dbpedia.fr

Ce document propose la création d'une version francophone de la base DBPedia utilisée dans de nombreuses applications anglophones, notamment pour la publication de collections culturelles.

Qu'est-ce que DBPedia ?

DBPedia est une base de données publique extraite de Wikipédia, projet d'encyclopédie collective sur le Web, fonctionnant sur le principe désormais bien connu du wiki. L'encyclopédie Wikipédia offre un contenu librement réutilisable, que chacun peut modifier et améliorer dans les limites d'un processus éditorial défini. Elle est essentiellement à destination humaine : seuls des individus peuvent véritablement lire et comprendre le contenu de ses pages. Or celles-ci n'en recèlent pas moins des données pouvant s'avérer utiles aux applications informatiques, à la condition qu'elles leur deviennent accessibles (ex. : dans le contexte croissant de la mobilité, la longitude et la latitude des musées et autres monuments…).

A l'instar de son modèle, DBPedia est également un effort communautaire ayant pour but d'extraire des informations structurées de Wikipédia afin de rendre ces données disponibles sur le Web. En tant que base de connaissances, DBPedia bénéficie donc du gigantesque corpus de Wikipédia et décrit actuellement plus de 3,5 million de d'objets, dont 364 000 personnes, 462 000 lieux, 99 000 albums de musique, 54 000 films, 17 000 jeux vidéo, 148 000 organisations, 169 000 espèces et 5 200 maladies. Les données de DBPedia décrivent en outre ces 3,5 millions d'éléments dans 97 langues différentes et proposent 1 850 000 liens vers des images et 5 900 000 liens vers des pages Web externes.

En extrayant les données de Wikipédia puis en les publiant conséquemment dans un format structuré (standards ouverts du Web Sémantique), DBPedia les rend accessibles à tout un chacun et favorise par là même l'émergence de nouvelles applications et de nouveaux usages, dans des domaines aussi variés que la recherche, l'industrie ou encore, bien entendu, la culture.

A quoi sert DBPedia ?

En tant que base de connaissances, DBPedia a plusieurs avantages sur les bases existantes. Elle couvre de très nombreux domaines, capture un véritable consensus collectif qui évolue automatiquement en fonction des transformations de Wikipédia et peut ainsi suivre les nouvelles tendances, pierre d'achoppement de la plupart des référentiels et autres thésaurus peuplant les systèmes d'information traditionnels. Elle fournit dès lors un large référentiel vivant pour d'autres collections de ressources sur le Web et rend possible, entre autre et dans le désordre : leur identification, indexation, références croisées, intégration, interrogation structurée, et même certaines formes de raisonnement automatique utiles, par exemple, à la recherche d'information.

DBPedia permet donc de répondre automatiquement à des requêtes structurées complexes sur les données de Wikipédia et de les lier à d'autres ensembles de données sur le Web.

La figure 1 est un agrandissement de la constellation des sources recensées sur le web de données (Figure 2). Les exemples d'application pullulent dans l'espace anglophone, aucune autre source de données ne bénéficiant à l'heure actuelle d'une pareille centralité. Tout un pan de ces applications concerne la culture, et la figure 1 montre des collections liées à DBPedia dans le domaine de la musique, de la presse, des documentaires, etc. Mentionnons à titre d'exemple les nouveaux mécanismes de navigation mis en place sur de nombreux sites, comme en témoigne notamment la plateforme consacrée aux documentaires animaliers de la BBC, qui exploite les données et catégories de DBPedia afin de proposer une meilleure structuration de ses contenus et une navigation enrichie.

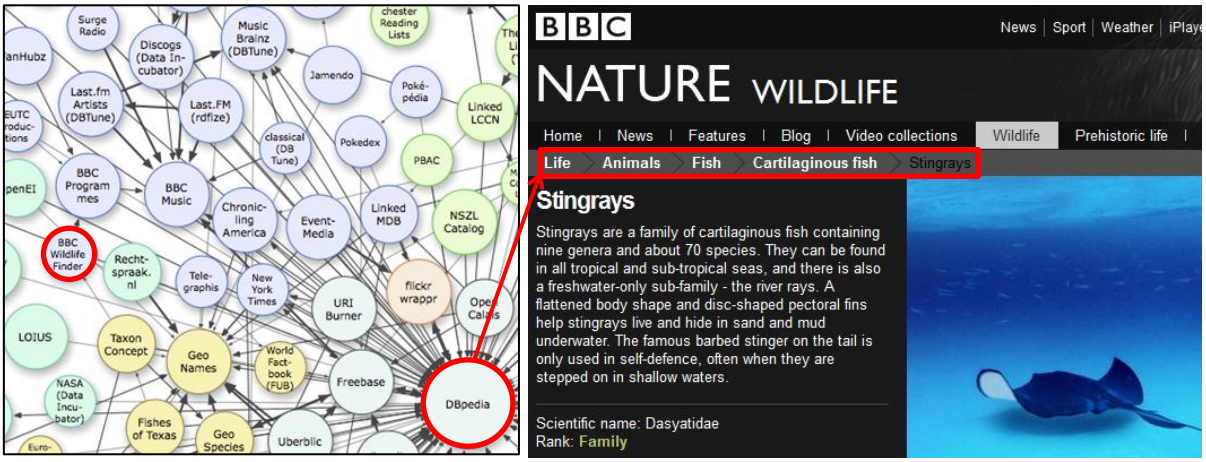


Figure 1 Quelques bases culturelles dépendant de DBPedia et un exemple de « BBC Programmes » dont l'organisation interne (Life>Animals>Fish>Cartilaginous Fish>Stingrays) est basée sur des connaissances issues de DBPedia.

En s'appuyant sur DBPedia les institutions culturelles voient désormais s'offrir à elles la possibilité d'ouvrir tout ou partie de leur catalogue et de l'augmenter en même temps avec de nouveaux moyens d'accès. L'expérience "Picture Book Mashup" du Brooklyn Museum associe DBPedia au catalogue du musée pour créer un album interactif et structuré de la collection. De son côté, la collection complète du Musée d'Amsterdam[1] est disponible sur le « Linked Open Data » (5 millions de triplets RDF décrivant plus de 70 000 objets du patrimoine culturel lié à la ville d'Amsterdam), fournissant des liens vers un thésaurus (AATNed), une liste d'artistes (ULAN), une base de lieux (Geonames) et les ressources de DBPedia pour en enrichir la structuration et l'exploitation. Citons encore le « Museum Finland » dont l'intégration du catalogue à d'autres bases permet notamment la navigation en anglais dans une collection finlandaise...

Des problématiques et exemples similaires existent tant pour les bibliothèques que les archives comme l'INA par exemple. Chaque lien créé entre les bases dégage un accès supplémentaire à la collection permettant aux utilisateurs et à leurs applications d'y entrer et d'en sortir selon autant de nouveaux parcours. La constellation de la Figure 2 témoigne non seulement de ce que DBPedia joue un rôle pivot vis-à-vis de nombreuses autres bases mais aussi et surtout qu'au-delà, elle offre, en définitive, une source de données utiles à une immense variété d'applications sur le Web, et ce quels que soient les domaines.

[1] http://semanticweb.cs.vu.nl/lod/am/

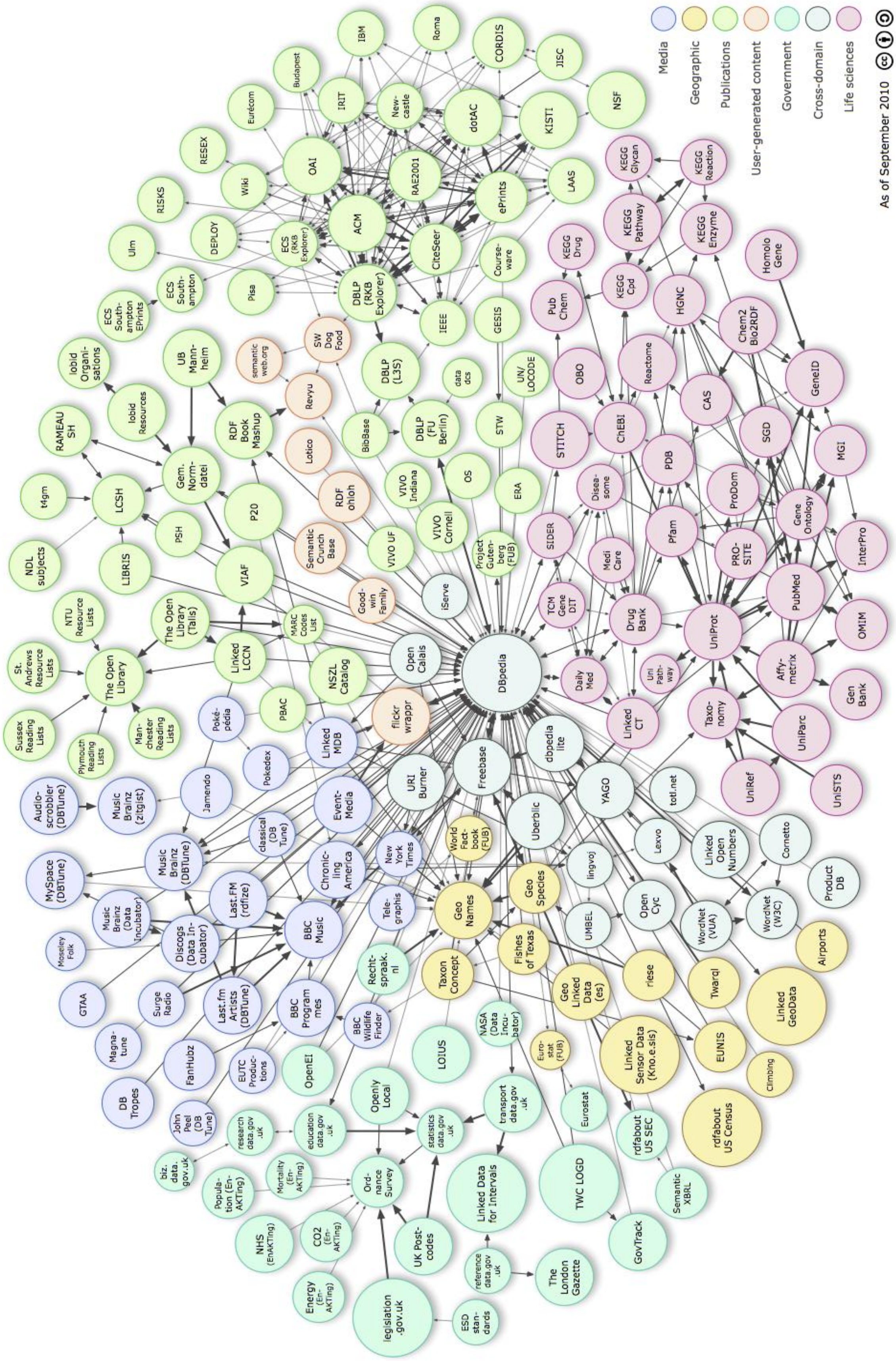


Figure 2 Représentation des sources des données du LOD (Linking Open Data). On notera que DBPedia est très nettement au centre et permet ainsi à de nombreuses applications de se développer. Ce schéma permet d'appréhender le réseau au sein duquel Dbpedia joue d'ores et déjà un rôle de référence.

Les utilisateurs de DBPedia incluent donc, entre autres, aussi bien les détenteurs d'autres jeux de données dont la valeur s'accroît dès lors qu'ils pointent vers DBPedia (principe fondateur du Web de données basé sur les externalités positives), que les entreprises dont les développeurs créent des applications consommant des informations destinées à leur permettre de répondre aux besoins de leurs utilisateurs, sans compter les gestionnaires de référentiels en tous genres (classifications bibliothéconomiques, thésauri, etc.), susceptibles d'y puiser de nouveaux descripteurs partagés, à l'échelle du Web.

Pourquoi un DBpedia Français ?

Dans sa version Française, Wikipédia compte très précisément (au moment où nous écrivons ces lignes) 1 114 361 articles traitant de culture[2], géographie, histoire, sciences, divertissement, société ou technologie. Malheureusement, DBPedia, centré sur la version anglaise de Wikipédia, ignore par conséquent les articles en français ne bénéficiant pas d'équivalents anglais et n'en expose donc pas les données.

Ainsi, le célèbre quatuor « Les Frères Jacques » n'est pas identifié dans DBPedia car l'article décrivant ces artistes est absent de la version anglaise. Ce défaut signifie que pour toutes les ressources culturelles francophones dans le même cas, aucunes données n'existent dans DBPedia qui permettraient de les référencer, de les indexer, de les interroger, etc.

Nous pensons que la publication de telles données francophones est une prérogative régalienne de l'État favorisant le maintien d'un accès public à la culture. Elle faciliterait également la préservation et le rayonnement, au plan international, des collections culturelles françaises. Investir l'espace numérique à cette fin étant une nécessité.

Partant, il est extrêmement important de garantir la pérennité d'un site qui publierait de telles données afin d'assurer la stabilité des liens qu'il autoriserait (dans tous les sens du terme) ; la stabilité constituant la condition *sine qua non* pour que d'autres applications et d'autres jeux de données fondent leur déploiement et leur expansion sur cette base.

Pourquoi maintenant ?

A ce jour, l'Allemagne, la Grèce et la Corée ont toutes trois mis en place des versions de DBPedia dans leurs langues nationales respectives[3]. La version française est actuellement absente, ce qui retarde l'intégration de nombreuses collections françaises, comme francophones, au Web de données.

---

[2] Selon le rapport de Wikimédia France pour le rapport au parlement de la Délégation générale à la langue française et aux langues de France, en date du 31 mai 2011, les pages culturelles représentent à elles seules 39 % des pages de l'encyclopédie Wikipédia (Culture et arts, 17 %, Société et sciences sociales, 12 %, Histoire, 10 %) : http://upload.wikimedia.org/wikipedia/meta/b/b3/Rapport_de_WMFR_sur_l%27utilisation_de_la_langue_fran%C3%A7aise.pdf

[3] http://wiki.dbpedia.org/Internationalization/Chapters?v=ygn (dernière modification 24/10/2010 10:30:23)

Il faut aussi noter l'importance d'une gouvernance publique en première ligne sur ces questions pour éviter une prise de contrôle par des acteurs ne partageant pas le souci de l'accès ouvert aux données relevant de l'utilité publique.

Enfin, en s'inscrivant dans un calendrier à court terme, l'initiative pourra bénéficier de la dynamique de Datalift[4], projet financé par l'Agence Nationale de la Recherche (ANR), dont le but est de développer une plateforme de publication et d'interconnexion des jeux de données liés (*Linked data*) sur le Web (outre la publication de données gouvernementales, la spécificité de Data Lift étant la prise en compte des besoins des acteurs de la culture et de la recherche). Datalift propose ainsi un ensemble d'outils facilitant ce processus de publication. Un jeu de données extrait de Wikipédia.fr pourrait donc d'ores et déjà bénéficier de l'infrastructure de cette plateforme pour devenir DBPedia.fr.

Un agenda de 18 mois :

- 7 mois de comparatif, d'extraction et prototypage :
  - Étude et adaptation des outils d'extraction de la version anglaise
  - Premières expériences d'extraction de données de Wikipédia.fr
  - Configuration d'une plateforme de publication sur la base de Datalift
  - Preuve de concept et validation de la chaîne technique
  - Nettoyage des données et amélioration de la qualité d'extraction
  - Mise en ligne d'une version de test
- 7 mois de déploiement, tests et passage à l'échelle
  - Validation avec des utilisateurs choisis
  - Mise en place d'une infrastructure opérationnelle
  - Mise en place du site public et de sa documentation
  - Conception et réalisation des interfaces utilisateur
  - Tests de volumétrie de données
  - Tests de charge d'interrogation
  - Passage en mode de production
  - Ouverture au public et communication
- 4 mois de pérennisation, stabilisation et passage en maintenance
  - Automatisation et stabilisation de la chaine d'extraction et publication.
  - Maintenance et intégration des retours d'utilisation
  - Étude des problèmes d'évolutions et de versions
  - Activités de transfert et pérennisation

Investissement INRIA : *38 303,47 EUR HT*

- Conception, suivi et direction du projet:
  7% temps CR1, Fabien Gandon (responsable scientifique de la proposition, co-responsable équipe de recherche Edelweiss et contact Datalift), sur 18 mois
  9 840,83 €
- Expertise technique et suivi des extensions:
  7% temps CR1, Olivier Corby (expert RDF/SPARQL, responsable scientifique équipe de

[4] Projet ANR-10-CORD-009 http://datalift.org/fr/

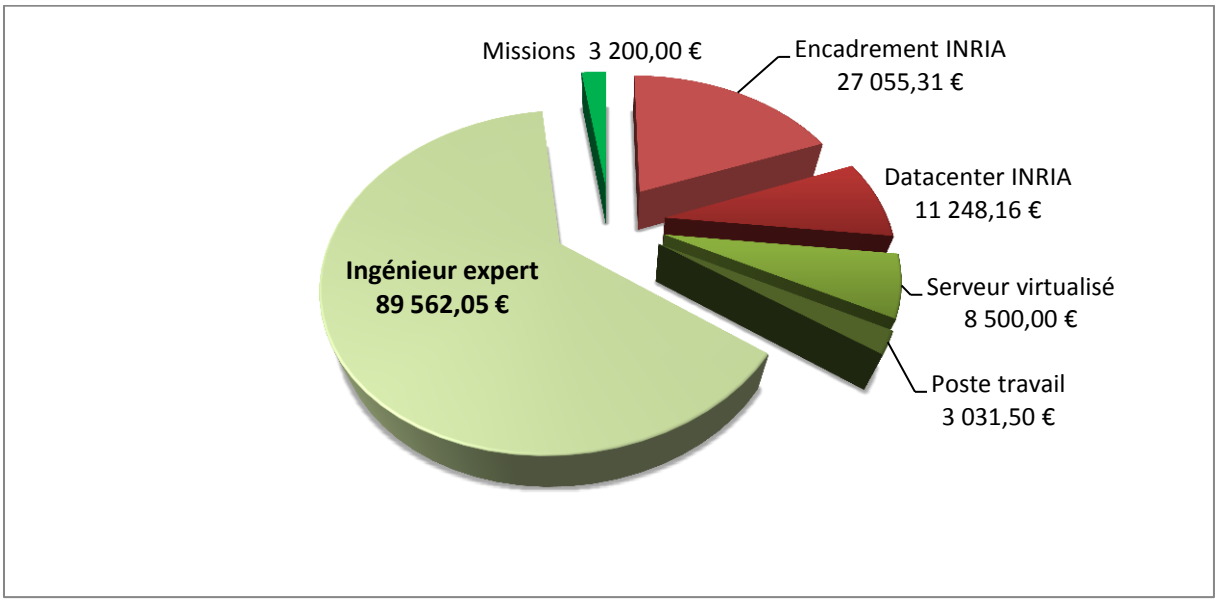
recherche Edelweiss, et concepteur de la plateforme CORESE/KGRAM), sur 18 mois
9 840,83 €

- Expertise de la plateforme de publication Datalift:
  10% temps Post-Doc, Serena Villata (contact technique Datalift pour l'équipe Edelweiss), sur 18 mois
  7 373,65 €
- Expertise en architecture du Web, philosophie et organisation de collections culturelles :
  10% temps Collaborateur Extérieur INRIA, Alexandre Monnin (Paris 1/IRI/CNAM) , sur 18 mois
- Mise à disposition d'un emplacement du Datacenter VMWare INRIA, et installation, maintenance et intégration du serveur DELL - PE R610 dans le cluster pour répondre aux critères de disponibilité et de fiabilité du service sur 18 mois (intégration physique du serveur dans le Datacenter, Intégration logique du serveur dans l'infrastructure VMware, Supervision du serveur, Contribution à la mise au point de la plate-forme, Accompagnement et suivi de la plateforme).
  7% temps IR1, contact service des moyens informatiques de d'INRIA (SEMIR, DSI)
  11 248,16 €

Aide demandée : *104 293,55 EUR HT*

- Achat du serveur et de son environnement logiciel : 1 config serveur DELL - PE R610 - bi-pro - 64 GB RAM + licences VMware Sphère 4 Enterprise+ licence Red Hat + garantie 5 ans J+1
  8 500 EUR HT
- MacBook Pro 13,3'' LED/Core i7 2-Core à 2,70 GHz/8 Go (2x2)/250 Go Serial SSD /SuperDrive 8x DL/Intel HD Graphics 3000/Bluetooth 2.1 EDR/AirPort 802.11n/Go Eth, moniteur 27'' LED Display
  3 031,50 EUR HT
- 18 mois d'ingénieur expert de recherche, expérience > 5 ans
  89 562,05 EUR HT
- Missions Paris – Nice Sophia Antipolis:
  3 200,00 EUR HT

Total du montage financier : *142 597,02 EUR HT*



Résumés biographiques des intervenants :

Fabien Gandon (vice responsable équipe de recherche Edelweiss, porteur de cette proposition) est chercheur en informatique dans l'équipe Edelweiss de l'Inria de Sophia-Antipolis. En 2002, son doctorat en informatique est au carrefour entre l'intelligence artificielle distribuée et les systèmes d'information pour la gestion des connaissances, explorant les formalismes basés sur les ontologies informatiques et les architectures multi-agents pour matérialiser et gérer des mémoires organisationnelles sous la forme de web sémantiques d'entreprise. En post-doc à l'Université de Carnegie Mellon (CMU) il supervise le projet myCampus exploitant les formalismes du web sémantique et les services web pour intégrer la captation du contexte et le respect de la vie privée dans les accès mobiles aux services en ligne. En 2008 il défend une HDR sur la gestion des connaissances à base de graphes. Fabien s'intéresse maintenant aux formalismes et aux architectures du web sémantique pour assister le cycle de vie des communautés en ligne dans leurs interactions et leur gestion de l'information. Ce domaine l'a amené à étudier l'analyse des réseaux sociaux et du social tagging à travers les langages du web sémantique. Il a participé à plusieurs projets européens et nationaux, et est membre de plusieurs comités de programme pour des journaux, conférences et ateliers. Fabien est co-auteurs de nombreux articles et enseigne le web sémantique à l'EPU de Nice. Enfin, il est membre de plusieurs groupes de travail du Consortium World Wide Web (W3C), où il a contribué à des standards tels que GRDDL et RDFa. Voir http://fabien.info

Olivier Corby (responsable équipe de recherche Edelweiss) a un doctorat en informatique de l'Université de Nice-Sophia-Antipolis et effectue des recherches sur les environnements de développement logiciel pour la modélisation des connaissances et sur les représentations XML pour la gestion de mémoires organisationnelles. Ses thèmes principaux de recherche sont l'ingénierie des connaissances et leWeb sémantique et il est notamment le concepteur et le développeur principal du moteur web sémantique Corese / KGRAM, une machine abstraite pour manipuler les graphes de connaissances. Olivier a publié de nombreux articles dans des revues, des livres ou des conférences. Il est membre de plusieurs comités de programme de conférences ou ateliers. Olivier participe aux groupes de travail W3C sur RIF, RDF et SPARQL. Il est aussi responsable d'un cours de Master sur le Web sémantique à l'EPU de Nice - Sophia Antipolis où il a également enseigné l'ingénierie des connaissances.

Serena Villata (Post-doc dans l'équipe Edelweiss pour le projet Datalift) a été assistante de recherche au Département d'Informatique de l'Université de Turin et a obtenu son doctorat en informatique dans le même département en 2010, avec une thèse sur la théorie de l'argumentation, et en particulier sur la méthodologie des méta-argumentations appliquée aux systèmes multi-agents. Serena s'intéresse principalement à l'intelligence artificielle, aux systèmes multi-agents et la théorie de l'argumentation, en particulier sur la coopération et les normes, sur le dialogue et les ontologies. Serena est co-auteur de quatre articles de journaux, et de nombreux articles de conférences. Serena est aussi présidente et organisatrice d'ateliers internationaux. Enfin Serena est maintenant en charge de l'implication de l'équipe Edelweiss (INRIA) dans le projet Datalift visant à développer une plateforme pour publier et interconnecter des jeux de données sur le web de données.

Alexandre Monnin (Collaborateur Extérieur Edelweiss) est doctorant en philosophie à l'université Paris 1 Panthéon-Sorbonne, résident à l'Institut de Recherche et d'Innovation (IRI) du Centre Pompidou en qualité de chef de projet métadonnées, et doctorant associé au CNAM (équipe DICEN). Il s'intéresse en particulier au tagging (projet NiceTag), aux ontologies et au Web social et sémantique. Son travail de thèse entend illustrer l'idée d'une philosophie du Web, par la mise en évidence des concepts philosophiques à l'œuvre au sein même de l'architecture du Web. A cet égard, il s'attache à produire une théorie des principes fondamentaux du Web susceptible d'enrichir d'autres démarches, notamment dans le domaine de l'ingénierie. Dans le cadre de ses fonctions à

l'IRI, Alexandre est également en charge du projet d'enrichissement des ressources du portail HDA (Histoire de Arts) par le tagging sémantique, projet issu d'une convention entre le Ministère de la Culture et l'IRI. Alexandre a travaillé à l'École nationale des chartes, comme Conservateur à l'Urfist de Paris-École des chartes et il fut doctorant invité à l'INRIA et au LIRMM de Montpellier. Enfin, il a donné des cours dans divers Master 2, notamment sur le Web sémantique.