Faire parler les données de la recherche grâce au Web sémantique : le projet VIVO

Marie Dominique Heusse

Faire parler les données sur la recherche grâce au Web sémantique : le projet VIVO

Marie-Dominique Heusse

1. IRIT, Université de Toulouse

RESUME. VIVO est un outil provenant des technologies du web sémantique, open source, pour repérer et caractériser l'activité de recherche d'un établissement. Il permet de réutiliser des informations ou des données présentes dans plusieurs bases de référence externes et les différentes bases de signalement bibliographique. L'objectif est de rassembler et exploiter quatre ensembles de données pertinentes sur l'activité de recherche : les structures de recherche, les auteurs, les productions de recherche et les moyens.

ABSTRACT. VIVO is an open source tool from semantic web technologies, used to identify and characterize the research activity of an institution. It allows the reuse of information or data existing in several external reference bases and different bibliographic databases. The aim is to collect and use four sets of relevant data on research activity: research structures, authors, research outputs and means.

MOTS-CLES : VIVO, DONNEES SUR LA RECHERCHE, WEB SEMANTIQUE

KEYWORDS: VIVO, RESEARCH DATA, SEMANTIC WEB

Les systèmes d'information recherche[1] sont très rares en France dans les établissements d'enseignement supérieur (Delemontez, 2017[2]). On entend par là des outils qui mettent en perspective des données sur les structures de recherche, sur les équipes, sur les moyens et sur les résultats des recherches menées afin de mieux connaître l'activité de recherche, mieux la gérer et la valoriser. Il s'agit donc de connecter des données qui sont éclatées entre différentes sources pour améliorer la connaissance de cette activité, et également pouvoir répondre aux besoins d'information des tutelles mais aussi des instances d'évaluation ou de financement. Ces outils peuvent en outre offrir une plus grande visibilité publique de l'activité de recherche en permettant le repérage de compétences et d'expertises. C'est le cas de VIVO, qui a été choisi par l'Université de Toulouse.

1. Le contexte

C'est celui de la construction de l'université de Toulouse, qui a vocation à rassembler de façon plus structurée qu'aujourd'hui les établissements membres de l'actuelle communauté d'universités et d'établissements (Comue). La recherche menée au sein des quelques 150 unités de recherche du site est déjà très largement inter-établissements : ainsi, un grand nombre d'unités mixtes de recherche (UMR) associent un (voire plusieurs) organismes de recherche nationaux (Établissements publics à caractère scientifique et technologique, EPST) à plusieurs universités ou écoles d'ingénieurs.

La mise en œuvre du « grand établissement » Université de Toulouse est tout particulièrement l'occasion de construire une stratégie scientifique de site : il s'agit d'élaborer une politique scientifique partagée pour consolider les secteurs d'excellence, et également développer l'interdisciplinarité.

Pour autant, la visibilité globale de l'activité de recherche est aujourd'hui assez faible. La perception des points forts s'appuie sur la connaissance de l'activité et de la performance des différents laboratoires, dans une logique verticale ; mais il est plus difficile d'avoir une perspective transversale, par grand domaine scientifique par exemple, surtout si ce domaine est commun à plusieurs établissements.

Le projet d'observatoire de la recherche porté par l'Université de Toulouse doit permettre de caractériser l'activité scientifique, de réunir les données pertinentes pour en rendre compte, de proposer des interfaces de visualisation et de cartographie, et de rendre ces informations accessibles aux différents acteurs concernés, ainsi qu'au grand public.

[1] http://www.eurocris.org/

[2] Les bibliothèques universitaires face aux systèmes d'information recherche: nouveaux outils, nouveaux rôles? Renaud DELEMONTEZ, Lyon, Enssib, 2017

2. Le projet

L'objectif est donc de se doter d'un outil permettant de rassembler et traiter l'ensemble des données qui peuvent concourir à cette visibilité. Très rapidement l'utilisation d'une solution Web sémantique s'est imposée : elle permet d'exploiter toute une série de sources de données externes de structures - et de formats - différents, et d'utiliser le potentiel que permettent les données liées. Une fois ce premier choix effectué, celui de l'outil lui-même a été rapide également : au sein des *triplestores* (bases de données conçues pour gérer des données en format RDF) qui pouvaient être candidats, VIVO[3] présentait l'avantage décisif d'être doté d'une ontologie correspondant assez exactement à nos besoins.

Ce n'était pas sa seule supériorité : développé en *open source* par l'université de Cornell aux États-Unis, VIVO est porté aujourd'hui par un consortium actif d'universités, et compte près de 150 instances installées dans 26 pays, dont une vingtaine en Europe (2 pour le moment en France : à l'INRA et le projet de Toulouse).

3. Les données

Pour chaque « famille » de données, il existe plusieurs sources possibles, parfois concurrentes ou redondantes, parfois spécifiques et complémentaires des autres. Le candidat/source idéal devrait réunir les qualités suivantes en matière de données : normalisées, homogènes, fondées sur des référentiels, des identifiants uniques et pérennes. C'est parfois très loin d'être le cas, ce qui oblige à divers traitements pour améliorer cette qualité.

- Les structures (Unités de recherche, Établissements, EPST…). L'utilisation de jeux de données publiés en *Open Data* par le Ministère de l'enseignement supérieur – jeu de données sur les écoles doctorales, puis Répertoire National des structures de recherche (RNSR) est apparu comme la meilleure solution : ces sources sont contrôlées, dotées d'identifiants nationaux et d'une indexation sujet, les jeux de données disposent d'une API. Cependant, certaines informations importantes pour le site n'y figurent pas, par exemple le rattachement des unités à six Pôles de coordination de la recherche, qui représentent les bases privilégiées pour l'analyse de l'activité de recherche. En outre, on constate parfois un retard dans la mise à jour du RNSR par les établissements et dans ce cas une source locale, le Guide de la Recherche, est utilisée pour compléter ou corriger l'information si la donnée dans le RNSR n'est pas à jour. Le nombre total de structures recensées est d'environ deux cents.

Les personnes (enseignants-chercheurs, chercheurs, BIATSS publiants), soit environ 5000 personnes: pour pouvoir repérer l'ensemble des chercheurs de chaque unité de recherche, le choix a été fait d'utiliser les fichiers des personnels établis lors des évaluations par le Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur (HCERES). Les champs disponibles sont : nom, prénom,

[3] Vivoweb.org

date de naissance (utilisée pour identifier les homonymes), statut, discipline, établissement de rattachement, date d'arrivée dans l'unité. Un identifiant est rajouté : le choix s'est porté sur IdRef[4], le référentiel maintenu par l'Agence bibliographique de l'enseignement supérieur (ABES), car c'est le seul à pouvoir être attribué à la totalité des auteurs[5] ; en outre l'ABES a développé des *webservices* qui permettent un alignement entre IdRef et d'autres identifiants comme Orcid ou IdHal : IdRef constitue bien un système pivot. Comme pour les structures, on constate parfois des problèmes de qualité des données dans les fichiers HCERES : nom ou prénom mal orthographié, informations manquantes. Ces erreurs ou lacunes rendent plus difficile la recherche de l'identifiant dans la base IdRef mais, lorsque la recherche a abouti, elle concourt à améliorer la qualité du fichier des auteurs.

- Les moyens : à la différence de certains systèmes d'information recherche à l'étranger, il n'est pas prévu actuellement de gérer l'information sur les budgets. Les données sur les moyens concernent deux types:
    - Les contrats. Ces informations sont présentes dans de multiples sources : bases de données Contrats tenues à jour par chaque établissement, par les organismes de recherche, par la Comue pour les contrats européens ; en outre l'information est désormais mentionnée dans certaines bases bibliographiques comme le Web of Science. Un travail de rapprochement et de dédoublonnage entre ces diverses sources de données est donc nécessaire.
    - Les équipements. L'objectif est de valoriser les grands équipements ou équipements rares présents sur le site, généralement très mal connus sauf de leurs utilisateurs directs. La solution passe par la création d'une base locale contenant les champs nécessaires.
- Les productions scientifiques peuvent être extrêmement diverses et ne prennent pas toutes la forme d'une publication même si celle-ci constitue le modèle dominant :
    - Les publications regroupent elles-mêmes des formes différentes : livres, chapitres de livres, thèses, divers types d'articles, *proceedings*, *working papers*, communications à des colloques, etc.
    - Les brevets.

---

[4] Identifiants et référentiels pour l'enseignement supérieur et la recherche, https://www.idref.fr/

[5] Selon une évaluation faite début 2017, moins de 15% des enseignants-chercheurs et chercheurs de l'université de Toulouse ont un identifiant ORCID

- Les corpus, bases de données, jeux de données (*data papers*), réalisations informatiques ou instrumentales.
- D'autres formes de productions de recherche sont importantes dans certaines disciplines ALSHS : expositions, installations, réalisations artistiques, multimédia.

Sur le plan des données, cette profusion de formes de production scientifique se traduit par un foisonnement de sources. Certaines bases bibliographiques sont spécialisées dans un type de documents (par ex. livres et thèses pour la base du Système universitaire de documentation (Sudoc), d'autres dans un ensemble de disciplines, d'autres encore par le statut *open access* des documents signalés. De ce fait, une publication peut être présente dans plusieurs sources, mais sans être forcément signalée de la même façon, ni au même moment, ni avec le même identifiant… En outre, les outils de caractérisation disciplinaire diffèrent également, chaque base ayant son propre vocabulaire pour l'indexation des publications. Pour les corpus ou les données sur les manifestations, il n'existe pas de sources externes structurées, ni de système d'identifiants construits ; les informations sont souvent gérées directement (et seulement) par le chercheur ou l'équipe en charge du projet, avec une très grande variété de formats mais pas toujours une garantie d'interopérabilité. Cela va conduire à un travail d'identification, mais aussi de normalisation, en reprenant dans toute la mesure du possible les formats et les référentiels professionnels utilisés par les services d'archives et les musées.

4. Les traitements : ontologies et dédoublonnage

Comme toute application web sémantique, VIVO utilise des ontologies pour représenter les données. L'ontologie intégrée de VIVO réutilise la terminologie d'ontologies largement connues (*SKOS, bibo, foaf, dublin core, event, geo, vcard*, etc.). Ses modèles fournissent un ensemble de types (les classes) et de relations (les propriétés) pour représenter les différents types de données gérées par l'outil : les chercheurs, les structures, les moyens, les productions de recherche. Des modèles d'ontologie associant classes et propriétés ont été définis pour chaque grand type de données. L'adaptation de l'ontologie de VIVO, né dans le contexte d'universités américaines, à notre situation nécessite quelques ajustements, par ex. pour représenter les EPST.

Le *mapping* (alignement) des ontologies consiste à faire correspondre les champs spécifiques de chacune des sources de données avec les classes de l'ontologie de VIVO. Les alignements ne se font pas forcément terme à terme : dans certains cas, les meta-données sont plus détaillée dans la source de données, dans d'autres certaines classes de l'ontologie n'ont pas de correspondance dans la source.

Une même donnée peut être présente dans plusieurs sources. C'est particulièrement le cas pour les sources bibliographiques, le même document pouvant être décrit dans une base de données commerciale (ex. le Web of Science), une base en archive ouverte, ou un outil collectif national comme le Sudoc. L'utilisation d'algorithmes de rapprochement est nécessaire, et ces algorithmes fonctionneront d'autant mieux que les données concernées s'appuient sur des référentiels et des identifiants – ce qui n'est pas toujours le cas. Ils imposent

également la rédaction de règles de gestion : lorsqu'une même notice est repérée dans deux sources différentes, laquelle est privilégiée, quel ordre est précisé ?

5. Les résultats

L'interface web de VIVO propose des fonctions de recherche, de navigation et de présentation des données. Elle offre également la visualisation en RDF de toutes les données contenues dans la base, et constitue donc un outil de publication de ces données. L'utilisation du moteur SolR permet d'améliorer les fonctionnalités de la recherche en générant des facettes à partir des méta-données présentes : par exemple en affinant une recherche par concept, ou par type de document. L'objectif est de permettre de décrire finement les compétences des chercheurs, leur environnement, leurs interactions et leurs réseaux, et d'offrir des possibilités de recherche telles que :

- Repérage d'experts (ex. évaluation de projets, actions de médiation scientifique),
- Repérage des structures de recherche (labos, Écoles doctorales, fédération, pôles, etc.) et leurs relations,
- Repérage des co-signatures, des coopérations internationales,
- Repérage des équipements scientifiques,
- Autres productions de recherche : expositions, installations, évènements.

Outre ces fonctions de type « découverte », il est prévu également un outil tableau de bord qui permettra de réaliser des rapports. Pour notre site, les informations et indicateurs attendus sont par exemple :

- Repérage des enseignants-chercheurs et chercheurs par secteur disciplinaire et sous-domaine,
- Repérage des publications (nombre total, par thèmes, type, etc.),
- Nombre de thèses (par an, par discipline, etc.),
- Nombre de brevets,
- Repérage des projets (ERC, H2020, ANR, PCRD, etc.),
- Évolution des thématiques de recherche, repérage des «signaux faibles», des thématiques émergentes.

VIVO propose également des interfaces cartographiques, qui permettent des représentations graphiques des données de publications :

- *Co-author network*
- *Co-investigator network*
- *Map of Science.*

Si l'objectif premier du projet est de mettre en œuvre un dispositif pour caractériser l'activité de recherche du site, VIVO est aussi conçu comme un "outil de découverte", voire de mise en réseau et de repérage de compétences (*research*

profiling system [6]). Au démarrage du projet, l'acceptation de cette dimension par la communauté des chercheurs nous a semblé devoir être questionnée, alors qu'elle ne pose visiblement pas de problème dans les universités anglo-saxonnes. Les réponses à ce sujet, qu'elles résultent de contacts informels ou de circonstances plus institutionnelles comme la Commission recherche d'une université, ont toutes montré une volonté d'ouverture et un désir de visibilité – qui sont déjà manifestes dans l'usage des réseaux sociaux académiques, mais avec la double différence qu'il s'agit dans le deuxième cas de démarches individuelles, et d'outils venant d'initiatives à caractère commercial[7]. La question de l'appropriation reviendra lorsqu'il faudra arbitrer sur le niveau d'intervention que l'on attribuera aux acteurs de la recherche sur les données : VIVO est conçu pour permettre aux chercheurs d'effectuer des modifications et compléments sur les données qui les concernent. Cette possibilité est assurément une richesse, mais elle peut constituer aussi un facteur de risque en matière de qualité des données : un équilibre délicat et nécessaire doit être trouvé.

Conclusion

Les opportunités et les risques du projet VIVO sont les deux faces d'une problématique commune. Un des facteurs décisif du succès est la qualité des données, ce qui suppose un travail très attentif sur les identifiants et les référentiels. Mais le pari repose aussi sur l'utilisation de sources de données qui n'ont pas toutes été prévues pour une réutilisation dans une logique de données liées. Et même lorsque c'est le cas (par exemple le RNSR, la base IdRef), on constate que la qualité des données n'est pas toujours optimale…

Un enjeu pour l'avenir, outre le nécessaire développement de ce type d'outils en Europe, sera d'en faire des démonstrateurs de l'importance et des utilisations potentielles des données sur la recherche.

---

[6] https://en.wikipedia.org/wiki/Comparison_of_research_networking_tools_and_research_profiling_systems

[7] *#DeleteAcademicSocialNetworks?* Les réseaux sociaux académiques en 2016 http://urfistinfo.hypotheses.org/3033