". À cet égard, il nous semble que cette *killer-app* est le Web Sémantique lui-même. En effet, à partir du moment où celui-ci permet une mise en commun et un accès universel à l'information, celle-ci étant l'essence même du savoir, l'application n'est en réalité qu'un moyen d'y accéder, de la visualiser, de l'interroger. Il faut certes encore du temps pour pouvoir l'exploiter à sa juste mesure. Du temps pour que les données soient accessibles et interconnectées, ce en quoi le projet *Linking Open Data* contribue grandement. Du temps également pour que certains challenges, comme les possibilités de requêtes ou d'inférence à grande échelle puissent être pris en compte. Du temps peut-être aussi pour que l'on prenne conscience du potentiel et de la rupture technologique et sociale que le Web Sémantique peut engranger, au même titre que le Web l'a lui-même engrangée en tant que médium de communication.

Le Web arrive à une certaine maturité, comme le montre l'initiative *Web Science*⁴³ qui envisage celui-ci comme une science à part entière, combinant sociologie, droit, informatique, etc. là où celui-ci a longtemps été considéré comme un sous-ensemble de cette dernière. Malgré tout, le Web est encore jeune, et les technologies du Web Sémantique le sont encore plus. Laissons lui ainsi du temps ; après tout, comme le chantaient certains, "*It's a long way to the top (If you wanna Rock'n'Roll)*" .

⁴³<http://webscience.org>

Annexe A

Requête SPARQL pour la traduction de données RSS vers SIOC

Requête SPARQL permettant la traduction de flux RSS 1.0 en données représentées avec SIOC. Une explication complète du processus est disponible à l'URL <http://apassant.net/blog/2006/10/05/from-rss-to-sioc-using-sparql/>.

```
CONSTRUCT {
  ?channel rdf:type sioc:Forum .
  ?channel sioc:link ?channel_url .
  ?channel dc:title ?channel_title .
  ?channel dc:description ?channel_description .
  ?channel sioc:container_of ?item .
  ?item rdf:type sioc:Post .
  ?item sioc:link ?item_url .
  ?item dc:title ?item_title .
  ?item dcterms:created ?item_created .
  ?item sioc:content ?item_content .
  ?item content:encoded ?item_content_encoded .
  ?item dc:subject ?item_subject .
  ?item foaf:maker _:foaf .
  _:foaf foaf:name ?item_creator .
  _:foaf foaf:holdsAccount _:sioc .
  _:foaf rdf:type foaf:Person .
  ?item sioc:has_creator _:sioc .
  _:sioc rdf:type sioc:User .
  _:sioc sioc:name ?item_creator .
} WHERE {
  ?channel rdf:type rss:channel .
  ?channel rss:link ?channel_url .
  ?channel rss:title ?channel_title .
  ?channel rss:description ?channel_description .
  ?channel rss:items ?items .
  ?items ?li ?item .
  ?item rdf:type rss:item .
  ?item rss:link ?item_url .
  ?item rss:title ?item_title .
  ?item rss:description ?item_content .
  OPTIONAL {
```

REQUÊTE SPARQL POUR LA TRADUCTION DE DONNÉES RSS VERS SIOC

```
?item dc:date ?item_created
} . OPTIONAL {
?item content:encoded ?item_content_encoded
} . OPTIONAL {
?item dc:subject ?item_subject
} . OPTIONAL {
?item dc:creator ?item_creator
}
}
```

Annexe B

Ontologie des rôles

Ontologie pour la représentation des rôles, domaines et métiers associés à un agent (foaf:Agent).

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns="http://athena.der.edf.fr/ontologies/role#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xml:base="http://athena.der.edf.fr/ontologies/role">

  <rdf:Description rdf:about="http://athena.der.edf.fr/
    ontologies/Role">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#
      Class"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/2004/02/skos
    /core#Concept">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#
      Class"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://xmlns.com/foaf/0.1/
    Agent">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#
      Class"/>
  </rdf:Description>

  <owl:Class rdf:ID="Role">
    <rdfs:subClassOf rdf:resource="http://www.w3.org
      /2004/02/skos/core#Concept"/>
    <rdfs:comment>Le rôle associe à un Agent</rdfs:comment>
  </owl:Class>

  <owl:Class rdf:ID="RoleType">
    <rdfs:subClassOf rdf:resource="http://www.w3.org
      /2004/02/skos/core#Concept"/>
    <rdfs:label>Metier</rdfs:label>
    <rdfs:comment>Le métier associé au rôle</rdfs:comment>
```

```

        </owl:Class>

        <owl:Class rdf:ID="RoleDomain">
            <rdfs:subClassOf rdf:resource="http://www.w3.org
                /2004/02/skos/core#Concept"/>
            <rdfs:label>Domaine</rdfs:label>
            <rdfs:comment>Le domaine associé au rôle</rdfs:comment>
        </owl:Class>

        <owl:ObjectProperty rdf:ID="hasRole">
            <rdfs:label>rôle</rdfs:label>
            <rdfs:domain rdf:resource="http://xmlns.com/foaf/0.1/
                Agent"/>
            <rdfs:range rdf:resource="#Role"/>
        </owl:ObjectProperty>

        <owl:ObjectProperty rdf:ID="type">
            <rdfs:label>type de rôle</rdfs:label>
            <rdfs:domain rdf:resource="#Role"/>
            <rdfs:range rdf:resource="#RoleType"/>
        </owl:ObjectProperty>

        <owl:ObjectProperty rdf:ID="domain">
            <rdfs:label>domaine associé au rôle</rdfs:label>
            <rdfs:domain rdf:resource="Role"/>
            <rdfs:range rdf:resource="#RoleDomain"/>
        </owl:ObjectProperty>

    </rdf:RDF>

```

Annexe C

Analyse de propriétés DBpedia

Analyse des propriétés les plus couramment associées à la notion d'artiste sous DBpedia, suivant un échantillon aléatoire de 400 instances. Tableau extrait de [Passant et Raimond, 2008].

Position	Propriété	Nombre de relations
1	skos:subject	1930
2	rdf:type	882
3	dbpedia:reference	847
4	dbpedia:genre	450
5	dbpedia:page	400
6	dbpedia:hasPhotoCollection	400
7	dbpedia:origin	355
8	dbpedia:wikiPageUsesTemplate	333
9	dbpedia:label	265
10	dbpedia:wordnet_type	194
11	dbpedia:associatedActs	189
12	foaf:homepage	178
13	dbpedia:currentMembers	151
14	dbpedia:url	114
15	dbpedia:pastMembers	108
16	dbpedia:occupation	97
17	owl:sameAs	95
18	foaf:depiction	89
19	foaf:img	89
20	dpedia:wikidata-de	85

Annexe D

Exemple d'annotations métier produites avec UfoWiki

Graphe d'annotations métier produit avec UfoWiki et relatif à l'*Association des Maires de France*.

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
    xmlns:sioc="http://rdfs.org/sioc/ns#"
    xmlns:content="http://purl.org/rss/1.0/modules/content"
    /
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns"
    #
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:sioct="http://rdfs.org/sioc/types#"
    xmlns:direct="http://triplestore.aktors.org/direct/#"
    xmlns:tstore="http://triplestore.aktors.org/ontology"
    #
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:geonames="http://www.geonames.org/ontology#"
    xmlns:geo84="http://www.w3.org/2003/01/geo/wgs84_pos#"
    xmlns:ical="http://www.w3.org/2002/12/cal/ical#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:company="http://athena.der.edf.fr/ontologies/
        company#"
    xmlns:tags="http://athena.der.edf.fr/ontologies/tags#"
    xmlns:event="http://purl.org/NET/c4dm/event.owl#"
    xmlns:topic="http://athena.der.edf.fr/ontologies/topic
        #"
    xmlns:foafplus="http://athena.der.edf.fr/ontologies/
        foafplus#"
    xmlns:athena="http://athena.der.edf.fr/ontologies/
        athena#"
    xmlns:role="http://athena.der.edf.fr/ontologies/role#"
    xmlns:moat="http://moat-project.org/ns#"
```

```
xmlns:tag="http://www.holygoat.co.uk/owl/redwood/0.1/
tags/"
xmlns:doap="http://usefulinc.com/ns/doap#"
xmlns:admin="http://webns.net/mvcb/"
xmlns:dbprop="http://dbpedia.org/property/"
xmlns:partenariat="http://athena.der.edf.fr/ontologies
/partenariat#"
xmlns:wkn="http://athena.der.edf.fr/ontologies/wkn#"
>
<rdf:Description rdf:about="http://athena.der.edf.fr/
ontologies/athena#AMFAssociationDesMairesDeFrance">
<athena:name><![CDATA[AMF - Association des Maires de
France]]></athena:name></rdf:Description>
<rdf:Description rdf:about="http://athena.der.edf.fr/
ontologies/athena#AMFAssociationDesMairesDeFrance">
<foaf:name><![CDATA[Association des Maires de France
]]></foaf:name></rdf:Description>
<rdf:Description rdf:about="http://athena.der.edf.fr/
ontologies/athena#AMFAssociationDesMairesDeFrance">
<foafplus:acronym><![CDATA[AMF]]></foafplus:acronym></
rdf:Description>
<rdf:Description rdf:about="http://athena.der.edf.fr/
ontologies/athena#AMFAssociationDesMairesDeFrance">
<geonames:locatedIn rdf:resource="http://sws.geonames.
org/2988507/">
</rdf:Description>
<foafplus:Association rdf:about="http://athena.der.edf.fr/
ontologies/athena#AMFAssociationDesMairesDeFrance"/>
<rdf:Description rdf:about="http://athena.der.edf.fr/
ontologies/athena#AMFAssociationDesMairesDeFrance">
<role:hasRole rdf:resource="http://athena.der.edf.fr/
ontologies/athena#_483ab98e2a7ee"/>
</rdf:Description>
<rdf:Description rdf:about="http://athena.der.edf.fr/
ontologies/athena#AMFAssociationDesMairesDeFrance">
<role:hasRole rdf:resource="http://athena.der.edf.fr/
ontologies/athena#_483ab98e5151b"/>
</rdf:Description>
<rdf:Description rdf:about="http://athena.der.edf.fr/
ontologies/athena#RoleDomainActivitesAssociatives">
<athena:name><![CDATA[ActivitÃ s Associatives]]></athena
:name></rdf:Description>
<role:RoleDomain rdf:about="http://athena.der.edf.fr/
ontologies/athena#RoleDomainActivitesAssociatives"/>
<rdf:Description rdf:about="http://athena.der.edf.fr/
ontologies/athena#RoleDomainActivitesAssociatives">
<skos:broader rdf:resource="http://athena.der.edf.fr/
ontologies/athena#RoleDomainActivitesAssociatives"/>
</rdf:Description>
```

```

<rdf:Description rdf:about="http://athena.der.edf.fr/
    ontologies/athena#RoleDomainAdministrationPublique">
    <athena:name><![CDATA[Administration Publique]]></athena
        :name></rdf:Description>
<role:RoleDomain rdf:about="http://athena.der.edf.fr/
    ontologies/athena#RoleDomainAdministrationPublique"/>
<rdf:Description rdf:about="http://athena.der.edf.fr/
    ontologies/athena#RoleDomainAdministrationPublique">
    <skos:broader rdf:resource="http://athena.der.edf.fr/
        ontologies/athena#RoleDomainAdministrationPublique"/>
</rdf:Description>
<rdf:Description rdf:about="http://athena.der.edf.fr/
    ontologies/athena#RoleTypeAssociatif">
    <athena:name><![CDATA[Associatif]]></athena:name></rdf:
        Description>
<role:RoleType rdf:about="http://athena.der.edf.fr/
    ontologies/athena#RoleTypeAssociatif"/>
<rdf:Description rdf:about="http://athena.der.edf.fr/
    ontologies/athena#RoleTypeAssociatif">
    <skos:broader rdf:resource="http://athena.der.edf.fr/
        ontologies/athena#RoleTypeAssociatif"/>
</rdf:Description>
<rdf:Description rdf:about="http://athena.der.edf.fr/
    ontologies/athena#RoleTypeRepresentation">
    <athena:name><![CDATA[ReprÃsentation]]></athena:name></
        rdf:Description>
<role:RoleType rdf:about="http://athena.der.edf.fr/
    ontologies/athena#RoleTypeRepresentation"/>
<rdf:Description rdf:about="http://athena.der.edf.fr/
    ontologies/athena#RoleTypeRepresentation">
    <skos:broader rdf:resource="http://athena.der.edf.fr/
        ontologies/athena#RoleTypeRepresentation"/>
</rdf:Description>
<rdf:Description rdf:about="http://athena.der.edf.fr/
    ontologies/athena#_483ab98e2a7ee">
    <geonames:locatedIn rdf:resource="http://sws.geonames.
        org/3017382"/>
</rdf:Description>
<rdf:Description rdf:about="http://athena.der.edf.fr/
    ontologies/athena#_483ab98e2a7ee">
    <role:domain rdf:resource="http://athena.der.edf.fr/
        ontologies/athena#RoleDomainActivitesAssociatives"/>
</rdf:Description>
<rdf:Description rdf:about="http://athena.der.edf.fr/
    ontologies/athena#_483ab98e2a7ee">
    <role:type rdf:resource="http://athena.der.edf.fr/
        ontologies/athena#RoleTypeAssociatif"/>
</rdf:Description>
<rdf:Description rdf:about="http://athena.der.edf.fr/

```

EXEMPLE D'ANNOTATIONS MÉTIER PRODUITES AVEC UFOWIKI

```
    ontologies/athena#_483ab98e5151b">
<geonames:locatedIn rdf:resource="http://sws.geonames.org/3017382"/>
</rdf:Description>
<rdf:Description rdf:about="http://athena.der.edf.fr/
    ontologies/athena#_483ab98e5151b">
    <role:domain rdf:resource="http://athena.der.edf.fr/
        ontologies/athena#RoleDomainAdministrationPublique"/>
</rdf:Description>
<rdf:Description rdf:about="http://athena.der.edf.fr/
    ontologies/athena#_483ab98e5151b">
    <role:type rdf:resource="http://athena.der.edf.fr/
        ontologies/athena#RoleTypeRepresentation"/>
</rdf:Description>
<rdf:Description rdf:about="http://sws.geonames.org
    /2988507/">
    <athena:name><![CDATA[Paris, France]]></athena:name></
        rdf:Description>
<geonames:Feature rdf:about="http://sws.geonames.org
    /2988507"/>
<rdf:Description rdf:about="http://sws.geonames.org
    /3017382/">
    <athena:name><![CDATA[France]]></athena:name></rdf:
        Description>
<geonames:Feature rdf:about="http://sws.geonames.org
    /3017382"/>
</rdf:RDF>
```

Annexe E

Exemple d'annotations socio-structurelles produites avec UfoWiki

Graphe d'annotations socio-structurelles produit avec UfoWiki et relatif à l'*Association des Maires de France*. Le contenu textuel a volontairement été supprimé de cette annexe pour des raisons de lisibilité.

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
    xmlns:sioc="http://rdfs.org/sioc/ns#"
    xmlns:content="http://purl.org/rss/1.0/modules/content/"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:sioc="http://rdfs.org/sioc/types#"
    xmlns:direct="http://triplestore.aktors.org/direct/#"
    xmlns:tstore="http://triplestore.aktors.org/ontology/#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:geonames="http://www.geonames.org/ontology#"
    xmlns:geo84="http://www.w3.org/2003/01/geo/wgs84_pos#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:foafplus="http://athena.der.edf.fr/ontologies/
        foafplus#"
    xmlns:athena="http://athena.der.edf.fr/ontologies/athena
        #"
    xmlns:role="http://athena.der.edf.fr/ontologies/role#"
    xmlns:moat="http://moat-project.org/ns#"
    xmlns:tag="http://www.holygoat.co.uk/owl/redwood/0.1/
        tags/"
    xmlns:doap="http://usefulinc.com/ns/doap#"
    xmlns:admin="http://webns.net/mvcb/"
    xmlns:dbprop="http://dbpedia.org/property/"
    xmlns:partenariat="http://athena.der.edf.fr/ontologies/
        partenariat#"
    xmlns:wkn="http://athena.der.edf.fr/ontologies/wkn#"
```

```
>
<foaf:Document rdf:about=  >
  <dc:title>SIOC profile for Hermes </dc:title>
  <dc:description>A SIOC profile describes the structure and
    contents of a community site (e.g., weblog) in a
    machine processable form. For more information refer to
    the &lt;a href="http://rdfs.org/sioc"&gt;;
    SIOC project page&lt;/a&gt;</dc:description>
  <foaf:primaryTopic rdf:resource="http://athena.der.edf.fr/
    hermes/?q=node/16853"/>
  <admin:generatorAgent rdf:resource="http://drupal.org/
    project/sioc"/>
</foaf:Document>
<sioct:WikiArticle  rdf:about="http://athena.der.edf.fr/
  hermes/?q=node/16853">
  <dc:creator><![CDATA[Ariane Bouchet]]></dc:creator>
  <dc:title><![CDATA[AMF - Association des Maires de France
  ]]></dc:title>
  <dc:description><![CDATA[...]]></dc:description>
  <content:encoded><![CDATA[...]]></content:encoded>
  <dcterms:created>2007-11-14T15:36:00+01:00</dcterms:
    created>
  <dcterms:modified>2008-05-26T15:22:22+02:00</dcterms:
    modified>
  <sioc:link rdf:resource="http://athena.der.edf.fr/hermes/?q=node/16853" />
  <sioc:has_creator rdf:resource="http://athena.der.edf.fr/hermes/?q=user/630" rdfs:seeAlso="http://athena.der.edf.fr/hermes/?q=sioc/user/630" />
  <sioc:has_container rdf:resource="http://athena.der.edf.fr/hermes/?q=wiki/80" rdfs:seeAlso="http://athena.der.edf.fr/hermes/?q=sioc/wiki/80" />
  <wkn:embedsKnowledge rdf:resource="http://athena.der.edf.fr/hermes/?q=rdfdata/node/16853" rdfs:seeAlso="http://athena.der.edf.fr/hermes/?q=rdfdata/node/16853"/>
  <foaf:primaryTopic rdf:resource="http://athena.der.edf.fr/ontologies/athena#AMFAssociationDesMairesDeFrance"/>
  <sioc:topic rdf:resource="http://athena.der.edf.fr/ontologies/athena#RoleTypeRepresentation"/>
  <sioc:topic rdf:resource="http://sws.geonames.org/3017382"/>
  <sioc:topic rdf:resource="http://athena.der.edf.fr/ontologies/athena#RoleDomainAdministrationPublique"/>
  <sioc:topic rdf:resource="http://athena.der.edf.fr/ontologies/athena#RoleDomainActivitesAssociatives"/>
  <sioc:topic rdf:resource="http://sws.geonames.org/3017382"/>
  <sioc:topic rdf:resource="http://athena.der.edf.fr/ontologies/athena#RoleTypeAssociatif"/>
```

```
<sioc:topic rdf:resource="http://sws.geonames.org
/2988507/">
</sioc:WikiArticle>
<tag:RestrictedTagging>
  <tag:taggedResource rdf:resource="http://athena.der.edf.fr
/hermes/?q=node/16853"/>
  <tag:associatedTag>
    <moat:Tag rdf:about="http://athena.der.edf.fr/tags/tag
/hp%C3%A9dia">
      <moat:name><![CDATA [hpÃ dia]]></moat:name>
    </moat:Tag>
  </tag:associatedTag>
</tag:RestrictedTagging>
</rdf:RDF>
```

Bibliographie

- [Abel, 2008] Fabian Abel (2008). The benefit of additional semantics in folksonomy systems. *In PIKM '08 : Proceeding of the 2nd PhD workshop on Information and Knowledge Management*, pages 49–56. ACM Press.
- [Abel *et al.*, 2007] Fabian Abel, Mischa Frank, Nicola Henze, Daniel Krause, Daniel Lapappert et Patrick Siehndel (2007). GroupMe! – Where Semantic Web Meets Web 2.0. *In Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007)*, volume 4825 de *Lecture Notes in Computer Science*, pages 871–878. Springer.
- [Adida et Birbeck, 2008] Ben Adida et Mark Birbeck, éditeurs (2008). RDFa Primer 1.0. W3C Working Group Note 14 October 2008, World Wide Web Consortium. <http://www.w3.org/TR/xhtml-rdfa-primer/>.
- [Akhtar *et al.*, 2008] Waseem Akhtar, Jacek Kopecky, Thomas Krennwallner et Axel Polleres (2008). XSPARQL : Traveling between the XML and RDF worlds and avoiding the XSLT pilgrimage. *In Proceedings of the 5th European Semantic Web Conference (ESWC 2008)*, volume 5021 de *Lecture Notes in Computer Science*, pages 432–447. Springer.
- [Amardeilh, 2007] Florence Amardeilh (2007). *Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle*. Thèse de doctorat, Université Paris-X.
- [Amardeilh *et al.*, 2005] Florence Amardeilh, Philippe Laublet et Jean-Luc Minel (2005). Annotation documentaire et peuplement d'ontologie à partir d'extractions linguistiques. *In IC2005, 16èmes Journées Francophones d'Ingénierie des Connaissances*.
- [Angeletou, 2008] Sofia Angeletou (2008). Semantic Enrichment of Folksonomy Tagspaces. *In International Semantic Web Conference*, volume 5318 de *Lecture Notes in Computer Science*, pages 889–894. Springer.
- [Anicic *et al.*, 2006] Nenad Anicic, Nenad Ivezic et Albert Jones (2006). *An Architecture for Semantic Enterprise Application Integration Standards*, In Dimitri Konstantas, Jean-Paul Bourrières, Michel Léonard et Nacer Boudjlida, éditeurs : *Interoperability of Enterprise Software and Applications*, chapitre 3, pages 25–34. Springer.

- [Ankolekar et al., 2008] Anupriya Ankolekar, Markus Krötzsch, Duc Thanh Tran et Denny Vrandecic (2008). The Two Cultures : Mashing up Web 2.0 and the Semantic Web. *Journal of Web Semantics*, 6(1):70–75.
- [Ankolekar et Vrandecic, 2008] Anupriya Ankolekar et Denny Vrandecic (2008). Kalpana – enabling client-side web personalization. In *HYPertext 2008, Proceedings of the 19th ACM Conference on Hypertext and Hypermedia*, pages 21–26. ACM Press.
- [Auer, 2005] Sören Auer (2005). Powl - A Web Based Platform for Collaborative Semantic Web Development. In *First Workshop on Scripting for the Semantic Web (SFSW2005)*, volume 135 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Auer et al., 2007] Sören Auer, Chris Bizer, Jens Lehmann, Georgi Kobilarov, Richard Cyganiak et Zachary Ives (2007). Dbpedia : A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007)*, volume 4825 de *Lecture Notes in Computer Science*, pages 715–728. Springer.
- [Auer et al., 2006] Sören Auer, Sebastian Dietzold et Thomas Riechert (2006). OntoWiki - A Tool for Social, Semantic Collaboration. In *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, volume 4273 de *Lecture Notes in Computer Science*. Springer.
- [Auillans et al., 2002] Pascal Auillans, Patrice Ossona de Mendez, Pierre Rosenstiehl et Bernard Vatant (2002). A Formal Model for Topic Maps. In *The Semantic Web - ISWC 2002. First International Semantic Web Conference*, volume 2342 de *Lecture Notes in Computer Science*, pages 69–83. Springer.
- [Ayers et Völkel, 2008] Danny Ayers et Max Völkel, Leo Sauermann et Richard Cyganiak, éditeurs (2008). Cool URIs for the Semantic Web. W3C Interest Group Note 03 December 2008, World Wide Web Consortium. <http://www.w3.org/TR/cooluris/>.
- [Baader et al., 2003] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi et Peter F. Patel-Schneider (2003). *The Description Logic Handbook : Theory, Implementation, and Applications*. Cambridge University Press.
- [Bachimont, 2000] Bruno Bachimont (2000). *Engagement Sémantique et Engagement Ontologique : Conception et Réalisation D'ontologies En Ingénierie Des Connaissances*, In Manuel Zakklad, Jean Charlet, Gilles Kassel et Didier Bourigault, éditeurs : *Ingénierie des connaissances : Évolutions récentes et nouveaux défis*, chapitre 19, pages 305–324. Eyrolles.
- [Bechhofer et al., 2004] Sean Bechhofer, Frank van Harmelen, James A. Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider et Lynn Andrea Stein, Mike Dean et Guus Schreiber, éditeurs (2004). OWL Web Ontology Language Reference. W3C Recommendation 10 February 2004, World Wide Web Consortium. <http://www.w3.org/TR/owl-ref/>.
- [Beck, 1999] Kent Beck (1999). *Extreme Programming Explained : Embrace Change*. Addison-Wesley Professional.

- [Beckett, 2004] David Beckett, éditeur (2004). RDF/XML Syntax Specification (Revised). W3C Recommendation 10 February 2004, World Wide Web Consortium. <http://www.w3.org/TR/rdf-syntax-grammar/>.
- [Beckett et Berners-Lee, 2008] David Beckett et Tim Berners-Lee (2008). Turtle - Terse RDF Triple Language. W3C Team Submission 14 January 2008, World Wide Web Consortium. <http://www.w3.org/TeamSubmission/turtle/>.
- [Begelman *et al.*, 2006] Grigory Begelman, Philipp Keller et Frank Smadja (2006). Automated Tag Clustering : Improving search and exploration in the tag space. In *Proceedings of the WWW2006 Workshop on Collaborative Tagging*.
- [Berendt et Hanser, 2007] Bettina Berendt et Christoph Hanser (2007). Tags are not meta-data, but "just more content" - to some people. In *Proceedings of the First International Conference on Weblogs and Social Media (ICWSM2007)*.
- [Bernardi *et al.*, 2008] Ansgar Bernardi, Stefan Decker, Ludger van Elst, Gunnar Grimnes, Tudor Groza, Siegfried Handschuh Mehdi Jazayeri, Cedric Mesnage, Knud Moeller, Gerald Reif et Michael Sintek (2008). *The Social Semantic Desktop : A New Paradigm Towards Deploying the Semantic Web on the Desktop*, In Jorge Cardoso et Miltiadis D. Lytras, éditeurs : *Semantic Web Engineering in the Knowledge Society*, chapitre 7, pages 290–312. IGI Global.
- [Berners-Lee, 1989] Tim Berners-Lee (1989). Information Management : A Proposal. Rapport technique, CERN. <http://www.w3.org/History/1989/proposal.html>.
- [Berners-Lee, 2005a] Tim Berners-Lee (2005a). Putting the web back in semantic web. [http://www.w3.org/2005/Talks/1110-iswc-tbl/\(1\)](http://www.w3.org/2005/Talks/1110-iswc-tbl/(1)).
- [Berners-Lee, 2005b] Tim Berners-Lee (2005b). Tim Berners-Lee Podcast at ISWC2005. [www. http://esw.w3.org/topic/IswcPodcast](http://esw.w3.org/topic/IswcPodcast).
- [Berners-Lee, 2006a] Tim Berners-Lee (2006a). Linked Data. Design issues for the world wide web, World Wide Web Consortium. <http://www.w3.org/DesignIssues/LinkedData.html>.
- [Berners-Lee, 2006b] Tim Berners-Lee (2006b). Notation 3. <http://www.w3.org/DesignIssues/Notation3.html>.
- [Berners-Lee *et al.*, 2006] Tim Berners-Lee, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer et David Sheets (2006). Tabulator : Exploring and Analyzing linked data on the Semantic Web. In *Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI2006)*.
- [Berners-Lee *et al.*, 2005] Tim Berners-Lee, Roy Fielding, U.C. Irvine et Larry Masinter (2005). Uniform Resource Identifiers (URI) : Generic Syntax. Request for comments : 3986, Internet Engineering Task Force. <http://www.ietf.org/rfc/rfc3986.txt>.
- [Berners-Lee et Fischetti, 1999] Tim Berners-Lee et Mark Fischetti (1999). *Weaving the Web : The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. Harper Collins Publishers, New York.

- [Berners-Lee *et al.*, 2001] Tim Berners-Lee, James A. Hendler et Ora Lassila (2001). The Semantic Web. *Scientific American*, 284(5):34–43.
- [Berrueta *et al.*, 2007] Diego Berrueta, Dan Brickley, Stefan Decker, Sergio Fernández, Christoph Görn, Andreas Harth, Tom Heath, Kingsley Idehen, Kjetil Kjernsmo, Alistair Miles, Alexandre Passant, Axel Polleres, Luis Polo et Michael Sintek, Uldis Bojārs et John G. Breslin, éditeurs (2007). SIOC Core Ontology Specification. W3C Member Submission 12 June 2007, World Wide Web Consortium. <http://www.w3.org/Submission/sioc-spec/>.
- [Berrueta *et al.*, 2008] Diego Berrueta, Jose E. Labra. et Ivan Herman (2008). XSLT+SPARQL : Scripting the Semantic Web with SPARQL embedded into XSLT stylesheets. In *4th Workshop on Scripting for the Semantic Web (SFSW2008)*, volume 368 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Bibikas *et al.*, 2008] Dimitris Bibikas, Dimitrios Kourtesis, Iraklis Paraskakis, Ansgar Bernardi, Leo Sauermann, Dimitris Apostolou, Gregoris Mentzas et Ana Cristina Vasconcelos (2008). Organisational Knowledge Management Systems in the Era of Enterprise 2.0 : The case of OrganiK. In *BIS 2008 Workshops Proceedings*, volume 333 de *CEUR Workshop Proceedings*, pages 45–53. CEUR-WS.org.
- [Biezunski *et al.*, 2002] Michel Biezunski, Martin Bryan et Steven R. Newcomb, éditeurs (2002). ISO/IEC 13250, Topic Maps (Second Edition). Rapport technique, ISO/IEC.
- [Bizer et Cyganiak, 2007] Christian Bizer et Richard Cyganiak (2007). The TriG Syntax. Rapport technique, Freie Universität Berlin. <http://www4.wiwiiss.fu-berlin.de/bizer/TriG/>.
- [Bizer *et al.*, 2007a] Christian Bizer, Richard Cyganiak et Tobias Gauss (2007a). The rdf book mashup : From web apis to a web of data. In *3rd Workshop on Scripting for the Semantic Web (SFSW2007)*, volume 248 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Bizer *et al.*, 2007b] Chris Bizer, Richard Cyganiak et Tom Heath (2007b). How to Publish Linked Data on the Web. Rapport technique. <http://www4.wiwiiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/>.
- [Bizer *et al.*, 2008] Christian Bizer, Tom Heath, Kingsley Idehen et Tim Berners-Lee, éditeurs (2008). First Workshop on Linked Data on the Web (LDOW2008). volume 369 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Bizer et Schultz, 2008] Christian Bizer et Andreas Schultz (2008). Benchmarking the Performance of Storage Systems that expose SPARQL Endpoints. In *Proceedings of the 4th International Workshop on Scalable Semantic Web knowledge Base Systems (SSWS2008)*.
- [Bojārs, 2009] Uldis Bojārs (2009). *Establishing a Multipurpose Ontology for Describing User-Generated Content on the Semantic Web*. Thèse de doctorat, National University of Ireland, Galway. À paraître.

- [Bojārs et Breslin, 2007] Uldis Bojārs et John G. Breslin (2007). ResumeRDF : Expressing Skill Information on the Semantic Web. In *Proceedings of the 1st International ExpertFinder Workshop*.
- [Bojārs et al., 2006] Uldis Bojārs, John G. Breslin et Alexandre Passant (2006). SIOC Browser – Towards a Richer Blog Browsing Experience. In *Proceedings of the 4th Blogtalk Conference (Blogtalk Reloaded)*. Books on demand.
- [Bojārs et al., 2007a] Uldis Bojārs, John G. Breslin, Alexandre Passant et Axel Polleres, éditeurs (2007a). SIOC Ontology : Related Ontologies and RDF Vocabularies. W3C Member Submission 12 June 2007, World Wide Web Consortium. <http://www.w3.org/Submission/sioc-related/>.
- [Bojārs et al., 2008a] Uldis Bojārs, Alexandre Passant, John G. Breslin et Stefan Decker (2008a). Social Network and Data Portability using Semantic Web Technologies. In *BIS 2008 Workshops Proceedings*, volume 333 de *CEUR Workshop Proceedings*, pages 5–19. CEUR-WS.org.
- [Bojārs et al., 2008b] Uldis Bojārs, Alexandre Passant, Richard Cyganiak et John G. Breslin (2008b). Weaving sioc into the web of linked data. In *Proceedings of the WWW2008 Workshop Linked Data on the Web (LDOW2008)*, volume 369 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Bojārs et al., 2007b] Uldis Bojārs, Alexandre Passant, Frederick Giasson et John G. Breslin (2007b). An architecture to discover and query decentralized RDF data. In *3rd Workshop on Scripting for the Semantic Web (SFSW2007)*, volume 248 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Bonabeau et Theraulaz, 1994] Eric Bonabeau et Guy Theraulaz (1994). *Intelligence collective*. Hermès Science Publications.
- [Bottollier et al., 2007] Virginie Bottollier, Olivier Corby et Priscille Durville, Fabien L. Gandon, éditeur (2007). RDF/XML Source Declaration. W3C Member Submission 5 September 2007, World Wide Web Consortium. <http://www.w3.org/Submission/rdfsource/>.
- [Boyd, 2008] Danah M. Boyd (2008). *Taken Out of Context : American Teen Sociality in Networked Publics*. Thèse de doctorat, University of California, Berkeley.
- [Breslin et al., 2008] John G. Breslin, Uldis Bojārs, Alexandre Passant et Sergio Fernández, éditeurs (2008). First Workshop on Social Data on the Web (SDoW2008). volume 405 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Breslin et Decker, 2006] John G. Breslin et Stefan Decker (2006). Semantic Web 2.0 : Creating Social Semantic Information Spaces. Tutorial at the 15th International World Wide Web Conference (WWW2006).
- [Breslin et Decker, 2007] John G. Breslin et Stefan Decker (2007). The Future of Social Networks on the Internet : The Need for Semantics. *IEEE Internet Computing*, 11(6):86–90.

- [Breslin *et al.*, 2005] John G. Breslin, Andreas Harth, Uldis Bojārs et Stefan Decker (2005). Towards Semantically-Interlinked Online Communities. In *Proceedings of the 2nd European Semantic Web Conference (ESWC2005)*, volume 3532 de *Lecture Notes in Computer Science*, pages 500–514. Springer.
- [Breslin *et al.*, 2009] John G. Breslin, Alexandre Passant et Stefan Decker (2009). *The Social Semantic Web*. Springer.
- [Brickley, 2003] Dan Brickley, éditeur (2003). Basic Geo (WGS84 lat/long) Vocabulary. Rapport technique, World Wide Web Consortium. <http://www.w3.org/2003/01/geo/>.
- [Brickley et Guha, 2004] Dan Brickley et Ramanatgan V. Guha, éditeurs (2004). RDF Vocabulary Description Language 1.0 : RDF Schema. W3C Recommendation 10 February 2004, World Wide Web Consortium. <http://www.w3.org/TR/rdf-schema/>.
- [Brickley et Miller, 2004a] Dan Brickley et Libby Miller (2004a). FOAF Vocabulary Specification. Namespace Document 2 Sept 2004. <http://xmlns.com/foaf/0.1/>.
- [Brickley et Miller, 2004b] Dan Brickley et Libby Miller (2004b). FOAF Vocabulary Specification. Namespace Document 2 Sept 2004, FOAF Project. <http://xmlns.com/foaf/0.1/>.
- [Brin et Page, 1998] Sergey Brin et Lawrence Page (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.
- [Broekstra et Kampman, 2005] Jeen Broekstra et Arjohn Kampman (2005). The SeRQL query language (revision 1.2). Rapport technique, Aduna. <http://www.openrdf.org/doc/sesame/users/ch06.html>.
- [Buffa *et al.*, 2008] Michel Buffa, Fabien L. Gandon, Guillaume Eretéo, Peter Sander et Catherine Faron (2008). SweetWiki : A semantic wiki. *Journal of Web Semantics*, 6(1):84–97.
- [Bush, 1945] Vannevar Bush (1945). As We May Think. *The Atlantic Monthly*, 176(1):101–108.
- [Cao *et al.*, 2003] Tuan-Dung Cao, Fabien L. Gandon et Rose Dieng-Kuntz (2003). Intégration de sources extérieures dans un Web sémantique d'entreprise géré par un système multiagents. In *IC2003, 14èmes Journées Francophones d'Ingénierie des Connaissances*.
- [Cardon *et al.*, 2007] Dominique Cardon, Hélène Delaunay-Teterel, Cédric Fluckiger et Christophe Prieur (2007). Sociological Typology of Personal Blogs. In *Proceedings of the First International Conference on Weblogs and Social Media (ICWSM2007)*.
- [Caroll, 2003] Jeremy J. Caroll (2003). Signing RDF graphs. In *Proceedings of International Semantic Web Conference 2003 (ISWC03)*, volume 2870 de *Lecture Notes in Computer Science*, pages 369–384. Springer.
- [Caroll et Stickler, 2004] Jeremy J. Caroll et Patrick Stickler (2004). TriX : RDF Triples in XML. Technical Report HPL-2004-56, HP Labs.
- [Carroll *et al.*, 2005] Jeremy Carroll, Christian Bizer, Patrick Hayes et Patrick Stickler (2005). Named Graphs, Provenance and Trust. In *Proceedings of the 14th International World Wide Web Conference (WWW2005)*, pages 613–622.

- [Cayzer, 2004] Steve Cayzer (2004). Semantic blogging and decentralized knowledge management. *Communications of the ACM*, 47(12):47–52.
- [Cayzer, 2006] Steve Cayzer (2006). What next for Semantic Blogging ? Technical Report HPL-2006-149, HP Labs.
- [Cayzer et Castagna, 2005] Steve Cayzer et Paolo Castagna (2005). How to build a snippet manager. In *Proceedings of the 1st Workshop on The Semantic Desktop, 4th International Semantic Web Conference*, volume 175 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Cayzer et Shabajee, 2003] Steve Cayzer et Paul Shabajee (2003). Semantic Blogging and Bibliography Management. In *BlogTalk Proceedings*.
- [Charlet *et al.*, 2000] Jean Charlet, Manuel Zacklad, Gilles Kassel et Didier Bourigault, éditeurs (2000). *Ingénierie des connaissances*. Eyrolles.
- [Christensen *et al.*, 2001] Erik Christensen, Francisco Curbera, Greg Meredith et Sanjiva Weerawarana (2001). Web Service Description Language (WSDL) 1.1. W3c note 15 march 2001, World Wide Web Consortium. <http://www.w3.org/TR/wsdl1>.
- [Ciccarese *et al.*, 2008] Paolo Ciccarese, Elizabeth Wu, Gwen Wong, Marco Ocana, June Kinoshita, Alan Ruttenberg et Tim Clark (2008). The SWAN biomedical discourse ontology. *Journal of Biomedical Informatics*, 41(5):739–751.
- [Clark, 1999] James Clark, éditeur (1999). XSL Transformations (XSLT) Version 1.0. W3c recommendation 16 november 1999, World Wide Web Consortium. <http://www.w3.org/TR/xslt>.
- [Clark *et al.*, 2008] Kendall Grant Clark, Lee Feigenbaum et Elias Torres, éditeurs (2008). SPARQL Protocol for RDF. W3C Recommendation 15 January 2008, World Wide Web Consortium. <http://www.w3.org/TR/rdf-sparql-protocol/>.
- [Claudio *et al.*, 2005] Masolo Claudio, Guarino Nicola, Oltramari Alessandro et Shneider Luc (2005). The WonderWeb Library of Foundational Ontologies. Projet WonderWeb, Délivrable D18.
- [Cohen *et al.*, 2004] David Cohen, Mikael Lindvall et Patricia Costa (2004). *An introduction to agile methods*, In Marvin V. Zelkowitz, éditeur : *Advances in Computers*, volume 62, pages 2–67. Elsevier Academic Press.
- [Cointet *et al.*, 2007] Jean-Philippe Cointet, Emmanuel Faure et Camille Roth (2007). Inter-temporal Topic Correlations in Online Media : A Comparative Study on Weblogs and News Websites. In *Proceedings of the First International Conference on Weblogs and Social Media (ICWSM2007)*.
- [Connolly, 2007] Dan Connolly, éditeur (2007). Gleaning Resource Descriptions from Dialects of Languages (GRDDL). W3C Recommendation 11 September 2007, World Wide Web Consortium. <http://www.w3.org/TR/grddl1/>.
- [Corby *et al.*, 2004] Olivier Corby, Rose Dieng-Kuntz et Catherine Faron-Zucker (2004). Querying the Semantic Web with Corese Search Engine. pages 705–709. IOS Press.

- [Cyganiak *et al.*, 2008] Richard Cyganiak, Holger Stenzhorn, Renaud Delbru, Stefan Decker et Giovanni Tummarello (2008). Semantic Sitemaps : Efficient and Flexible Access to Datasets on the Semantic Web. In *Proceedings of the 5th European Semantic Web Conference (ESWC 2008)*, volume 5021 de *Lecture Notes in Computer Science*, pages 690–704. Springer.
- [d'Aquin *et al.*, 2008] Mathieu d'Aquin, Marta Sabou, Enrico Motta, Sofia Angeletou, Laurian Gridinoc, Vanessa Lopez et Fouad Zablith (2008). What Can be Done with the Semantic Web ? An Overview Watson-based Applications. In *Proceedings of the 5th Workshop on Semantic Web Applications and Perspectives (SWAP2008)*, volume 426 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Davenport et Prusak, 1997] Thomas H. Davenport et Laurence Prusak (1997). *Information Ecology : Mastering the Information and Knowledge Environment*. Oxford University Press.
- [Davis, 2005] Ian Davis (2005). An Introduction to RDF. <http://research.talis.com/2005/rdf-intro/>.
- [Decker *et al.*, 1999] Stefan Decker, Michael Erdmann, Dieter Fensel et Rudi Studer (1999). Ontobroker : Ontology Based Access to Distributed and Semi-Structured Information. In *Database Semantics : Semantic Issues in Multimedia System*, pages 351–369. Kluwer Academic Publisher.
- [Desclés, 1997] Jean-Pierre Desclés (1997). *Systèmes d'exploration contextuelle*, In Claude Guimier, éditeur : *Co-texte et Calcul du sens*, pages 215–232. Presses Universitaires de Caen.
- [Ding *et al.*, 2004] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal Doshi et Joel Sachs (2004). Swoogle : a search and metadata engine for the semantic web. In *Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management (CIKM '04)*, pages 652–659. ACM Press.
- [Djioua *et al.*, 2006] Brahim Djioua, Jorge J. García Flores, Antoine Blais, Jean-Pierre Desclés, Gaëll Guibert, Agata Jackiewicz, Florence Le Priol, Leila Nait-Baha et Benoît Sauzay (2006). EXCOM : An Automatic Annotation Engine for Semantic Information. In *Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 285–290. AAAI Press.
- [Domingue et Dzbor, 2004] John Domingue et Martin Dzbor (2004). Magpie : supporting browsing and navigation on the semantic web. In *Proceedings of the 9th International conference on Intelligent user interface*, pages 191–197. ACM Press.
- [Dublin Core Metadata Initiative, 2006] Dublin Core Metadata Initiative (2006). Dcmi metadata terms.
- [Echarte *et al.*, 2007] Francisco Echarte, José Javier Astrain, Alberto Córdoba et Jesús Villa-dangos (2007). Ontology of Folksonomy : A New Modeling Method. In *Proceedings of the Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM2007)*, volume 289 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Engelbart, 1962] Douglas C. Engelbart (1962). Augmenting Humain Intellect : A Conceptual Framework. Rapport technique, Stanford Research Institute.

- [Engelbart, 1990] Douglas C. Engelbart (1990). Knowledge-Domain Interoperability and an Open Hyperdocument System. In *Proceedings of the 1990 ACM conference on Computer-supported cooperative work*, pages 143–156. ACM Press.
- [Erling et Mikhailov, 2007] Orri Erling et Ivan Mikhailov (2007). RDF Support in the Virtuoso DBMS. In *SABRE Conference on Social Semantic Web (CSSW 2007)*, volume 113 de *Lecture Notes in Informatics*, pages 59–68. GI-EDITION.
- [Fellbaum, 1998] Christiane Fellbaum, éditeur (1998). *Wordnet, an Electronic Lexical Database*. MIT Press.
- [Fensel *et al.*, 2000] Dieter Fensel, Ian Horrocks, Frank van Harmelen, Stefan Decker, Michael Erdmann et Michel Klein (2000). OIL in a nutshell. In *Proceedings of the European Knowledge Acquisition Conference (EKAW-2000)*, volume 1937 de *Lecture Notes in Computer Science*, pages 1–16. Springer.
- [Fernández *et al.*, 2007a] Sergio Fernández, Diego Berrueta et Jose E. Labra (2007a). Mailing Lists Meet The Semantic Web. In *Proceedings of the BIS 2007 Workshop on Social Aspects of the Web (SAW2007)*, volume 245 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Fernández *et al.*, 2007b] Sergio Fernández, Frédéric Giasson et Kingsley Idehen, Uldis Bojārs, John G. Breslin et Alexandre Passant, éditeurs (2007b). SIOC Ontology : Applications and Implementation Status. W3C Member Submission 12 June 2007, World Wide Web Consortium. <http://www.w3.org/Submission/sioc-applications/>.
- [Fielding, 2000] Roy Thomas Fielding (2000). *REST : Architectural Styles and the Design of Network-based Software Architectures*. Thèse de doctorat, University of California, Irvine.
- [Franz et Staab, 2005] Thomas Franz et Steffen Staab (2005). SAM : Semantics Aware Instant Messaging for the Networked Semantic Desktop. In *Proceedings of the 1st Workshop on The Semantic Desktop, 4th International Semantic Web Conference*, volume 175 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Frappaolo et Keldsen, 2008] Carl Frappaolo et Dan Keldsen (2008). Enterprise 2.0 : Agile, Emergent Integrated. Rapport technique, AIIM – The Enterprise Content Management Association.
- [Fuchs *et al.*, 2000] Norbert E. Fuchs, Uta Schwertel et Sunna Torge (2000). Attempto Controlled English (ACE). *Journal of Language and Computation*, 1(2):199–214.
- [Fukazawa *et al.*, 2006] Yusuke Fukazawa, Takefumi Naganuma, Kunihiro Fujii et Shoji Kurakake (2006). Construction and Use of Role-Ontology for Task-Based Service Navigation System. In *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, volume 4273 de *Lecture Notes in Computer Science*, pages 806–819. Springer.
- [Gandon, 2002] Fabien L. Gandon (2002). *Intelligence artificielle distribuée et gestion des connaissances : ontologies et systèmes multi-agents pour un web sémantique organisationnel*. Thèse de doctorat, INRIA Sophia-Antipolis.

- [Gandon, 2006] Fabien L. Gandon (2006). Le web sémantique n'est pas antisocial. In *IC2006, 17èmes Journées Francophones d'Ingénierie des Connaissances*, pages 131–140.
- [Gandon, 2007] Fabien L. Gandon, éditeur (2007). GRDDL Use Cases : Scenarios of extracting RDF data from XML documents. W3c working group note 6 april 2007, World Wide Web Consortium. <http://www.w3.org/TR/grddl-scenarios/>.
- [Gandon et Giboin, 2008] Fabien L. Gandon et Alain Giboin (2008). Vers des ontologies à l'état sauvage. In *Atelier Ingénierie des Connaissances 2.0*.
- [Garey et Johnson, 1979] Michael R. Garey et David S. Johnson (1979). *Computers and Intractability – A Guide to the Theory of NP-Completeness*. W. H. Freeman And Company.
- [Giboin et al., 2008] Alain Giboin, Alexandre Passant, Philippe Laublet, Nathalie Aussenac-Gilles et Yannick Prié, éditeurs (2008). Atelier IC 2.0 : Vers une ingénierie "sociale" des connaissances : Dans quelle mesure les usages du Web 2.0 font-ils évoluer les pratiques d'IC ?
- [Gillmor, 2004] Dan Gillmor (2004). *We the Media*. O'Reilly.
- [Golder et Huberman, 2006] Scott Golder et Bernardo A. Huberman (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208.
- [Gómez-Pérez et Corcho, 2002] Asunción Gómez-Pérez et Oscar Corcho (2002). Ontology languages for the Semantic Web. *IEEE Intelligent Systems*, 17(1):54–60.
- [Gruber, 2007] Thomas Gruber (2007). Ontology of Folksonomy : A Mash-up of Apples and Oranges. *International Journal on Semantic Web and Information Systems*, 3(2):1–11.
- [Gruber, 1995] Thomas. R. Gruber (1995). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies*, 43(5–6):907–928.
- [Gruber, 2008] Thomas R. Gruber (2008). Collective Knowledge Systems : Where the Social Web Meets the Semantic Web. *Journal of Web Semantics*, 6(1):4–13.
- [Guarino, 1992] Nicola Guarino (1992). Concepts, attributes and arbitrary relations. *Data Knowledge Engineering*, 8(3):249–261.
- [Guarino et Giaretta,] Nicolas Guarino et Pierdaniele Giaretta. Ontologies and Knowledge Bases : Towards a Terminological Clarification. In *Towards Very Large Knowledge Bases (Proceedings of the 2nd International Conference on Building and Sharing of Very-Large Scale Knowledge Bases)*, pages 25–32. IOS Press.
- [Guha et al., 2003] Ramanatgan V. Guha, Rob McCool et Eric Miller (2003). Semantic Search. In *Proceedings of the 12th International World Wide Web Conference (WWW2003)*, pages 700–709. ACM Press.
- [Haarslev et Möller, 2001] Volker Haarslev et Ralf Möller (2001). Description of the RACER System and its Applications. In *Proceedings of the 2001 International Workshop on Description Logics (DL-2001)*, volume 49 de *CEUR Workshop Proceedings*. CEUR Workshop Proceedings.

- [Halpin *et al.*, 2006] Harry Halpin, Valentin Robu et Hana Shepard (2006). The Dynamics and Semantics of Collaborative Tagging. In *Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW'06)*, volume 209 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Halpin *et al.*, 2007] Harry Halpin, Valentin Robu et Hana Shepherd (2007). The Complex Dynamics of Collaborative Tagging. In *Proceedings of the 16th international conference on World Wide Web (WWW2007)*, pages 211–220.
- [Harris et Gibbins, 2003] Steve Harris et Nicholas Gibbins (2003). 3store : Efficient Bulk RDF Storage. In *Proceedings of the First International Workshop on Practical and Scalable Semantic Systems*, volume 89 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Harth *et al.*, 2005] Andreas Harth, Hannes Gassert, Ina O'Murchu, John G. Breslin et Stefan Decker (2005). WikiOnt : An Ontology for Describing and Exchanging Wikipedia Articles. In *Proceedings of Wikimania 2005 – The First International Wikimedia Conference*.
- [Harth *et al.*, 2007] Andreas Harth, Aidan Hogan, Jürgen Umbrich et Stefan Decker (2007). SWSE : Objects before documents ! In *Semantic Web Challenge 2008, collocated with the 7th International Semantic Web Conference (ISWC)*.
- [Harth *et al.*, 2006] Andreas Harth, Jürgen Umbrich et Stefan Decker (2006). MultiCrawler : A Pipelined Architecture for Crawling and Indexing Semantic Web Data. In *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, volume 4273 de *Lecture Notes in Computer Science*, pages 258–271. Springer.
- [Hartmann *et al.*, 2004] Jens Hartmann, York Sure, Alain Giboin, Diana Maynard, Mari del Carmen Suárez-Figueroa et Roberta Cuel (2004). Methods for ontology evaluation. Projet KWeb, Délivrable 1.2.3.
- [Hausenblas *et al.*, 2008] Michael Hausenblas, Wolfgang Halb et Yves Raimond (2008). Scripting User Contributed Interlinking. In *Proceedings of the WWW2008 Workshop Linked Data on the Web (LDOW2008)*, volume 369 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Hayes *et al.*, 2007] Conor Hayes, Paolo Avesani et Sriharsha Veeramachaneni (2007). An Analysis of the Use of Tags in a Blog Recommender System. In *Twentieth International Joint Conferences on Artificial Intelligence*, pages 2772–2777.
- [Hayes, 2004] Patrick Hayes, éditeur (2004). RDF Semantics. W3C Recommendation 10 February 2004, World Wide Web Consortium. <http://www.w3.org/TR/rdf-mt/>.
- [Heath et Motta, 2007] Tom Heath et Enrico Motta (2007). Revyu.com : A Reviewing and Rating Site for the Web of Data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007)*, volume 4825 de *Lecture Notes in Computer Science*, pages 895–902. Springer.
- [Heflin et Hendler, 2000] Jeff Heflin et James A. Hendler (2000). Searching the Web with SHOE. In *Artificial Intelligence for Web Search. Papers from the AAAI Workshop*. WS-00-01., pages 35–40. AAAI Press.

- [Hendler et Golbeck, 2008] James A. Hendler et Jenifer Golbeck (2008). Metcalfe's law, Web 2.0, and the Semantic Web. *Journal of Web Semantics*, 6(1):14–20.
- [Herman *et al.*, 2000] Ivan Herman, Guy Melançon et M. Scott Marshall (2000). Graph Visualization and Navigation in Information Visualization : a Survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43.
- [Hildebrand *et al.*, 2007] Michiel Hildebrand, Jacco van Ossenbruggen, Alia K. Amin, Lora Aroyo, Jan Wielemaker et Lynda Hardman (2007). The Design Space Of A Configurable Autocompletion Component. Rapport technique, CWI Amsterdam.
- [Hogan *et al.*, 2008] Aidan Hogan, Andreas Harth et Axel Polleres (2008). SAOR : Authoritative Reasoning for the Web. In *Proceedings of the 3rd Asian Semantic Web Conference (ASWC 2008)*, volume 5367 de *Lecture Notes in Computer Science*, pages 76–90. Springer.
- [Horrocks, 2002] Ian Horrocks (2002). DAML+OIL : a Description Logic for the Semantic Web. *IEEE Data Engineering Bulletin*, 25(1):4–9.
- [Huynh *et al.*, 2007] David F. Huynh, David R. Karger et Robert C. Miller (2007). Exhibit : Lightweight structured data publishing. In *Proceedings of the 16th international conference on World Wide Web*, pages 737–746.
- [Huynh-Kim-Bang et Dané, 2008] Benjamin Huynh-Kim-Bang et Eric Dané (2008). Social bookmarking et tags structurés. In *IC2008, 19èmes Journées Francophones d'Ingénierie des Connaissances*.
- [Idehen et Erling, 2008] Kingsley Idehen et Orri Erling (2008). Linked Data Spaces Data Portability. In *Proceedings of the WWW2008 Workshop Linked Data on the Web (LDOW2008)*, volume 369 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Isaac, 2005] Antoine Isaac (2005). *Conception et utilisation d'ontologies pour l'indexation de documents audiovisuels*. Thèse de doctorat, Université Paris-IV, Paris, France.
- [Isaac *et al.*, 2007] Antoine Isaac, John Phipps et Daniel Rubin, éditeurs (2007). SKOS Use Cases and Requirements. W3C Working Draft 16 May 2007, World Wide Web Consortium. <http://www.w3.org/TR/2007/WD-skos-ucr-20070516/>.
- [Jäschke *et al.*, 2008] Robert Jäschke, Andreas Hotho, Christoph Schmitz, Bernhard Ganter et Gerd Stumme (2008). Discovering Shared Conceptualizations in Folksonomies. *Journal of Web Semantics*, 6(1):38–53.
- [Java *et al.*, 2007] Akshay Java, Xiaodan Song, Tim Finin et Belle Tseng (2007). Why We Twitter : Understanding Microblogging Usage and Communities. In *WebKDD/SNA-KDD '07 : Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM Press.
- [Kahan et Koivunen, 2001] José Kahan et Marja-Ritta Koivunen (2001). Annotea : an open rdf infrastructure for shared web annotations. In *Proceedings of the 10th international conference on World Wide Web (WWW 2001)*, pages 623–632.

- [Karger et Quan, 2004] David R. Karger et Dennis Quan (2004). What Would It Mean to Blog on the Semantic Web ? In *The Semantic Web - ISWC 2004 : Third International Semantic Web Conference*, volume 3298 de *Lecture Notes in Computer Science*. Springer.
- [Kassel et Perrette, 1999] Gilles Kassel et Sébastien Perrette (1999). Co-operative ontology construction needs to carefully articulate terms, notions and objects. In *Proceedings of the International Workshop on ontological Engineering on the Global Information Infrastructure*, pages 57–70.
- [Khushraj et Lassila, 2005] Deepali Khushraj et Ora Lassila (2005). Ontological Approach to Generating Personalized User Interfaces for Web Services. In *International Semantic Web Conference*, volume 3729 de *Lecture Notes in Computer Science*, pages 916–927. Springer.
- [Kiefer et al., 2007] Christoph Kiefer, Abraham Bernstein, Hong Joo Lee, Mark Klein et Markus Stocker (2007). Semantic Process Retrieval with iSPARQL. In *Proceedings of the 4th European Semantic Web Conference (ESWC 2007)*, volume 4519 de *Lecture Notes in Computer Science*, pages 609–623. Springer.
- [Kim et al., 2007] Hak Lae Kim, Sung-Kwon Yang, John G. Breslin et Hong-Gee Kim (2007). Simple algorithms for representing tag frequencies in the scot exporter. In *Proceedings of the 2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pages 536–539. IEEE Computer Society.
- [Kiryakov et al., 2004] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov et Damyan Ognyanoff (2004). Semantic Annotation, Indexing, and Retrieval. *Journal of Web Semantics*, 2(1):49–79.
- [Klinker et al., 1991] Georg Klinker, Carlos Bhola, Geoffroy Dallemagne, David Marques et John McDermott (1991). Usable and reusable programming constructs. *Knowledge Acquisition*, 3(2):117–136.
- [Klyne et Carroll, 2004] Graham Klyne et Jeremy J. Carroll (2004). Resource Description Framework (RDF) : Concepts and abstract syntax. W3C Recommendation 10 February 2004, World Wide Web Consortium. <http://www.w3.org/TR/rdf-concepts/>.
- [Knerr, 2006] Thomas Knerr (2006). Tagging Ontology - Towards a Common Ontology for Folksonomies. <http://code.google.com/p/tagont/>.
- [Kochut et Janik, 2007] Krys Kochut et Maciej Janik (2007). SPARQLeR : Extended Sparql for Semantic Association Discovery. In *Proceedings of the 4th European Semantic Web Conference (ESWC 2007)*, volume 4519 de *Lecture Notes in Computer Science*, pages 145–159. Springer.
- [Koivunen et al., 2001] Marja-Riitta Koivunen, Ralph Swick, Jose Kaha et Eric Prud'hommeaux (2001). An Annotea Bookmark Schema. Rapport technique, World Wide Web Consortium. <http://www.w3.org/2003/07/Annotea/BookmarkSchema-20030707>.
- [Kolari et al., 2007] Pranam Kolari, Tim Finin, Yelena Yesha, Yaacov Yesha, Kelly Lyons, Stephen Perelgut et Jen Hawkins (2007). On the Structure, Properties and Utility of Internal Corporate Blogs. In *Proceedings of the First International Conference on Weblogs and Social Media (ICWSM2007)*.

- [Kraft *et al.*, 2003] Tobias Kraft, Holger Schwarz, Ralf Rantzau et Bernhard Mitschang (2003). Coarse-Grained Optimization : Techniques for Rewriting SQL Statement Sequences. In *Proceedings of the 29th international conference on Very large data bases*, pages 488–499. Morgan Kaufmann.
- [Krötzsch *et al.*, 2006] Markus Krötzsch, Denny Vrandecic et Max Völkel (2006). Semantic MediaWiki. In *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, volume 4273 de *Lecture Notes in Computer Science*, pages 935–942. Springer.
- [Kuhn, 2008] Tobias Kuhn (2008). AceWiki : Collaborative Ontology Management in Controlled Natural Language. In *Third Semantic Wiki Workshop – The Wiki Way of Semantics*, volume 360 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Lee, 2004] Ryan Lee (2004). Scalability report on triple store applications. Rapport technique, MIT – Massachusetts Institute of Technology. <http://simile.mit.edu/reports/stores/index.html>.
- [Lenat *et al.*, 1990] Douglas B. Lenat, Ramanathan V. Guha, Karen Pittman, Dexter Pratt et Mary Shepherd (1990). Cyc : Toward Programs with Common Sense. *Communications of the ACM*, 33(8):30–49.
- [Leuf et Cunningham, 2001] Bo Leuf et Ward Cunningham (2001). *The Wiki Way : Collaboration and Sharing on the Internet*. Addison-Wesley Professional.
- [Lewis, 2007] Rhys Lewis (2007). Dereferencing http uris. Draft Tag Finding 31 May 2007, World Wide Web Consortium. <http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14.html>.
- [Libert *et al.*, 2007] Barry Libert, Jon Spector et Don Tapscott (2007). *We Are Smarter Than Me : How to Unleash the Power of Crowds in Your Business*. Wharton School Publishing.
- [Luke et Heflin, 2000] Sean Luke et Jeff Heflin (2000). Shoe 1.01. Rapport technique, Parallel Understanding Systems Group, Department of Computer Science, University of Maryland at College Park.
- [Maedche *et al.*, 2003] Alexander Maedche, Boris Motik, Ljiljana Stojanovic, Rudi Studer et Raphael Volz (2003). Ontologies for Enterprise Knowledge Management. *IEEE Intelligent Systems*, 18(2):26–33.
- [Marlow *et al.*, 2006] Cameron Marlow, Mor Naaman, Danah Boyd et Marc Davis (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *HYPertext '06 : Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40. ACM Press.
- [Martre, 1994] Henri Martre, Paris : La Documentation Française, éditeur (1994). Intelligence économique et stratégie des entreprises. Rapport technique, Commissariat général du Plan.
- [Mathes, 2004] Adam Mathes (2004). Folksonomies : Cooperative Classification and Communication Through Shared Metadata.

- [Mcafee, 2006] Andrew P. McAfee (2006). Enterprise 2.0 : The Dawn of Emergent Collaboration. *MIT Sloan Management Review*, 47(3):21–28.
- [McGuinness *et al.*, 2003] Deborah L. McGuinness, Richard Fikes, Lynn Andrea Stein et James A. Hendler (2003). *DAML-ONT : An Ontology Language for the Semantic Web*, In Dieter Fensel, James A. Hendler, Henry Lieberman et Wolfgang Wahlster, éditeurs : *Spinning the Semantic Web*, chapitre 3, pages 65–93. MIT Press.
- [Mika, 2005] Peter Mika (2005). Ontologies Are Us : A Unified Model of Social Networks and Semantics. In *Proceedings of the 4th International Semantic Web Conference (ISWC 2005)*, volume 3729 de *Lecture Notes in Computer Science*, pages 522–536. Springer.
- [Mika, 2008] Peter Mika (2008). Microsearch : An Interface for Semantic Search. In *Proceedings of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008)*, volume 334 de *CEUR Workshop Proceedings*. CEUR Workshop Proceedings.
- [Miles et Bechhofer, 2008] Alistair Miles et Sean Bechhofer (2008). SKOS Simple Knowledge Organization System Reference. W3C Working Draft 29 August 2008, World Wide Web Consortium. <http://www.w3.org/TR/2008/WD-skos-reference-20080829/>.
- [Milićić, 2008] Vuc Milićić (2008). Semantic tags. W3C SWEO Case Study, World Wide Web Consortium. <http://www.w3.org/2001/sw/sweo/public/UseCases/Faviki/>.
- [Möller *et al.*, 2006] Knud Möller, Uldis Bojārs et John G. Breslin (2006). Using Semantics to Enhance the Blogging Experience. In *Proceedings of the 3th European Semantic Web Conference (ESWC 2006)*, volume 4011 de *Lecture Notes in Computer Science*, pages 679–696. Springer.
- [Nakayama, 2008] Kotaro Nakayama (2008). Wikipedia Mining for Triple Extraction Enhanced by Co-reference Resolution. In *Proceedings of the ISWC2008 Workshop on Social Data on the Web (SDoW2008)*, volume 405 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Nelson, 1965] Theodor H. Nelson (1965). Complex information processing : a file structure for the complex, the changing and the indeterminate. In *Proceedings of the 1965 20th ACM national conference*, pages 84–100. ACM Press.
- [Newman *et al.*, 2005] Richard Newman, Danny Ayers et Seth Russell (2005). Tag ontology. <http://www.holygoat.co.uk/owl/redwood/0.1/tags/>.
- [Nickull *et al.*, 2008] Duane Nickull, Dion Hinchcliffe et James Governor (2008). *Web 2.0 Patterns : What entrepreneurs and information architects need to know*. O'Reilly.
- [Nottingham et Sayre, 2005] Mark Nottingham et Robert Sayre (2005). The Atom Syndication Format. Request for comments : 3986, Internet Engineering Task Force. <http://www.ietf.org/rfc/rfc4287.txt>.
- [Nowack, 2008] Benjamin Nowack (2008). Sparql+, sparqlscript, sparql result templates - sparql extensions for the mashup developer. In *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008)*, volume 401 de *CEUR Workshop Proceedings*. CEUR-WS.org.

- [O'Reilly, 2005] Tim O'Reilly (2005). O'Reilly Network : What Is Web 2.0 : Design Patterns and Business Models for the Next Generation of Software. <http://www.oreillynet.com/lpt/a/6228>.
- [Oren, 2005] Eyal Oren (2005). SemperWiki : a semantic personal Wiki. In *Proceedings of the 1st Workshop on The Semantic Desktop, 4th International Semantic Web Conference*, volume 175 de CEUR Workshop Proceedings. CEUR-WS.org.
- [Oren *et al.*, 2006] Eyal Oren, Renaud Delbru et Stefan Decker (2006). Extending faceted navigation for rdf data. In *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, volume 4273 de Lecture Notes in Computer Science, pages 559–572. Springer.
- [Oren *et al.*, 2007] Eyal Oren, Renaud Delbru, Sebastian Gerke, Armin Haller et Stefan Decker (2007). Activerdf : Object-oriented semantic web programming. In *Proceedings of the 16th international conference on World Wide Web (WWW2007)*, pages 817–824.
- [Osterfeld *et al.*, 2005] Frank Osterfeld, Malte Kiesel et Sven Schwarz (2005). Nabu – a semantic archive for xmpp instant messaging. In *Proceedings of the 1st Workshop on The Semantic Desktop, 4th International Semantic Web Conference*, volume 175 de CEUR Workshop Proceedings. CEUR-WS.org.
- [Pan *et al.*, 2008] Jeff Z. Pan, Giorgos Stamou, Giorgos Stoilos, Edward Thomas, et Stuart Taylor (2008). Scalable Querying Service over Fuzzy Ontologies. In *Proceedings of the 17th International World Wide Web Conference (WWW 2008)*, pages 575–584.
- [Passant, 2006] Alexandre Passant (2006). FOAFMap : Web2.0 meets the Semantic Web. In *2nd Workshop on Scripting for the Semantic Web (SFSW2006)*, volume 181 de CEUR Workshop Proceedings. CEUR-WS.org.
- [Passant, 2007a] Alexandre Passant (2007a). Linked Data tagging with LODr. In *Semantic Web Challenge 2008, collocated with the 7th International Semantic Web Conference (ISWC)*.
- [Passant, 2007b] Alexandre Passant (2007b). A user-friendly interface to browse and find DOAP project with doap :store. In *3rd Workshop on Scripting for the Semantic Web (SFSW2007)*, volume 248 de CEUR Workshop Proceedings. CEUR-WS.org.
- [Passant, 2007c] Alexandre Passant (2007c). Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs. In *Proceedings of the First International Conference on Weblogs and Social Media (ICWSM2007)*.
- [Passant, 2008a] Alexandre Passant (2008a). Enhancement and Integration of Corporate Social Software Using the Semantic Web. W3C SWEO Case Study, World Wide Web Consortium. <http://www.w3.org/2001/sw/sweo/public/UseCases/EDF/>.
- [Passant, 2008b] Alexandre Passant (2008b). :me owl :sameAs flickr :33669349@N00. In *Proceedings of the WWW2008 Workshop Linked Data on the Web (LDOW2008)*, volume 369 de CEUR Workshop Proceedings. CEUR-WS.org.
- [Passant *et al.*, 2008] Alexandre Passant, Tuukka Hastrup, Uldis Bojārs et John G. Breslin (2008). Microblogging : A Semantic Web and Distributed Approach. In *4th Workshop on*

Scripting for the Semantic Web (SFSW2008), volume 368 de *CEUR Workshop Proceedings*. CEUR-WS.org.

[Passant *et al.*, 2009a] Alexandre Passant, Jacek Kopecký, Stéphane Corlosquet, Diego Berrueta, Davide Palmisano et Axel Polleres, éditeurs (2009a). XSPARQL : Use cases. Rapport technique. <http://xsparql.deri.org/spec/xsparql-use-cases.html>.

[Passant *et al.*, 2009b] Alexandre Passant, Philipp Kärger, Michael Hausenblas, Daniel Olmedilla, Axel Polleres et Stefan Decker (2009b). Enabling Trust and Privacy on the Social Web. In *W3C Workshop on the Future of Social Networking*.

[Passant et Laublet, 2008a] Alexandre Passant et Philippe Laublet (2008a). Combining Structure and Semantics for Ontology-Based Corporate Wikis. In *11th International Conference on Business Information Systems, BIS 2008*, volume 7 de *Lecture Notes in Business Information Processing*, pages 58–69. Springer.

[Passant et Laublet, 2008b] Alexandre Passant et Philippe Laublet (2008b). Meaning Of A Tag : A collaborative approach to bridge the gap between tagging and Linked Data. In *Proceedings of the WWW2008 Workshop Linked Data on the Web (LDOW2008)*, volume 369 de *CEUR Workshop Proceedings*. CEUR-WS.org.

[Passant et Laublet, 2008c] Alexandre Passant et Philippe Laublet (2008c). Ontologies et Web 2.0. In *IC2008, 19èmes Journées Francophones d'Ingénierie des Connaissances*.

[Passant et Laublet, 2008d] Alexandre Passant et Philippe Laublet (2008d). Towards an Interlinked Semantic Wiki Farm. In *Third Semantic Wiki Workshop – The Wiki Way of Semantics*, volume 360 de *CEUR Workshop Proceedings*. CEUR-WS.org.

[Passant et Laublet, 2008e] Alexandre Passant et Philippe Laublet (2008e). Wikis sémantiques : Le peuplement d’ontologies pour tous ? In *Atelier Ingénierie des Connaissances 2.0*.

[Passant et Raimond, 2008] Alexandre Passant et Yves Raimond (2008). Combining Social Music and Semantic Web for music-related recommender systems. In *Proceedings of the ISWC2008 Workshop on Social Data on the Web (SDoW2008)*, volume 405 de *CEUR Workshop Proceedings*. CEUR-WS.org.

[Passant *et al.*, 2009c] Alexandre Passant, Matthias Samwald, John G. Breslin et Stefan Decker (2009c). Federating Distributed Social Data to Build an Interlinked Online Information Society. In *WebSci'09 : Society On-Line*.

[Passant *et al.*, 2006] Alexandre Passant, Jean-David Sta et Philippe Laublet (2006). Folksonomies, Ontologies and Corporate Bloging. In *Proceedings of the 4th Blogtalk Conference (Blogtalk Reloaded)*. Books on demand.

[Patel-Schneider *et al.*, 2004] Peter F. Patel-Schneider, Patrick Hayes et Ian Horrocks, éditeurs (2004). OWL Web Ontology Language Semantics and Abstract Syntax. W3C Recommendation 10 February 2004, World Wide Web Consortium. <http://www.w3.org/TR/owl-semantics/>.

- [Pérez *et al.*, 2006] Jorge Pérez, Marcelo Arenas et Claudio Gutierrez (2006). Semantics and Complexity of SPARQL. In *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, volume 4273 de *Lecture Notes in Computer Science*, pages 30–43. Springer.
- [Polleres *et al.*, 2007] Axel Polleres, François Scharffe et Roman Schindlauer (2007). SPARQL++ for Mapping Between RDF Vocabularies. In *Proceedings of the 6th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2007)*, volume 4803 de *Lecture Notes in Computer Science*, pages 878–896. Springer.
- [Prud'hommeaux et Seaborne, 2008] Eric Prud'hommeaux et Andy Seaborne, éditeurs (2008). SPARQL query language for RDF. W3C Recommendation 15 January 2008, World Wide Web Consortium. <http://www.w3.org/TR/rdf-sparql-query/>.
- [Quan *et al.*, 2003a] Dennis Quan, Karun Bakshi et David R. Karger (2003a). A Unified Abstraction for Messaging on the Semantic Web. In *Proceedings of the 12th International World Wide Web Conference (WWW2003)*, page 231. ACM Press.
- [Quan *et al.*, 2003b] Dennis Quan, David Huynh et David R. Karger (2003b). Haystack : A Platform for Authoring End User Semantic Web Applications. In *Proceedings of International Semantic Web Conference 2003 (ISWC03)*, volume 2870 de *Lecture Notes in Computer Science*, pages 738–753. Springer.
- [Quillian, 1968] Ross Quillian (1968). *Semantic Memory*, In Marvin L. Minsky, éditeur : *Semantic Information Processing*, pages 216–270. MIT Press.
- [Rada *et al.*, 1989] R. Rada, H. Mili, E. Bicknell et M. Blettner (1989). Development and application of a metric on semantic nets. 19(1):17–30.
- [Rager *et al.*, 1997] David Rager, James A. Hendler et Alice M. Mulvehill (1997). ForMAT and Parka : A Technology Integration Experiment and Beyond. In *Case-Based Reasoning Research and Development : Proceedings of the Second International Conference on Case-Based Reasoning, (ICCBR'97)*, volume 1266 de *Lecture Notes in Computer Science*, pages 122–132. Springer.
- [Raimond *et al.*, 2008] Yves Raimond, Christopher Sutton et Mark Sandler (2008). Automatic Interlinking of Music Datasets on the Semantic Web. In *Proceedings of the WWW2008 Workshop Linked Data on the Web (LDOW2008)*, volume 369 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Rehatschek et Hausenblas, 2007] Herwig Rehatschek et Michael Hausenblas (2007). Enhancing the Exploration of Mailing List Archives Through Making Semantics Explicit. In *Semantic Web Challenge 2007, collocated with the 6th International Semantic Web Conference (ISWC)*.
- [Rousset *et al.*, 2002] Marie-Christine Rousset, Alain Bidault, Christine Froidevaux, Hélène Gagliardi, François Goasdoué, Chantal Reynaud et Brigitte Safar (2002). Construction de Médiateurs pour Intégrer des Sources d'information multiples et hétérogènes. *Revue I3*, 2:9–59.

- [Russell et Norvig, 2003] Stuart J. Russell et Peter Norvig (2003). *Artificial Intelligence : A Modern Approach*. Pearson Education.
- [Salton et McGill, 1986] Gerard Salton et Michael J. McGill (1986). *Introduction to Modern Information Retrieval*. McGraw Hill Computer Science Series.
- [Sanderson et Croft, 1999] Mark Sanderson et William Bruce Croft (1999). Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'99*, pages 206–213. ACM Press.
- [Scerri et al., 2007] Simon Scerri, Michael Sintek, Ludger van Elst et Siegfried Handschuh, Simon Scerri, éditeur (2007). NEPOMUK Annotation Ontology Specification. Rapport technique. <http://www.semanticdesktop.org/ontologies/nao/>.
- [Schaffert, 2006] Sebastian Schaffert (2006). IkeWiki : A Semantic Wiki for Collaborative Knowledge Management. In *First International Workshop on Semantic Technologies in Collaborative Applications (STICA 06)*.
- [Schmitz, 2006] Patrick Schmitz (2006). Inducing Ontology from Flickr Tags. In *Proceedings of the WWW2006 Workshop on Collaborative Tagging*.
- [Scott et al., 2008] Tom Scott, Yves Raimond, Patrick Sinclair et Nicholas Humfrey (2008). The Programmes Ontology. In *XTech 2008 : The Web on the Move*.
- [Seaborne, 2004] Andy Seaborne (2004). RDQL – A Query Language for RDF. W3c member submission 9 january 2004, World Wide Web Consortium. <http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/>.
- [Seaborne et al., 2008] Andy Seaborne, Geetha Manjunath, Chris Bizer, John G. Breslin, Sou-ripriya Das, Ian Davis, Steve Harris, Kingsley Idehen, Olivier Corby, Kjetil Kjernsmo et Benjamin Nowack (2008). SPARQL Update – A language for updating RDF graphs. W3C Member Submission 15 July 2008, World Wide Web Consortium. <http://www.w3.org/Submission/2008/SUBM-SPARQL-Update-20080715/>.
- [Servant, 2006] François-Paul Servant (2006). Semanlink. In *Jena User Conference (JUC)*.
- [Sheth et al., 2002] Amit P. Sheth, Clemens Bertram, David Avant, Brian Hammond, Krys Kochut et Yashodhan S. Warke (2002). Managing Semantic Content for the Web. *IEEE Internet Computing*, 6(4):80–87.
- [Sirin et al., 2007] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur et Yarden Katz (2007). Pellet : A practical OWL-DL reasoner. *Journal of Web Semantics*, 5(2):51–53.
- [Sowa, 1984] John F. Sowa (1984). *Conceptual Structures : Information processing in mind and machine*. Addison-Wesley Longman Publishing Co., Inc.
- [Specia et Motta, 2007] Lucia Specia et Enrico Motta (2007). Integrating Folksonomies with the Semantic Web. In *Proceedings of the 4th European Semantic Web Conference (ESWC 2007)*, volume 4519 de *Lecture Notes in Computer Science*, pages 624–639. Springer.

- [Staab, 2002] Steffen Staab (2002). Emergent semantics. *IEEE Intelligent Systems*, 17(1):78–86.
- [Steimann, 2000] Friedrich Steimann (2000). On the representation of roles in object-oriented and conceptual modelling. *Data Knowledge Engineering*, 35(1):83–106.
- [Stocker *et al.*, 2008] Markus Stocker, Christoph Kiefer Andy Seaborne, Abraham Bernstein et Dave Reynolds (2008). SPARQL Basic Graph Pattern Optimization Using Selectivity Estimation. In *Proceedings of the 17th International World Wide Web Conference (WWW 2008)*, pages 595–604.
- [Suchanek *et al.*, 2007] Fabian M. Suchanek, Gjergji Kasneci et Gerhard Weikum (2007). Yago : A Core of Semantic Knowledge – Unifying WordNet and Wikipedia. In *Proceedings of the 16th international conference on World Wide Web (WWW2007)*, pages 697–706. ACM Press.
- [Sure *et al.*, 2005] York Sure, Stephan Bloehdorn, Peter Haase, Jens Hartmann et Daniel Oberle (2005). The SWRC ontology – Semantic Web for research communities. In *Progress in Artificial Intelligence – Proceedings of the 12th Portuguese Conference on Artificial Intelligence (EPIA 2005)*, volume 3808 de *Lecture Notes in Computer Science*. Springer.
- [SVG Working Group, 2003] SVG Working Group, Jon Ferraiolo, Jun Fujisawa et Dean Jackson, éditeurs (2003). Scalable Vector Graphics (SVG) 1.1 Specification. W3C Recommendation 14 January 2003, World Wide Web Consortium. <http://www.w3.org/TR/SVG11/>.
- [Tanaka et Taylor, 1991] James W. Tanaka et Marjorie Taylor (1991). Object categories and expertise : Is the basic level in the eye of the beholder ? *Cognitive Psychology*, 23(3):457–482.
- [Tanasescu et Streibel, 2007] Vlad Tanasescu et Olga Streibel (2007). Extreme Tagging : Emergent Semantics through the Tagging of Tags. In *Proceedings of the International Workshop on Emergent Semantics and Ontology Evolution (ESOE2007)*, volume 292 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Tapscott et Williams, 2007] Don Tapscott et Anthony D. Williams (2007). *Wikinomics : How Mass Collaboration Changes Everything*. Pearson Education.
- [Taylor, 1999] Arlene G. Taylor (1999). *The Organization of Information*. Libraries Unlimited.
- [Tazzoli *et al.*, 2004] Roberto Tazzoli, Paolo Castagna et Stefano Emilio Campanini (2004). Towards a Semantic WikiWikiWeb. In *The Semantic Web - ISWC 2004 : Third International Semantic Web Conference*, volume 3298 de *Lecture Notes in Computer Science*. Springer.
- [Terziev *et al.*, 2005] Ivan Terziev, Atanas Kiryakov et Dimitar Manov (2005). Base Upper-level Ontology (BULO) Guidance. Projet SEKT, Délivrable 1.8.1.
- [Troncy, 2004] Raphaël Troncy (2004). *Formalisation des connaissances documentaires et des connaissances conceptuelles à l'aide d'ontologies : application à la description de documents audiovisuels*. Thèse de doctorat, Université Joseph Fourier-INPG.

- [Tummarello *et al.*, 2007] Giovanni Tummarello, Renaud Delbru et Eyal Oren (2007). Sindice.com : Weaving the Open Linked Data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007)*, volume 4825 de *Lecture Notes in Computer Science*, pages 552–565. Springer.
- [Van Damme *et al.*, 2007] Céline Van Damme, Martin Hepp et Katharina Siorpaes (2007). FolksOntology : An Integrated Approach for Turning Folksonomies into Ontologies. In *Proceedings of the ESWC'2007 workshop Bridging the Gap between Semantic Web and Web 2.0*.
- [Vander Wal, 2007] Thomas Vander Wal (2007). Folksonomy Coinage and Definition. url{<http://www.vanderwal.net/folksonomy.html>}.
- [Vitvar *et al.*, 2008] Tomas Vitvar, Jacek Kopecky, Jana Viskova et Dieter Fensel (2008). WSMO-Lite Annotations for Web Services. In *Proceedings of the 5th European Semantic Web Conference (ESWC 2008)*, volume 5021 de *Lecture Notes in Computer Science*, pages 674–689. Springer.
- [Völkel et Oren, 2006] Max Völkel et Eyal Oren (2006). Towards a Wiki Interchange Format (WIF) - Opening Semantic Wiki Content and Metadata. In *Proceedings of the First Workshop on Semantic Wikis - From Wiki to Semantics (SemWiki-2006)*, volume 206 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [Völkel et Schaffert, 2006] Max Völkel et Sébastien Schaffert, éditeurs (2006). First Workshop on Semantic Wikis – From Wiki to Semantics. volume 206 de *CEUR Workshop Proceedings*. CEUR-WS.org.
- [W3C Technical Architecture Group, 2004] W3C Technical Architecture Group, Ian Jacobs et Norman Walsh, éditeurs (2004). Architecture of the World Wide Web, Volume One. W3C Recommendation 15 December 2004, World Wide Web Consortium. <http://www.w3.org/TR/webarch/>.
- [Welty et Guarino, 2001] Christopher A. Welty et Nicola Guarino (2001). Supporting ontological analysis of taxonomic relationships. *Data Knowledge Engineering*, 39(1):51–74.
- [West, 2000] Douglas B. West (2000). *Introduction to Graph Theory (Second Edition)*. Prentice Hall.
- [Wiederhold, 1992] Gio Wiederhold (1992). Mediators in the Architecture of Future Information Systems. *IEEE Computer*, 25(3):38–49.
- [Wilensky, 1967] Harold L. Wilensky (1967). *Organizational intelligence*. Basic Books.
- [Wu et Weld, 2008] Fei Wu et Daniel S. Weld (2008). Automatically refining the wikipedia infobox ontology. In *Proceedings of the 17th International World Wide Web Conference (WWW 2008)*, pages 635–644.
- [Xyleme, 2001] Lucie Xyleme (2001). A dynamic warehouse for XML Data of the Web. *IEEE Data Engineering Bulletin*, 24(2):40–47.

- [Yee *et al.*, 2003] Ka-Ping Yee, Kirsten Swearingen, Kevin Li et Marti Hearst (2003). Faceted Metadata for Image Search and Browsing. In *CHI '03 : Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 401–408. ACM Press.
- [Zacklad, 2005] Manuel Zacklad (2005). Introduction aux ontologies sémiotiques dans le Web Socio Sémantique. In *IC2005, 16èmes Journées Francophones d'Ingénierie des Connaissances*.
- [Zacklad, 2007] Manuel Zacklad (2007). Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information (roi). In *CAIS/ACSI 2007, 35e Congrès annuel de l'Association Canadienne des Sciences de l'Information*.

Technologies du Web Sémantique pour l'Entreprise 2.0

Les travaux présentés dans cette thèse proposent différentes méthodes, réflexions et réalisations associant Web 2.0 et Web Sémantique. Après avoir introduit ces deux notions, nous présentons les limites actuelles de certains outils, comme les blogs ou les wikis, et des pratiques de tagging dans un contexte d'Entreprise 2.0. Nous proposons ensuite la méthode *SemSLATES* et la vision globale d'une architecture de médiation reposant sur les standards du Web Sémantique (langages, modèles, outils et protocoles) pour pallier à ces limites. Nous détaillons par la suite différentes ontologies (au sens informatique) développées pour mener à bien cette vision : d'une part, en contribuant activement au projet SIOC – *Semantically-Interlinked Online Communities* –, des modèles destinés aux métadonnées socio-structurelles, d'autre part des modèles, étendant des ontologies publiques, destinés aux données métier. De plus, la définition de l'ontologie MOAT – *Meaning Of A Tag* – nous permet de coupler la souplesse du tagging et la puissance de l'indexation à base d'ontologies. Nous revenons ensuite sur différentes implémentations logicielles que nous avons mises en place à EDF R&D pour permettre de manière intuitive la production et l'utilisation d'annotations sémantiques afin d'enrichir les outils initiaux : wikis sémantiques, interfaces avancées de visualisation (navigation à facettes, mash-up sémantique, etc.) et moteur de recherche sémantique. Plusieurs contributions ont été publiées sous forme d'ontologies publiques ou de logiciels libres, contribuant de manière plus large à cette convergence entre Web 2.0 et Web Sémantique non seulement en entreprise mais sur le Web dans son ensemble.

Mot-clés : Web 2.0, Entreprise 2.0, Web Sémantique, Ontologies, Folksonomies, Wikis, SIOC, MOAT

Semantic Web technologies for Enterprise 2.0

The work described in this thesis provides different methods, thoughts and implementations combining Web 2.0 and the Semantic Web. After introducing those terms, we present the current shortcomings of tools such as blogs and wikis as well as tagging practices in an Enterprise 2.0 context. We define the *SemSLATES* methodology and the global vision of a middleware architecture based on Semantic Web technologies (languages, models, tools and protocols) to solve these issues. Then, we detail the various ontologies (as in computer science) that we build to achieve this goal: on the one hand models dedicated to socio-structural meta-data, by actively contributing to SIOC – *Semantically-Interlinked Online Communities* –, and on the other hands models extending public ontologies for domain data. Moreover, the MOAT ontology – *Meaning Of A Tag* – allows us to combine the flexibility of tagging and the power of ontology-based indexing. We then describe several software implementations, at EDF R&D, dedicated to easily produce and use semantic annotations to enrich original tools: semantic wikis, advanced visualization interfaces (faceted browsing, semantic mash-ups, etc.) and a semantic search engine. Several contributions have been published as public ontologies or open-source software, contributing more generally to this convergence between Web 2.0 and the Semantic Web, not only in enterprise but on the Web as a whole.

Keywords: Web 2.0, Enterprise 2.0, Semantic Web, Ontologies, Folksonomies, Wikis, SIOC, MOAT

Discipline : Informatique

Laboratoire : Langues, Logiques, Informatique, Cognition (LaLIC), Équipe d'Accueil (EA 4350),
Maison de la Recherche, 28 rue Serpente, 75006 Paris, France