Archives ouvertes et bases de publications : exploration et analyse des sources de données pour la recherche et ses environnements

Lundi 23 mai 2016, 9h--17h
Institut de recherche et d'histoire des textes, 40 avenue d'Iéna, 75116 Paris

Appel à communications

L'évolution numérique majeure de la recherche scientifique et de ses impacts societaux, économiques et industriels permet maintenant d'avoir accès aux données scientifiques tels que les textes publiés dans des archives ouvertes, des revues ou des conférences ainsi que les données d'expérimentation ou les résultats de simulation, mais également, et c'est fondamental, aux données d'usage des différents services qui se mettent en place.
Le développement de méthodes d'analyse de ces données, ou l'application de méthodes existantes, est une étape inévitable de cette mutation. De la même manière que le monde du business a intégré avec succès les méthodes d'analyse de ses données, le monde académique envisage maintenant les nombreuses possibilités offertes par ces méthodes sur les données scientifiques. Ces méthodes couvrent tout le processus de valorisation des données, leur préparation, leur analyse (apprentissage, fouille, statistiques, recommandation...) jusqu'à l'interprétation des résultats, ainsi que leur visualisation. Les enjeux pour les données de publication sont cruciaux par la valeur que ces méthodes peuvent ajouter au monde de la recherche. Ces enjeux peuvent concerner l'aide aux chercheurs, l'ouverture au grand public (avec la mise à disposition d'indicateurs transparents), ou encore la gestion de la recherche ou la prospective scientifique. Tels sont les constats formulés lors du colloque "Publication scientifique, innovation et services à la recherche" des 9 et 10 novembre 2015 à Meudon, organisé conjointement par l'ADBU, Couperin, EPRIST et la DIST du CNRS.

L'aide aux chercheurs peut prendre la forme de recommandations (e.g. quels articles concernent un sujet particulier pour constituer une bibliographie, en relation avec des requêtes similaires ? Quels collègues sont actifs sur ce sujet, et quels sont les co-auteurs dans le graphe de relations ? Qui sont les auteurs dont les publications sont souvent consultées ensemble dans des requêtes des usagers d'une plateforme comme HAL ? etc.). Elle peut aussi venir de la détection de tendances dans les mots clés enregistrés dans les publications d'un domaine, d'une meilleure compréhension des facteurs d'impact et de visibilité des travaux d'un chercheur, ou encore de la corrélation entre jeux de données disponibles publiquement pour permettre une plus large diffusion de ces derniers.

Le grand public pourrait disposer d'indicateurs transparents sur les activités de recherche d'un territoire (département, région, pays) en lien avec les données disponibles (e.g. les travaux sont-ils issus d'un laboratoire privé, public, ou une collaboration entre les deux ? Quelle est la source du financement ? Quelle est l'ancienneté de l'équipe sur le sujet ? etc.).

Enfin, la gestion de la recherche peut se voir suggérer, par la communauté des chercheurs analysant ces données, de nouveaux descripteurs qui permettent, par exemple, d'évaluer l'impact d'un appel à projet et de son orientation sur les publications qui ont suivi dans les années suivantes ; de comprendre les collaborations locales, nationales ou internationales ; de mieux situer la recherche publique et la recherche privée (en termes de sujets, de collaborations, de relations internationales, etc.) ; ou encore de situer les laboratoires entre eux selon les domaines de publications, les conférences auxquelles ils participent ou les interactions entre auteurs.

L'objectif de cette journée organisée conjointement par l'Inria et le CNRS, est triple :

- Présenter des corpus de données réelles préparées et/ou annotées, permettant d'explorer et d'analyser les données de la recherche. Ces corpus évolueront selon les échanges de cette journée, puis seront mis à disposition dans le cadre d'un appel à projet ultérieur. Cette journée regroupera donc les chercheurs et les fournisseurs de services et de données scientifiques pour mieux comprendre ces données et comment les utiliser pour mettre à disposition des chercheurs, des équipes et des organismes de recherche des services à haute valeur ajoutée.

- Présenter des travaux (les communications retenues pour cette journée) permettant de mieux connaitre les interactions possibles entre le paysage actuel de la recherche en analyse de données et celui des données de la recherche. Les présentations auront pour objectif d'expliquer ces travaux et d'en dessiner une prospective sur des applications possibles aux données de la recherche.
- Présenter un appel à projet, en cours de réflexion, autour de ces données. Les participants et les travaux présentés enrichiront les thèmes de l'appel afin d'assurer la meilleure adéquation avec les possibilités offertes par l'analyse de données.

Le principal corpus présenté lors de cette journée, et qui sera au centre de l'appel à projet à venir, concerne les données de HAL. Il représente environ 300 000 articles, liés à plus d'un million de notices métadonnées. Ce jeu de données sera téléchargeable pour être utilisé localement. On pourra également considérer les extractions faites à partir des pdf comme les images, les figures d'expérimentations, etc. Les données d'usage (consultation des articles, pages auteurs, etc.) seront également présentées et mises à disposition dans un cadre éthique approprié.

Nous pourrons aussi considérer les données suivantes (et les présenter, selon confirmation des intervenants) :
* les données ISTEX représentent plus de 16 millions d'articles de collections rétrospectives couvrant tous les domaines de la littérature scientifique. Des sous-corpus peuvent être construits et extraits à travers une API (https://api.istex.fr/documentation/).
* de façon plus générale, des corpus d'étude ou des données primaires de la recherche mutualisés au sein d'entrepôts de données nationaux ou internationaux.

Les communications attendues sont liées aux questions d'analyse de données de manière générale (constitution des corpus, apprentissage, fouille, statistiques, recommandation, visualisation, etc.). Elles pourront être généralistes (présenter un domaine, un état de l'art, une vision) ou ciblées (des cas d'études ou des applications sur, par exemple, des données scientifiques, des données textes, des graphes issus de réseaux sociaux... la liste n'est pas restrictive). L'objectif étant de créer une dynamique en ouvrant le plus largement possible cette journée aux différentes équipes qui, par la suite, pourront répondre à un appel à projet autour de ces données de la recherche.

Modalités
Les propositions sont à envoyer à data4ist@inria.fr. Elles mentionneront le titre, les auteurs et leur affiliation, un résumé de 10 à 15 lignes et un court développement (entre 1 et 2 pages) reprenant ou référençant éventuellement des éléments déjà publiés. Elles pourront être rédigées en anglais ou en français.

Vendredi 1 avril 2016 : réception des propositions de communications
Mercredi 13 avril 2016: notification
Lundi 23 mai 2016: déroulement de la journée à Paris

Les communications retenues donneront lieu à une présentation de 15 minutes. À la fin de chaque session, une discussion générale se tiendra sur la base des présentations données afin de mieux préciser les possibilités envisageables sur les données de la recherche.

Lieu : Institut de recherche et d'histoire des textes, 40 avenue d'Iéna, 75116 Paris

Organisateurs de la journée :
- Patrice Bellot (Aix-Marseille Université - OpenEdition)
- Christine Berthaud (CNRS)
- Daniel Egret (PSL)
- Renaud Fabre (CNRS)
- Odile Hologne (INRA)
- Claude Kirchner (inria)
- Florent Masseglia (Inria)
- Jean-Marie Pierrel (Université de Lorraine)
- Laurent Romary (Inria)
- Ken Takeda (CNRS)