RAPPORT DE MISSION DIST

DÉPLACEMENT ELSEVIER LONDRES JEUDI 6 AVRIL 2017

« Les systèmes de recommandation et le data management »

Mission composée de :

Renaud FABRE, Directeur de l'Information Scientifique et Technique (DIST) du CNRS,

Dominique DUNON-BLUTEAU, Directeur adjoint scientifique à l'Institut National des Sciences Biologiques (INSB) du CNRS,

Claire FRANÇOIS, Directrice du département de l'offre de service à l'Institut de l'Information Scientifique et Technique (INIST),

Elsa VERBRUGGHE, Expert juridique IST numérique à la Direction des Affaires Juridiques (DAJ) du CNRS,

Vincent GIACOBBI, Stagiaire - chargé de mission sur les modèles économiques de l'IST à la DIST du CNRS.

CONTENU

PRÉAMBULE

Contexte des systèmes de recommandation et du data management

La recherche d'information produit des documents « recommandés » aux utilisateurs, en fournissant une réponse à une requête. Pour toute requête sur un mot-clé, les systèmes de recommandation ajustent les besoins des utilisateurs et des documents correspondants. Les recherches sur l'information scientifique et technologique (IST) sont fortement structurées et produites dans une communauté d'information : les utilisateurs sont facilement identifiés les documents appartiennent à de grandes bases de données spécialisées avec des vocabulaires indexés. Les communautés d'utilisateurs sont dynamiques et interactives et le contenu innovant prévaut. L'IST semble être une candidate prometteuse pour d'autres utilisations des systèmes de recommandation. Cependant, alors que la publication augmente rapidement, la conception actuelle des systèmes disponibles (basés sur de longues listes de synonymes) ne répond guère à la nécessité d'une recherche d'information sémantique à l'échelle requise par les communautés de chercheurs : leur principal besoin est d'identifier et de partager des contenus innovants intégrés dans les articles publiés en vigueur.

Une correspondance plus précise des besoins de l'utilisateur et du contenu du document est en cours dans la recherche pour une meilleure recommandation, qui prend deux directions complémentaires dans l'optimisation de la gestion des requêtes : l'une sécurise la qualité du mot-clé et l'autre est d'offrir un meilleur choix pour l'utilisateur. Le principal défi décisif est que le contenu du mot-clé doit s'adapter régulièrement aux changements novateurs journaliers. Il peut être atteint de différentes façons avec de nouvelles architectures des systèmes, comme par exemple, avec les nouvelles bibliothèques dynamiques numériques. Les approches dominantes soutiennent une réponse efficace à la surcharge d'information par un traitement efficient de l'information. Le nouveau *data management* permet de répondre à cet enjeux. Le croisement de données n'est qu'au début de sa révolution.

Le CNRS justifie son déplacement chez Elsevier le 6 avril dernier afin de se positionner sur ces deux enjeux majeurs.

VUE RAPIDE DU DÉPLACEMENT

Objet du déplacement

Rencontre cher Elsevier, leader mondial de l'édition scientifique, sur les systèmes de recommandation et le *data management*.

Objectifs

- Partager les nouveaux projets du CNRS et d'Elsevier dans les deux domaines précités,
- Entretenir des rapports constructifs et professionnels entre le CNRS et Elsevier,
- Visiter les nouveaux locaux d'Elsevier à Londres.

Lieu

À Londres dans les bureaux d'Elsevier.

Date

Jeudi 6 avril 2016.

Durée de la rencontre

10h30 à 14h30.

Membres de la délégation

- Renaud FABRE, Directeur de l'Information Scientifique et Technique (DIST) du CNRS.
- Dominique DUNON-BLUTEAU, Directeur adjoint scientifique à l'Institut National des Sciences Biologiques (INSB) du CNRS,
- Claire FRANÇOIS, Directrice du département de l'offre de service à l'Institut de l'Information Scientifique et Technique (INIST),
- Elsa VERBRUGGHE, Expert juridique IST numérique à la Direction des Affaires Juridiques (DAJ) du CNRS,
- Vincent GIACOBBI, Stagiaire - chargé de mission sur les modèles économiques de l'IST à la DIST du CNRS.

Personnes rencontrées

- Oliver DUMON, Managing Director, Research Applications and Platform chez Elsevier,
- Gaby APPLETON, Managing Director, Mendeley chez Elsevier,
- Alexander CRAIG, VP Software Development (Content & Data Technology) chez Elsevier ,
- Maya HRISTAKEVA, Lead Data Scientist , Mendeley chez Elsevier;
- Matt HOBY, Head of Data Science for Analytics chez Elsevier.

Programme

Time	Agenda item	Presenter
10.30 -10.45	Introductions and office walk-around	Gaby Appleton / Alexander Craig
10.45 – 11.30	Presentation of CNRS research objectives and challenges in French research	Renaud Fabre
11.30 – 12.15	Elsevier research strategy; RELX / Elsevier technology approach	Gaby Appleton & Oliver Dumon Alex Craig
12.15 – 12.45	Lunch break	
12.45 – 14.00	Demo of different technology applications (2 x 25 min) : - Recommenders - User Analytics	 Maya Hristakeva Matt Hobby
14.00	Wrap up and depart	All

OBJECTIFS GÉNÉRAUX : LES SYSTÈMES DE RECOMMANDATION ET LE DATA MANAGEMENT

Les systèmes de recommandation

Les systèmes de recommandation sont souvent résumés à un algorithme. La réalité est bien plus complexe. La recommandation existe depuis longtemps sous forme de citation ou de *peer review.* La révolution numérique en cours permet de croiser davantage les informations et de créer des flux plus importants comparativement à l'édition papier.

Ils peuvent se décliner sous différents aspects : citations, vues, partages, *peer review,* recherche de fonds, réseaux sociaux, propositions de travail, etc. Elsevier les regroupe sous quatre grandes catégories : lecture, connexion, possibilité de financement et possibilité de travail. Une des avancées majeures en la matière serait la possibilité de proposer des sujets de recherche pertinents sur la base des systèmes de recommandation. Ils permettent dans le cas du *peer review* par exemple d'intégrer « l'émotion » relative à un choix.

Le nouveau *business model* des systèmes de recommandation est stratégique. Les acteurs de l'édition auront une proportion plus importante de bénéfice à terme avec ces services et la création de flux.

Le data management

Le *data management*, ou gestion des données, est une nouvelle discipline qui met en valeur les informations comme les ressources numériques. Cette notion est très vaste. Elle prend tout son sens au regard de l'évolution des *big data*.

Le *data management* peut se décliner sous différents aspects : TDM, bibliothèques dynamiques numériques, stockage dynamique numérique, etc. Le *data manager* a ainsi en charge la conception et la mise en place d'architecture et de processus permettant de collecter, stocker, exploiter et sécuriser les données. Il est une aide à la décision et permet l'innovation. Les systèmes de recommandation s'inscrivent parfaitement dans cette logique. Le *data management* a des perspectives infinies.

SYNTHÈSE DES PRÉSENTATIONS

Presentation of CNRS research objectives and challenges in French research - Renaud Fabre

La DIST a présenté quelques les chiffres significatifs sur les publications du CNRS en 2014. Son directeur a rappelé que la part des publications françaises du CNRS entre 2000 et 2014 était en augmentation de plus de 14%. Un graphique d'indice d'activité relative permet de comprendre que l'activité de production dans Scopus est plus forte dans les domaines de publications tels que celui de la chimie ou encore celui des sciences des matériaux.

Les publications dans les journaux multidisciplinaires entre 2000 et 2014 sont comparables entre la France et l'Allemagne. La Chine et l'Inde sont également comparables en la matière.

Une analyse d'impact, montrant la fréquence des citations dans les publications françaises, illustre que, par exemple, les publications en sciences de l'environnement (2000-2014) ont une influence 2,12 fois supérieure aux publications mondiales du domaine sur la même période.

Il a également été rappelé les missions de la DIST et le catalogue des offres partagées en matière d'IST. Pour rappel, cinquante services d'IST sont destinés aux chercheurs (développés en majorité par l'INIST), regroupés en quatre catégories : se documenter, publier, analyser et être accompané.e.

Elsevier research strategy - Gaby Appleton & Oliver Dumon

Elsevier a présenté ses missions résumées ainsi « ouvrir la voie aux sciences avancées, à la technologie et à la santé ». Ils ont exposé également les caractéristiques de la productivité mondiale de la recherche. Ils ont précisé que les dépenses de R&D avaient augmenté de 3,4% en 2016. Selon eux, 70%-80% des instruments de recherche sont perdus et cela prend trop de temps pour quantifier. Pour passer de la recherche médicale aux pratiques cliniques, ils disent qu'il faut plus de 18 ans. C'est pourquoi ils souhaitent remodeler les pratiques de recherche par la technologie dans trois domaines : l'information, la data et la programmation ainsi que les logiciels de travail.

Ils ont également montré la carrière archétypale d'un chercheur qui passe du master, à docteur, à post-docteur, à assistant professeur, à professeur associé et à « professeur complet ». Ils souhaitent proposer des services adaptés au regard de l'évolution de la carrière et des besoins des chercheurs.

Ils identifient treize priorités dans le flux de recherche des chercheurs comme l'édition, l'écriture, la lecture, les réseaux sociaux, etc. A ce titre, ils ont présenté la nouvelle architecture (version bêta) de Mendeley, leur réseau social pour les chercheurs. Le nombre de profils est 6 millions de personnes. Ils croisent beaucoup de données dont celles de Scopus. A ce propos, Oliver Dumon a fait une proposition au CNRS : il est prêt à payer pour pouvoir utiliser le fichier des chercheurs du CNRS et/ou à créer un réseau dédié au CNRS. Les questions juridiques ont été soulevés mais restent en suspens et ferons l'objet d'un autre déplacement.

RELX / Elsevier technology approach - Alex Craig

Alex Craig a présenté les approches technologiques d'Elsevier. Il souhaite améliorer la prise de décision afin d'avoir de meilleur résultat en améliorant en amont le référencement. Pour cela, il distingue quatre axes : avoir une meilleure compréhension profonde des clients, avoir une meilleure utilisation des principaux contenus et des ensembles de données, avoir des analyses sophistiquées et avoir une technologie puissante pour le *big data* et le *machine learning*.

Il a également précisé des notions comme la filiation (TOP-DOWN/BOTTOM-UP), la qualité des données, la sensibilité des données et la puissance du *machine learning* ainsi que la bulle liée aux systèmes de recommandation.

Recommenders systems - Maya Hristakeva

Maya Hristakeva a commencé sa présentation en précisant qu'elle possède un blog sur la recommandation avec son équipe à l'adresse suivante : https://buildingrecommenders.wordpress.com . Elle en a profité également pour préciser ses publications (cf annexe).

Elle a illustré les différentes étapes (sans ordre) du périple du chercheur : « aidez-moi à publier rapidement », « aidez-moi à prendre des décisions éditoriales », « aidez-moi à exploiter mes données de recherche », « aidez-moi à écrire et à évaluer des articles », « aidez-moi à organisez mes textes », « aidez-moi à présenter mon travail », « aidez-moi à rendre l'évaluation par les pairs plus gratifiante », « aidez-moi à me connecter avec les bonnes personnes », « aidez-moi à rester à jour » et « aidez-moi à obtenir des financements ». Elle s'est arrêtée sur les trois dernières étapes.

Leurs systèmes de recommandation se distinguent en deux catégories : celle des articles de recherche et celle des autres chercheurs. Elle a présenté concrètement son système de recommandation dans ScienceDirect et dans la nouvelle version (version bêta) de Mendeley (« Articles for you », « People to follow », « Emails », « Carrers/Job opportunities », etc.).

Les chercheurs utilisent beaucoup Twitter pour recommander un article ou un auteur. A ce propos, la fonction d'actualisation de Twitter est la suivante : plus un article ou un auteur est cité, meilleur il est. La recommandation est un système dynamique.

Elle en a profité pour préciser comment se compose l'architecture et les flux de l'information et de la data. La recommandation est circonscrite chez Elsevier aux interfaces utilisateurs telles que Mendeley, Scopus ou encore ScienceDirect.

Il existe différents types de recommandation : implicite (basée sur les bibliothèques), liée à l'activité récente (basée sur les ajouts récents à la bibliothèque d'un utilisateur), liée aux intérêts des chercheurs (basée sur leurs tags) ou encore liée à la discipline (basée sur la discipline auto-identifiée). Dans trois types de recommandation, Maya Hristakeva présente sa modélisation (cf annexe).

Elle aborde également « l'évaluation » qui dispose d'outils : les tests d'utilisateur, les évaluations hors-lignes, la méthode appelée *A/B testing & Metrics* et le *Dogfooding* (utilisation des produits de l'entreprise par ses employés).

Enfin, elle propose un graphique sur la représentation du réseau social des sciences (cf annexe).

User Analytics - Matt Hobby

Matt Hobby a présenté une analyse des réseaux des utilisateurs. Sa présentation est au croisement du data management et des systèmes de recommandation.

Il a présenté une application supplémentaire sur ScienceDirect (version bêta) qui est une bibliothèque dynamique numérique sous forme de « constellation ».

Il en a profité pour expliquer la technologie utilisée qui se regroupe sous trois axes : l'extraction, la prédiction, et la disposition/attraction. Pour cela, ils ont besoin d'intégrer un *login*.

Renaud Fabre a posé la question de la recommandation *filtering* qui est un axe très important dans le cadre des bibliothèques numériques.

AVIS DES PARTIES PRENANTES

INSB - Oliver Dumon

Sur la présentation fournie par Elsevier :

Les présentations d'Elsevier ont été intéressantes. Le site Mendeley paraît équivalent à ResearchGate et il a été difficile d'identifier la valeur ajoutée de ce système. ScienceDirect existe depuis longtemps. Dans les deux cas ces applications sont utiles pour les chercheurs mais d'autres fournisseurs proposent des applications équivalentes. En revanche, pour ce qui est du développement chez Elsevier, il paraît très intéressant comme outil de « coaching » des chercheurs (aide à l'écriture, au funding et à l'élaboration des projets). Toutefois cette approche est à approfondir en termes de confidentialité des travaux réalisés à l'aide de ces plateformes mais également en termes d'interactivité entre les utilisateurs et les outils proposés (paramétrage, « recherche participative », recommandations, etc.)

Sur les suites à donner à ses présentations :

1- Clairement Elsevier souhaiterait enrichir sa base de Mendeley avec l'ensemble des chercheurs français pour consolider ses parts de marché au niveau mondial en attendant qu'un jour peut-être ce service devienne économiquement viable.

La question à poser est : que peut offrir Elsevier au CNRS si nous nous dirigeons dans cette direction ?

2- En ce qui concerne l'INSB, l'actualité est de finaliser l'analyse bibliométrique de sa production scientifique et de vérifier que les méthodes développées par Elsevier répondent à nos questions (non traité à Londres). De plus, en interaction avec eux comme nous l'avons fait à Londres, il s'agit de faire évoluer leurs approches méthodologiques.

En conclusion : des discussions internes au CNRS sont indispensables avant de donner suite.

L'implication de l'unité qui rédige le rapport :

Hors sujet – A ce stade je ne suis qu'un chercheur /utilisateur de l'INSB au mieux un CORIste.

INIST- Claire François

Sur la présentation fournie par Elsevier :

Sur la présentation d'Alexander Craig : utilisation des principes et techniques d'analyse de risque pour proposer une aide à la décision : si ce type d'outil peut faire gagner du temps, dans quels cas est-ce utile ? Comment se situe-t-il pour aider à la créativité ?

Les démonstrations « Network analytics data science » et « Recommender algorithms » montrent des développements intéressants mais basés sur des principes connus.

Sur les suites à donner à ses présentations :

Obtenir la présentation que nous n'avons pas eue : « Research data management » ?

Concernant les algorithmes de recommandation, obtenir une présentation des « dessous » de Research Gate, il sera utile de bien connaitre ces plateformes et surtout les enjeux juridiques concernant les documents déposés et leur ré-utilisation des « données utilisateurs ».

L'implication de l'unité qui rédige le rapport :

Un service que nous pouvons rendre à la communauté scientifique est de décoder les tenants et aboutissements des usages d'un outil comme Mendeley.

Comment préciser l'implication de l'INIST dans l'usage de Scival par le CNRS ?

DAJ - Elsa Verbrugghe

Sur la présentation fournie par Elsevier :

Les présentations étaient claires, agréables et m'ont permis de bien identifier les nouveaux services qu'Elsevier souhaite développer ces prochaines années et les problématiques juridiques qui pourraient voir le jour (données personnelles, confidentialité notamment).

Sur les suites à donner à ses présentations :

Peut-être une prochaine réunion avec le legal département d'Elsevier.

L'implication de l'unité qui rédige le rapport :

Il est très intéressant que la DAJ soit associée aux rencontres avec les éditeurs dans le but de connaitre leurs futurs développements et préparer les arguments juridiques qui pourront être déployés lors des futurs négociations ou révisions des contrats en cours eu égard aux importants changements induits par la Loi *Pour une République numérique*.

CONCLUSIONS

Gain de potentiel d'analyse associé à la bonne utilisation des bases et des outils

La visite effectuée chez Elsevier a confirmé la dynamique en œuvre chez le fournisseur et leur volonté affichée de proposer des services et des bases de données adaptées aux spécificités de la recherche française en matière de systèmes de la recommandation et de data management. Pour eux, le CNRS apparaît comme un client emblématique, premier producteur de recherche scientifique au niveau mondial par le nombre de ses publications.

Les enjeux pour la recherche française sont d'importance puisqu'il s'agit d'augmenter le rayonnement et l'impact de la recherche produite dans nos laboratoires en la rendant plus visible et plus repérable, et de fournir aux institutions (organismes de recherche, universités de recherche, établissements) et aux instances d'évaluation les outils et services qui permettront de mieux valoriser leur production scientifique (par exemple en la rendant visible au monde extérieur), d'effectuer des comparaisons internationales plus équitables et des analyses d'impact plus pertinentes. L'enjeu est également de favoriser le partage et la valorisation des informations bibliographiques et d'en exploiter les contenus pour mieux fonder notre compréhension des nouveaux processus de recherche et des pratiques en matière de production, utilisation et valorisation des résultats de la recherche. Enfin, fournir aux utilisateurs, quelle que soit leur fonction, des outils de maîtrise et d'analyse de l'impact des publications, ce sont les aider à formuler, avec des outils et services professionnels, les réponses aux enjeux, souvent perçus comme menaçants, de l'évaluation et de la comparaison internationale.

Des questions en suspens

L'offre d'Elsevier présente tout de même quelques réserves quant au gain immédiat et futur. La question de la concurrence n'a pas été abordée comme celle de Google avec ResearchGate.

Les questions essentielles en matière juridique du traitement des données à ce propos nécessitent beaucoup plus de précisions de la part d'Elsevier. A ce titre, il a été proposé d'organiser un déplacement à Amsterdam dans leurs locaux sur une thématique *legal*.

Le CNRS rédigera une étude sur les systèmes de recommandation à l'initiative de la DIST afin de se positionner très clairement sur cet enjeu majeur.

ANNEXE

Présentation de Maya Hristakeva sur les Recommenders systems