



L'entrepôt de données – un modèle coopératif

Ghislaine FILLIATREAU

Directrice de l'Observatoire des Sciences et des Techniques

Quatre préoccupations semblent émerger des débats de la journée : Quelles sont les indicateurs fournis par l'OST au ministère et qu'en fait-il ? Comment utiliser les indicateurs que l'OST transmet directement aux établissements ? Comment les mettre en relation avec ce qu'ils ont par ailleurs ? Comment évoluer tous ensemble et faire progresser les outils dans un référentiel commun ?

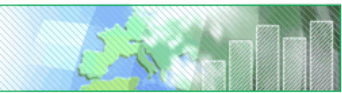
Nous nous inscrivons dans une démarche partenariale car nous souhaitons avant tout être utiles à l'ensemble des établissements sans lesquels nous ne pouvons rien faire.

Tout d'abord, concernant la qualité de l'interaction. Vous, établissements, évoluez très vite en ce qui concerne la collecte de données et nous devons, nous OST, progresser au-delà de ce que nous savons faire avec nos bases. Il s'agit de se pencher sur ce que vous faites afin d'améliorer les indicateurs que nous pouvons vous offrir en retour. La valeur ajoutée de l'OST dans cette boucle réside dans sa capacité à normaliser, à contrôler les données et à les rendre comparables. Nous sommes ensuite en mesure de vous proposer une restitution par le biais d'indicateurs permettant de vous situer les uns par rapport aux autres.

Vous avez également formulé quelques remarques sur la façon dont nous vous faisons travailler . En effet, vous « subissez » nos interfaces. Mais nous essayons d'améliorer les choses. Par exemple, nous pourrions vous donner les numéros de position des adresses afin de faciliter le repérage. Même si il y a des problèmes de licence, je m'engage à ce qu'on les résolve ou les négocie pour que vous ayez au plus vite ce type d'informations. De même, nous souhaitons développer avec vous, des outils de pré-repérage et peut être serait-il utile que nous mutualisions les méthodes développés par chacun de manière à les redistribuer. Nous pourrions, par exemple, constituer une plateforme commune portée par l'OST.

Nous devons donc anticiper vos demandes. Je vous ai entendu parler de HAL en SHS et nous pourrions organiser des groupes de veille afin d'organiser des tests sur les bases, en nous partageant le travail. Notre projet de constitution d'un groupe d'utilisateurs ne vise pas simplement à échanger des informations mais à imaginer des solutions opérationnelles efficaces et à organiser des travaux pratiques. Car l'univers des bases est complexe et il ne suffit pas qu'une base existe pour pouvoir la mettre en production de manière valable. Nous devons sans cesse tester les bases avec votre expertise sur la fiabilité de leur contenu. Mais ce travail demande du temps et des ressources. L'OST est bien placé pour offrir ce service là.

J'aimerais également revenir sur les relations avec le ministère. M. Maillet expliquait ce matin que le ministère est, lui aussi, en train d'évoluer. Son appropriation des indicateurs que nous lui fournissons et en particulier la manière de s'en servir nécessite un temps d'apprentissage. Le



Quatrième journée d'information du projet IPERU

partage se clarifie entre l'AERES et le ministère concernant les indicateurs. Le dialogue avec le ministère implique une réflexion préalable de chaque établissement sur les indicateurs le concernant afin de définir un plan d'action étalé sur plusieurs années pour faire éventuellement évoluer le ministère. Désormais, chaque établissement déploie sa propre stratégie et celle-ci est validée par le ministère et non plus le contraire. Les indicateurs sont là pour soutenir la stratégie, sont là pour structurer le dialogue. Ce que nous livrons aux ministères qui est moindre que ce que nous vous livrons n'a d'impact que parce qu'il y aura une boucle retour avec vous.

Actuellement, nous travaillons avec le ministère pour l'aider à s'approprier ces indicateurs.. C'est un travail que les établissements doivent faire aussi. La façon de faire passe vraisemblablement par des groupes de travail, qui permettent d'expliquer comment on fait pour mettre en perspective tel ou tel indicateur, comment on le met en lien avec le contexte et surtout dans l'avenir comment on va faire pour les mettre en lien avec vos données. C'est une expérience qu'ont déjà certains établissements comme à Orsay où les indicateurs de l'OST sont réellement mis en lien avec le programme scientifique et les orientations stratégiques de la présidence. Il s'agit en effet de mettre en perspective tous les indicateurs de manière à ce que les uns éclairent les autres. Les nôtres sont surtout des indicateurs de cadrage, de trajectoire. Ce sont les évolutions qui sont importantes.

Evoquons maintenant tout ce qui nous manque en termes de données. Deux solutions me paraissent envisageables. Si un établissement dispose déjà de sa propre base de données, on peut envisager à partir de cette base de consolider les données afin de les rendre comparables à d'autres niveaux. Ce système de consolidation implique une logique partenariale. Déjà, nous observons des progrès en matière de systèmes d'information notamment grâce à l'AMUE ce qui facilite votre travail. Notre manière de travailler est un peu différente, elle consiste à essayer de voir ce qui est « collectable » sur des bases coordonnées, puis de consolider les données pour leur donner une valeur ajoutée par l'ajout de nomenclatures et enfin la création d'indicateurs. La nomenclature permet d'éclairer certains points spécifiques que les établissements souhaitent mettre en valeur de façon collective.

L'autre solution est de décider que des données sont importantes et qu'il faut carrément les collecter car les bases n'existent pas. Je pense par exemple aux relations sciences/société ou aux contrats signés avec les entreprises. C'est ce que nous réalisons dans le cadre de « l'entrepôt de données ». L'OST participe à deux travaux de ce type, l'un qui s'appelle EREFIN pour le quel de nombreuses informations sont disponibles sur notre site. C'est un vrai travail de collecte de connaissances nouvelles : une base textuelle et documentaire recensant l'ensemble des contrats signés en leur attribuant une nomenclature approuvée par tous.

Je le répète, notre objectif est avant tout de progresser en partenariat avec vous et de nous organiser pour améliorer nos indicateurs. Cela peut être, par exemple, avec des établissements qui souhaitent être pilote, la mise en place d'un groupe de travail qui commence à prendre nos indicateurs pour les mettre en scène ou les mettre en lien avec les politiques de l'établissement...



Questions de la salle

Q- justement, ne pourrait-on pas bénéficier de votre base pour y ajouter nos propres données et la compléter tout en indiquant la source de ces nouvelles informations ? Ce qui permettrait des comparatifs à plusieurs niveaux.

R -L'idée me paraît bonne mais ne serait-il pas plus intéressant d'enrichir directement le Web of Science® (WoS) avec les bases thématiques existantes que vous nous auriez signalées ? Si je comprends bien, vous aimeriez bénéficier de deux jeux d'indicateurs : le premier basé sur nos données et un autre défini à partir de vos données d'enrichissement de la base ?

Q - Oui, tout en gardant une optique comparative et en évitant les doublons.

R - En définitive, vous vous prononcez pour la création d'une base consolidée par nos propres soins à partir des bases thématiques que vous nous recommanderiez. Vous auriez alors accès à une sorte de Web of Science® (WoS) enrichi dans lequel vous feriez vos repérages de la même manière qu'aujourd'hui. Un tel chantier vient d'être lancé par l'Allemagne et le ministère allemand de la Recherche y a consacré trois millions d'euros. C'est intéressant mais ce sont des travaux lourds.

Q – Il y a des établissements qui construisent leur propre outil, leur système d'information avec leur propre logiciel mais aussi avec HAL, des robots ne pourraient ils pas remonter ces données chez vous ?

R- Sauf que notre base est construite à partir des journaux et non des articles que vous ajouteriez. Le référencement d'articles est un métier de documentaliste et non d'ingénierie de base de données. Plus concrètement, si nous réalisions une base globale, à quoi vous serviraient les données reflétant l'ensemble des publications de votre établissement ?

Q- Cela nous servirait par exemple à connaître l'impact de nos publications au niveau international ou national et ainsi repérer les améliorations possibles en matière de publication dans certains journaux plus visibles que d'autres.

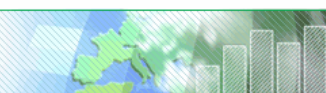
R- Si nous récupérons des publications nous nous situerons alors plutôt dans une logique descriptive permettant à chaque établissement de mesurer son activité mais qui n'est pas particulièrement adaptée à la production d'indicateurs consolidés et comparatifs. Nous n'aurons par exemple pas les données de citations et il faudra « nomenclaturer » à la main. Mais pourquoi pas, cela serait une sorte de base réservoir.

Q - Au niveau national, il existe déjà un projet dénommé GRAAL et que nous ne pouvons ignorer. Il faut donc rester cohérent. Mais les deux approches peuvent être complémentaires.

R - GRAAL consiste à collecter l'ensemble de la production scientifique auprès des établissements. Il serait donc plus simple, que les établissements la complètent ce qui permettrait ensuite de qualifier le Web of Science® (WoS).

Q- GRAAL n'est en aucune façon une base de données mutualisée. L'établissement se contente d'ajouter ses propres publications à GRAAL provenant de HAL en général, mais aucune mutualisation n'est prévue.

R- C'est un entrepôt de données mais non consolidé. J'en reviens donc à notre proposition d'entrepôt de données, cette fois-ci consolidé.



Quatrième journée d'information du projet IPERU

Q- Je reviens sur HAL. Plusieurs organismes ont donné des instructions pour que leurs publications soient dans HAL. Il ne me semble pas utile de remettre en cause cette politique, d'autant que cet outil mobilise une équipe d'ingénieurs importante. HAL permet déjà d'individualiser les données en fonction de la demande de chaque établissement et, à ce titre, constitue une base pertinente. Il faut fiabiliser cet outil et ne pas en créer d'autres sinon, on ne comprendra plus rien dans les établissements. Par contre, comment mettre en lien les entrepôts de données créés dans les établissements et HAL ?

R- HAL c'est exactement comme le Web of Science® (WoS) : c'est juste un accès à des données qui doivent ensuite être retraitées afin de créer des indicateurs et des nomenclatures. A ce titre, HAL sera très utile.

Q - Quels sont les indicateurs qui sont fournis au ministère par l'OST et quel usage en fait-il actuellement ?

R - Il me semble que l'usage qui est fait de ces indicateurs évolue constamment dans la mesure où le ministère est en phase d'apprentissage et d'appropriation. Ce sont des indicateurs de base descriptifs comme la part de production scientifique, la visibilité par discipline... qui permettent de positionner l'institution.

Intervention du ministère

Je le dis très clairement : un indicateur ne constitue pas un outil de classement destiné à déterminer les budgets lors de la négociation contractuelle ou à vous classer les uns par rapport aux autres. On a toujours conçu le chantier IPERU avec l'OST comme une aide qu'on fournissait aux établissements pour le dialogue contractuel. Nous réfléchissons sans arrêt à la pertinence des indicateurs utilisés. On vient de signer à nouveau la convention avec l'OST et on s'est mis d'accord pour tous les ans revoir les indicateurs que nous vous fournissons. L'essentiel est de vous aider à vous positionner, à vous comprendre et que tout cela vous aide dans la réflexion stratégique qu'on vous demande d'avoir avec nous. Par exemple, j'ai constaté au cours des années passées que les indicateurs que nous vous fournissions en matière de part mondiale de publication n'étaient pas utilisés dans le cadre des contrats quadriennaux des universités parce que pour les établissements un pourcentage infinitésimal du type 0,002 % n'avait guère d'utilité. Ainsi, pour qu'un établissement utilise véritablement un indicateur, il faut qu'il s'y retrouve et qu'il y voie une signification. De ce fait, nous nous sommes recentrés sur des indicateurs en part nationale voire européenne. Notre rôle consiste à vous aider à structurer votre positionnement en France, en Europe et à l'international et les indicateurs sont faits pour cela. Vous pourrez nous dire ainsi : ma stratégie c'est ça dans tel positionnement car aujourd'hui je suis là et je veux aller là.

Q – Le débat est très instructif. D'un côté, nous avons les indicateurs de l'OST avec ces positionnements internationaux par exemple et de l'autre, les attentes du ministère, de l'AERES avec sa codification propre, ce nouveau système de répartition des moyens, avec un bonus-malus en fonction du taux de publiants et de non publiants. D'un autre côté encore, on a des incitations de la part du CNRS pour déposer nos publications dans HAL. Il serait important que l'OST conserve sa principale mission de construction d'indicateurs afin d'évaluer la visibilité des établissements mais qu'il puisse également calculer des indicateurs à partir des bases de référence des établissements ou



Quatrième journée d'information du projet IPERU

de HAL pour qu'ils puissent avoir de façon presque automatique toute une batterie de chiffres qui les intéresse. De cette façon, chaque établissement pourrait se comparer aux autres selon un critère homogène défini par le ministère.

R- Je suis d'accord avec votre remarque.

Q- Il me semble que l'OST remplit bien sa mission. Au sein de notre établissement, nous avons incité nos chercheurs à entrer leurs publications dans HAL et nous les avons comparées avec les chiffres de l'OST. Au final, les deux méthodes aboutissent à des résultats très proches. En tout état de cause, il ne faut pas multiplier les nomenclatures. Les chercheurs doivent simplement prendre l'habitude de déposer leur notice bibliographique dans HAL.

R- Je vous remercie. A l'évidence, HAL constitue une préoccupation majeure pour nombre d'institutions. Je constate que vous vous en servez avant tout comme un entrepôt de données et non pas comme une source d'indicateurs normalisés et comparatifs.

Q- En 2006, HAL a fait l'objet d'une signature entre la CPU, le CNRS. Les organismes sont rentrés dans le jeu, les universités beaucoup moins. Sans doute faudrait-il les inciter à s'impliquer davantage dans ce projet. Par ailleurs, si nous adoptons HAL, il faudra que les classifications de HAL et de l'AERES concordent.

R - Nous disposons maintenant du DOI dans le Web of Science® (WoS) et nous pouvons donc faire un lien.

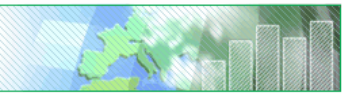
Q- Concernant HAL, il est souvent dit que chacun y fait n'importe quoi mais il existe une fonctionnalité de tamponnage qui permet de valider une publication. Par exemple, le directeur de laboratoire ou un organisme indépendant certificateur pourraient apposer une certification pour chaque article, notamment pour les sciences humaines et sociales.

R - C'est un sujet délicat dans la mesure où il pose la question de l'expertise et de l'indépendance de la personne qui valide. Si le Web of Science® (WoS) est devenu un standard de fait c'est aussi parce que cette base pratique le *cover to cover* : c'est-à-dire qu'elle est systématique ; le journal est l'atome de base : ou sélectionne mes journaux ou collecte un univers de référence. HAL, comme Google Scholar, posent encore des problèmes car la couverture est mouvante.

Q - HAL correspond à une logique de diffusion et non d'évaluation. L'idée d'un autocontrôle de la communauté scientifique semble prévaloir sur HAL. Un chercheur n'a donc pas intérêt à mettre en ligne un article de qualité médiocre sur HAL.

R - Certaines disciplines sont très régulées et d'autres beaucoup moins.

Q- HAL est une archive ouverte et son périmètre est beaucoup plus incertain que celui du *Web of Science*. Il revient à chaque établissement de mettre en place des garde-fous pour contrôler les publications. Concernant la codification, je sais que l'INSA de Lyon oblige ses chercheurs à déposer leurs publications dans HAL. Ainsi, lors des remontées des bilans scientifiques des unités, l'INSA demande à ce que les publications soient exclusivement extraites à partir de HAL. C'est un modèle jusqu'au-boutiste. Néanmoins, par rapport au Web of Science® (WoS), l'avantage de HAL



Quatrième journée d'information du projet IPERU

réside dans le dépôt de nombreux articles de sciences humaines et sociales. HAL apparaît sans doute comme la base la plus crédible pour les SHS.

R- Il ne faut pas perdre de vue les divers usages qui sont faits des informations. La réflexion stratégique n'impose pas d'avoir fait la mesure complète de l'activité de l'établissement. Ce qui est important ce n'est pas tout ce que j'ai fait mais ce qui me permet de me situer par rapport aux autres. Notre rôle consiste à irriguer cette réflexion stratégique grâce à nos indicateurs. Nous ne faisons pas d'évaluation scientifique. Nous nous attachons à la création d'indicateurs robustes pour l'évaluation stratégique. C'est un travail avant tout prospectif. Tout est normalisé et vous pouvez poursuivre les courbes de ce que nous vous fournissons. Ce sont des tendances lourdes même si elles peuvent s'infléchir.

Q - L'utilisation de HAL comme archive ouverte est différente de celle du dépôt de notices bibliographiques. Dans le premier cas, il s'agit avant tout de permettre aux chercheurs de communiquer entre eux tandis que, dans le second cas, nous parlons plutôt d'une activité administrative. La qualité des publications déposées et l'exactitude des notices bibliographiques sont de la responsabilité des tutelles ou des directeurs de laboratoires. Il n'y a aucune raison qui justifie le maintien de notices incorrectes dans HAL.

R- Pour ma part, je parlais plutôt du caractère incomplet des notices et non de leur éventuelle inexactitude. C'est pour cette raison que nous devons comparer HAL à d'autres bases afin de compléter les informations manquantes. Si vous voulez vous situer, vous ne pouvez pas partir seulement de vos publications. Elles doivent être réinsérées dans le journal et dans l'ensemble des publications de ce journal. Lui-même est repris dans l'ensemble du domaine thématique, et de la discipline.

Q - Concernant les indicateurs de l'OST, peut-on effectuer des zooms pour obtenir des informations précises au niveau des unités et non simplement des établissements ?

R- Nous pouvons même aller jusqu'aux articles

Q - Cela suppose-t-il un enrichissement de la base avec des métadonnées ? En effet, vous devez connaître la structure et l'organisation de chaque université pour mieux analyser les données et produire des indicateurs d'activité au niveau de chaque département ou unité de recherche.

R- Effectivement, la base doit être enrichie au moment du repérage afin que nous puissions identifier les données et leur provenance de la façon la plus fine. Ce travail doit être effectué avec l'aide des groupes de travail car la tâche est considérable. A ce titre, il me semble préférable d'enrichir des bases déjà existantes comme le Web of Science® (WoS), sauf dans les cas où les disciplines ou les bases sont défectueuses ou partielles. Mais, d'ores et déjà, notre base est très puissante, elle peut être enrichie mais il faut qu'on soit bien d'accord sur le niveau que l'on veut y introduire en plus via les interfaces.

Q - Que pensez-vous des études effectuées à partir de Google Scholar, de Scopus et d'autres systèmes similaires ? Quels sont les problèmes par rapport à vos méthodes ?

R - Pour prendre une analogie dans Google Scholar on lance son filet un peu au hasard tandis que nous recensons toutes les espèces de poisson existantes. C'est une question de critères de sélection. La logique qui préside à la sélection des journaux dans le Web of Science® (WoS) revient à



Quatrième journée d'information du projet IPERU

considérer la recherche comme un tissu de citations qui se font écho les unes aux autres. De ce fait, la recherche se bâtit à partir d'un ensemble de liens souterrains représentés par les citations. Dans chaque grand domaine disciplinaire, un certain nombre de personnes travaillent sur des problématiques similaires et accumulent les connaissances. Pour chaque discipline, il existe un noyau représentatif de la discipline incarné par un substrat d'articles les plus couramment cités. Même si ce cumul de citations ne reflète par mécaniquement la qualité intrinsèque des articles, cette méthode permet de bien mesurer l'impact de ce noyau d'articles sur la recherche actuelle. Il s'agit moins d'une logique d'évaluation de la recherche que de représentations des tendances scientifiques au sein de chaque discipline. Pour notre part, nous nous attachons avant tout à situer les établissements, discipline par discipline, les uns par rapport aux autres.

Q- Beaucoup de chercheurs regrettent les dérives induites par le *science citation index* du Web of Science® (WoS). L'idée originelle était bonne mais, par exemple, certains Américains qui ont travaillé dans des laboratoires de recherche français ne citent toujours pas les publications françaises.

R - Pour notre part, nous ne proposons jamais d'indicateurs individuels, par exemple sur les publiants. Nous adoptons une sorte de « vue aérienne » au niveau de chaque établissement.

Q - Certaines disciplines comme l'informatique organisent beaucoup de colloques avec actes et il serait sans doute intéressant, dans une optique de visibilité internationale, de considérer les communications avec actes de la même façon que les articles publiés dans des journaux à comité de lecture. Je vous rappelle que les communications avec actes sont comptabilisées comme des publications de rang A.

R - Nous venons d'ouvrir ce chantier. Nous achetons désormais les *Proceedings* du Web of Science® (WoS). Les limites actuelles sont liées à des problèmes d'ingénierie de base de données et de moyens plus que d'un manque de volonté.

Je retiens que parmi les différents groupes de travail que nous pourrions constituer, je pense notamment à un groupe dédié à la veille critique sur les autres bases, un autre consacré aux SHS, un groupe centré sur la mutualisation des méthodes de repérage et un autre portant sur l'amélioration de notre interface. Enfin, nous pourrions réfléchir à un moyen d'enrichir HAL.

Je vous remercie de votre participation à cette journée d'échanges très éclairants pour nous : nous allons nous efforcer d'en tirer le plus d'améliorations possibles.