Balisages
La revue de recherche de l'Enssib

3 | 2021
Penser les données par le territoire ?

Appréhender le périmètre de recherche à l'université Paris-Saclay

La mise en place de BiblioLabs

Luc Bellier, Henri Bretel et Laurence Gandois

Édition électronique
URL : https://journals.openedition.org/balisages/663
DOI : 10.35562/balisages.663
ISSN : 2724-7430

Éditeur
École nationale supérieure des sciences de l'information et des bibliothèques (ENSSIB)

Référence électronique
Luc Bellier, Henri Bretel et Laurence Gandois, « Appréhender le périmètre de recherche à l'université Paris-Saclay », *Balisages* [En ligne], 3 | 2021, mis en ligne le 16 novembre 2021, consulté le 05 juin 2024. URL : http://journals.openedition.org/balisages/663 ; DOI : https://doi.org/10.35562/balisages.663

APPRÉHENDER LE PÉRIMÈTRE DE RECHERCHE À L'UNIVERSITÉ PARIS-SACLAY

La mise en place de BiblioLabs

Luc Bellier

Directeur adjoint, Direction des bibliothèques de l'université Paris-Saclay
luc.bellier@universite-paris-saclay.fr

Henri Bretel

Chargé de bibliométrie, Direction des bibliothèques de l'université Paris-Saclay
henri.bretel@universite-paris-saclay.fr

Laurence Gandois

Chargée de traitement de données,
Direction des bibliothèques de l'université Paris-Saclay
laurence.gandois1@universite-paris-saclay.fr

Nous proposons de traiter la façon dont les questions de périmètre et de territoire institutionnel de la recherche à l'université Paris-Saclay (UP-Saclay) ont été appréhendées dans le but de mettre en place des outils de pilotage de la recherche. Au gré de la structuration de l'établissement, des reconfigurations institutionnelles, différents acteurs ont collaboré à la mise en données du paysage et de ses acteurs. Ils ont dû lever des obstacles bien souvent liés aux frontières institutionnelles et techniques héritées des établissements composant l'université et trouver des solutions afin de répondre à des besoins d'appréhension des périmètres de recherche de plus en plus précis et à des services de plus en plus diversifiés. Partant de questions bibliométriques, BiblioLabs devient aussi un outil au service de la science ouverte, s'appuyant sur des outils tels que les identifiants pérennes.

Mots-clés : territoire, recherche, université Paris-Saclay, BiblioLabs, identifiants

We propose to deal with the way in which issues of the scope and institutional territory of research at the Université Paris-Saclay have been understood in order to set up research management tools. As the establishment was structured and institutional reconfigurations, different actors collaborated to put the landscape and its actors into data. They had to remove obstacles very often linked to the institutional and technical boundaries inherited from the institutions that make up the university and find solutions in order to meet the needs of apprehending increasingly precise research perimeters and increasing services more diverse. Starting from bibliometric questions, BiblioLabs is also becoming a tool at the service of open science, relying on tools such as perennial identifiers.

Keywords: landscape, research, university Paris-Saclay, BiblioLabs, identifier

INTRODUCTION

BiblioLabs est un outil développé au sein de l'UPSaclay afin d'en appréhender les publications, alors qu'elle était organisée en Communauté d'universités et établissements (COMUE) sous l'impulsion des équipes en charge des questions documentaires et en lien avec celles en charge de la recherche. D'abord simple liste de structures, puis référentiel de signatures, BiblioLabs est devenu progressivement un outil de traitement des données liées aux publications et aux projets de recherche sous financements européens, qui traduit le paysage institutionnel de la recherche non seulement au sein de l'UPSaclay mais aussi dans les relations que cette dernière entretient avec d'autres acteurs de la recherche en France et à l'étranger. Aussi, la question du territoire est-elle centrale dans la construction de cet outil dont la fonction est de déterminer ce qui est dans le périmètre de l'université, ou de ses composantes, et ce qui, tout en étant en dehors de ce périmètre, participe des activités de recherche et de publications par des collaborations diverses. La notion de territoire, liée à celles de périmètre et de frontière, étant constitutive de la démarche à l'œuvre dans BiblioLabs, il nous a paru intéressant de partager la façon dont ce territoire se conçoit, se dessine et se redessine à mesure que l'institution s'organise et tout cela par les données, et la structure de ces données. Il s'agit donc de livrer un retour d'expérience sur la façon dont les questions de territoire et de périmètre ont pu être appréhendées et travaillées dans une base de données dans un contexte de profonde modification institutionnelle.

Pour les besoins de BiblioLabs[1], l'UPSaclay retient une définition juridique de son périmètre de recherche, sur un territoire donné, à cheval entre plusieurs institutions, et qui permet de définir ce qui est dans l'université et ce qui est hors de l'université. Si cette dernière notion binaire se traduit aisément dans une base de données, les périmètres au sein de l'institution peuvent être plus flous, plus mouvants et donc plus délicats à traduire sous forme de données.

La donnée, si l'on s'en tient à sa définition habituelle d'enregistrement ou mesure du réel, autrement dit d'enregistrement factuel[2], est bien sûr omniprésente sur un campus scientifique. Elle est également hétérogène. Ainsi par

1. BiblioLabs a fait l'objet d'une précédente publication à un stade plus précoce de son développement : Vincent Thébault, « BiblioLabs, un outil au service du pilotage de l'université Paris-Saclay », *Arabesques*, 2020, 96. [En ligne] < https://publications-prairial.fr/arabesques/index.php?id=1478 > (consulté le 19 septembre 2021).

2. « Les « données de la recherche » sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche », *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*, 2007, p. 17. < https://www.oecd.org/fr/science/inno/38500823.pdf >.

son travail d'analyse, tout chercheur produit et utilise des quantités importantes de données (à la fois publications originales, enregistrements ou productions intermédiaires plus couramment appelés données de la recherche). L'administration d'une université, par son travail d'enregistrement des activités de l'établissement, produit également des données massives.

Même si l'on se fonde sur une définition légèrement plus restrictive, qui associe toujours la donnée à un traitement automatisé[3], la méthodologie scientifique, le goût et le besoin de tous les services de soutien à la recherche pour la normalisation des remontées d'informations font des universités en général, et de l'UPSaclay en particulier, un territoire de données.

CONTEXTE TERRITORIAL ET INSTITUTIONNEL

Du plateau au cluster de recherche : vers la création de la COMUE

Le territoire géographique du plateau de Saclay a une longue tradition de recherche scientifique, si bien qu'on peut voir l'origine du cluster dès la sortie de la Seconde Guerre mondiale. Depuis l'installation du Commissariat à l'énergie atomique et aux énergies alternatives (CEA) en 1954 sur le plateau, Saclay a vu s'installer un nombre important de structures de recherche et d'enseignement dans les années 1960 et 1970 (HEC, Polytechnique et Supélec) qui se poursuit encore avec l'arrivée de l'École normale supérieure (ENS) en 2020[4].

Le cluster de recherche, enseignement supérieur et innovation, formé *de facto* autour de la vallée d'Orsay et du plateau de Saclay est structuré de diverses manières, l'important étant d'afficher sa volonté de regroupement ; apparaît d'abord un Pôle de recherche et d'enseignement supérieur (PRES), UniverSud Paris, doublé de deux réseaux thématiques de recherche avancée (RTRA) : Digitéo et Triangle de la physique. Vient ensuite, dans une logique

3. Amaël Cattaruzza, *Géopolitique des données numériques : pouvoir et conflits à l'heure du big data*, Paris, Le Cavalier Bleu, 2019.

4. Le début des années 1950 voit les premiers bâtiments scientifiques s'installer à Orsay, dont le futur Institut de physique nucléaire, sous l'impulsion d'Irène Joliot-Curie. Le terrain est choisi parce qu'il est à la fois grand et disponible, car sous séquestre depuis la Libération. La ligne de trains dite de Sceaux permet une certaine accessibilité des lieux depuis Paris. Dès lors, le besoin de place pour l'université de Paris se faisant sentir, le choix d'un campus à la fois accessible et proche de la réalité scientifique impose Orsay. Puis la même logique préside à la mise en place de campus au sud de Paris pour Supélec, Polytechnique, l'Institut d'optique Graduate School (IOGS), etc., au point que ce qui est déjà un cluster scientifique en 2006 devient un nouveau projet de regroupement universitaire, qui fait suite à la loi de programmation de la recherche de 2006.

de fusion de plus en plus nette, une Fondation de coopération scientifique (FCS), qui laisse la place dès 2014 à une COMUE, préfigurant l'UPSaclay.

Évolutivité du périmètre, permanence du paysage

Une fois le paysage de la recherche sur le plateau stabilisé, les instabilités institutionnelles, tributaires de nombreuses volontés politiques distinctes, ne font que commencer. Ce sont d'abord les évolutions du cadre dans lequel s'est successivement inscrit Saclay depuis la création du PRES UniverSud Paris en 2007, jusqu'à la création de l'établissement expérimental public à caractère scientifique, culturel et professionnel fin 2019. Ce sont ensuite les évolutions des établissements eux-mêmes avec la fusion de l'école Centrale Paris et de l'école Supélec, en 2015, puis le retrait de différentes écoles autour de Polytechnique du périmètre de l'UPSaclay afin de créer en 2017 l'Institut Polytechnique de Paris (IP Paris).

Figure 1. La composition de l'université Paris-Saclay en 2021

Les 10 composantes sont héritées de l'université Paris-Sud et sont rejointes par les 4 établissements-composantes. Cela implique qu'ils sont pleinement dans l'UPSaclay et, à ce titre, participent de sa gouvernance mais qu'ils conservent leur personnalité morale et juridique, notamment afin de poursuivre la pleine gestion des campus qui sont en dehors de l'université administrativement et géographiquement (Rennes, Bordeaux, etc.). *Source : Service communication de l'université Paris-Saclay.*

À tutelles différentes, correspondent des financements distincts. Ceci implique inévitablement un sentiment d'appartenance distinct, si bien que les points communs peuvent s'avérer insuffisants pour « faire territoire », donc pour habiter de manière commune un espace perçu comme collectif. Les 19 établissements qui convergeaient dans la FCS, puis la COMUE, avaient inévitablement des divergences de vision et de culture.

C'est ainsi que l'IP Paris s'est constitué autour de l'école Polytechnique, et en quittant l'UPSaclay, redessine le paysage de la recherche à Saclay. BiblioLabs a traduit ce changement de paysage en plaçant l'IP Paris hors du territoire de la recherche de l'université. Cependant, la scission avec l'IP Paris ne met pas fin aux initiatives partagées[5], que ce soit pour la recherche ou l'enseignement, mais scinde bien les données administratives et scientifiques.

Le passage de 19 à 14 établissements, parmi lesquels restent les trois universités et quatre grandes écoles, en plus des organismes nationaux de recherche et de l'Institut des hautes études scientifiques, ne résout pas toutes les questions de fusion, mais le sentiment de rapprochement suscité par le rétrécissement du périmètre permet des avancées. L'université Paris-Sud choisit de disparaître pour servir de base à l'UPSaclay : même s'il ne s'agit pas techniquement de la fusion réelle entre deux établissements d'enseignement supérieur et de recherche (ESR). Elle permet aux grandes écoles de devenir les établissements-composantes de la nouvelle université, inclus dans le périmètre de recherche mais sans perdre leur personnalité morale et juridique, et conservant par conséquent leur autonomie budgétaire.

Les choix de périmètre faits par les établissements-composantes convergent vers une certaine cohérence « territoriale », à savoir que les campus non localisés géographiquement en Île-de-France sont également hors du périmètre de recherche de l'université. La logique du regroupement territorial, matérialisé par les déménagements successifs (IOGS, puis CentraleSupélec, puis ENS Paris-Saclay, enfin peut-être AgroParisTech) précède donc en grande partie celle du regroupement administratif. En revanche, l'université Paris-Sud s'étant fondue dans le nouvel établissement public expérimental, ses quelques équipes qui ne font pas partie des campus franciliens, sont bel et bien incluses dans les données. Les exceptions à la logique territoriale du regroupement, au-delà des difficultés statistiques qu'elles peuvent poser, sont intéressantes dans la construction d'un territoire numérique. En effet, plus ou moins connues et respectées par la communauté scientifique, les règles ont un impact sur la capacité des chercheurs à s'identifier à une communauté

5. On peut citer notamment le portail documentaire Focus qui reste commun aux différents établissements : < https://focus.universite-paris-saclay.fr >.

et à montrer leur appartenance dans les données. Or, les exceptions administratives sont souvent marginalisées par rapport au fonctionnement central, ce qui transparaît dans les données. C'est particulièrement visible à travers la signature scientifique.

Par la permanence de l'essentiel des structures de recherche, les habitudes de collaboration scientifique entre équipes de recherche et aussi par l'affirmation progressive mais certaine de l'UPSaclay dans le pilotage de la recherche et le travail de normalisation de la signature, commencés en 2016, se renforcent progressivement et permettent dès lors d'envisager par les données le paysage de la recherche.

Le besoin de pilotage

Le fait qu'une administration compte et mesure la production scientifique de l'université, constitutive de la bibliométrie[6], ne suffit pas à transformer cette donnée en un territoire. Il faut que la donnée soit considérée par ses producteurs et acteurs comme appartenant à l'université. Ceci passe en particulier, du point de vue de la recherche, par l'appropriation d'une signature commune. Le respect de la signature étant l'un des indicateurs les plus évidents de l'appropriation par une communauté de l'identité de l'UPSaclay, on remarque sans surprise une liste de facteurs d'éloignement corrélés à une baisse du taux d'acceptation de la signature commune, qu'on pourrait d'une certaine manière appeler des facteurs de marginalité dans la donnée :

- éloignement géographique du campus, pour une équipe de l'université localisée hors d'île de France ;
- éloignement disciplinaire du « cœur » de l'université. On peut citer les publications en SHS, qui ne se reconnaissent pas toujours dans l'historique d'une université construite autour de la chimie et de la physique des particules ;
- éloignement institutionnel de la tutelle communautaire, comme pour les publications des universités membres-associés en attente de fusion ;
- éloignement professionnel du centre de gravité constitué par la recherche scientifique. On peut citer en particulier les juristes ou de médecins.

Ces marginalités, subies ou choisies, sont toujours pour la communauté une source de lacunes dans le repérage bibliométrique, et par conséquent, des données absentes ou imprécises.

6. On entend par bibliométrie les techniques et méthodes consistant à compter l'ensemble des publications scientifiques de l'institution ainsi que les méthodes de mesure de l'impact scientifique de ces publications.

Ce sont donc, dans la définition de la communauté de l'UPSaclay, trois territoires qui évoluent ensemble et s'alignent peu à peu. Le premier est administratif, ce sont les définitions institutionnelles de l'université, que la communauté s'approprie plus ou moins à mesure qu'elle connaît les particularités de l'établissement et s'y sent intégrée ou non. Le deuxième est géographique, c'est l'espace large du sud de Paris, entre Versailles et Evry, autour du centre non géographique mais symbolique constitué par le plateau de Saclay. C'est vers ce centre que convergent les écoles d'ingénieur (Centrale, Supélec, ENS Paris-Saclay, IOGS, et prochainement AgroParisTech) et les centres de recherche d'organismes nationaux (CEA sur le plateau de Saclay, CNRS à Gif-sur-Yvette, Office national d'études et de recherches aérospatiales [ONERA] et Institut national de recherche en sciences et technologies du numérique [INRIA] à Palaiseau, etc.), si bien que le territoire géographique tend peu à peu vers un alignement avec le territoire institutionnel. Enfin, le troisième est le territoire scientifique, défini par la signature, et qui ne prend sens qu'à travers les données de publications rassemblées dans BiblioLabs. Les remontées d'information à travers diverses sources permettent de faire correspondre les trois territoires, mais aussi d'en connaître les particularités et les exceptions.

SIGNATURE ET TERRITOIRE : DES IDENTIFIANTS ET DES RÉFÉRENTIELS

Du référentiel signature aux données bibliométriques

La donnée scientifique et ses métadonnées sont, à l'image de tout espace matériel, une réalité complexe qu'il convient d'approcher sous de nombreux prismes : si l'analyse de l'apparition de nombreux centres de données sur le territoire peut s'apparenter à de la géographie physique de la donnée, le regard porté sur les différents systèmes d'information utilisés par les services de l'université, et les liens construits ou en construction entre eux, ont des analogies avec l'aménagement d'un territoire et son urbanisation.

Les frontières des données de l'université

La délimitation de la donnée et son périmètre pourraient sembler plus simples au premier abord, car une frontière institutionnelle se veut par définition binaire, les deux seuls états possibles étant « dedans » ou « dehors ». Or elle est rendue plus complexe par les multiples barrières qui peuvent exister entre les systèmes d'information « en silo », barrières techniques (formats et protocoles d'échanges), politiques (volonté d'échanger, même au sein d'une même

institution), institutionnelles. Toutes sont héritées des frontières antérieures entre les établissements préexistants.

Cependant, une frontière extérieure délimitant le « nous » des « autres » est nécessaire et doit être connue de tous les acteurs. Cette frontière est le périmètre de l'université, au-delà de la simple appartenance d'établissements à une COMUE. La difficulté vient du fait que l'UPSaclay n'est pas la même en tant qu'employeur qu'en tant qu'établissement d'enseignement supérieur, ni même qu'en tant que centre de recherche. La frontière est donc multiple. BiblioLabs a eu pour objectif de clarifier d'abord le périmètre encore vague, puis les différents périmètres identifiés de l'université, même si son premier enjeu est le périmètre de la recherche.

Pour cela, il a fallu définir et différencier les périmètres : institutionnellement, l'UPSaclay en a deux, le premier appelé communément « périmètre employeur » et qui comprend l'ensemble des unités et services dont les agents sont rémunérés par l'université, et le second, appelé « périmètre recherche », qui comprend l'ensemble des unités dont au moins une tutelle est membre de l'université et sur le territoire du sud parisien.

BiblioLabs s'est attaché à caractériser ce périmètre recherche. Selon le niveau d'intégration des structures de recherche, les publications émanent de l'un ou l'autre périmètre. Aussi, il a fallu construire plusieurs concepts afin de préciser les différents niveaux d'appartenance à l'université. La construction de ces concepts et de leur signification s'est effectuée à mesure que le paysage et la gouvernance de l'université se sont mis en place. Nous en citerons deux, déjà évoqués précédemment (voir figure 1) :

- établissements-composantes (tels qu'AgroParisTech, CentraleSupelec et l'ENS Paris-Saclay) qui sont dans l'UPSaclay fusionnée mais tout en conservant leur personnalité morale et juridique ;
- universités membres-associés (tels que l'université d'Évry Val d'Essonne [UEVE] ou l'université de Versailles Saint-Quentin-en-Yvelines [UVSQ]) qui participent du périmètre recherche mais ne sont pas encore fusionnés et conservent leur personnalité morale et juridique.

Une fois le périmètre acquis, la signature normalisée peut être traduite en données, puisque chaque signature renvoie à une structure, qui, à son tour, renvoie à un périmètre.

BiblioLabs, développé à la COMUE sous l'impulsion des professionnels de l'information en lien avec la Direction de la recherche, est aujourd'hui géré par la Direction des bibliothèques de l'université. Au sein de l'équipe, deux personnes sont plus particulièrement en charge de cet outil. Une personne, assure, parmi d'autres missions, la maintenance des données. BiblioLabs propose des outils de suivi et de pilotage à différentes directions (relations inter-

nationales, recherche principalement). Une autre personne est chargée de faire évoluer les fonctionnalités de l'outil pour l'adapter aux besoins des utilisateurs. Si la gestion des données et des référentiels est effectuée en grande partie par la Direction des bibliothèques, certaines données sont maintenues et qualifiées par les utilisateurs hors des bibliothèques.

Dans son principe premier, BiblioLabs est un référentiel de signatures validé par la communauté. Ces signatures normalisées sont plus facilement repérées dans les publications et BiblioLabs peut les associer aux structures de recherche et à leurs identifiants. Dès lors que le travail d'enregistrement concerne plusieurs milliers de publications chaque année et plusieurs centaines de structures de tailles et statuts différents, l'automatisation du processus est nécessaire. L'alignement de chaque entité à un ou plusieurs identifiants numériques capables de qualifier les données nominales est ainsi réalisé par des traitements automatisés.

Reste manuelle la gestion des structures et de leurs identifiants au gré des scissions, des fusions et des réorganisations, afin que le registre des structures traduise au plus près la réalité administrative dans laquelle il s'agit de faire entrer les activités de recherche.

Collecter, aligner, valider : les identifiants

La structuration du territoire, pour en appréhender les contours académiques, est passée par la mise en données des entités de recherches : les identifiants des laboratoires, instituts, Graduate Schools et établissements font l'objet de collecte et de maintenance régulière. Chaque identifiant ayant des relations avec d'autres, il a été nécessaire de concevoir une architecture capable de transformer la donnée brute, qu'elle soit publication, signature ou bien simplement activité de recherche, en une donnée travaillée de manière à servir la communauté.

Cette transformation, dans le cadre de BiblioLabs, passe par le travail de reconnaissance textuelle qui permet de lier une chaîne de caractères, représentant un nom ou une institution, à l'identifiant d'un chercheur, d'un laboratoire, ou bien d'une université. Puis la donnée ainsi obtenue peut être synthétisée dans divers regroupements, permettant la cartographie.

Choix des identifiants

L'alignement entre divers identifiants est l'un des éléments de fiabilisation de la donnée les plus efficaces et, dans un monde de données ouvertes, multiplier les sources d'information traçables est un gage d'efficacité des alignements. Par ailleurs, chaque identifiant répondant à un besoin spécifique, il aurait été

illusoire de croire répondre à toute la diversité des usages avec un seul identifiant. Mais les types d'objets à analyser ont, eux, fait l'objet d'autres choix.

Identifiants de publications

Dans le comptage statistique des publications scientifiques, il est impossible de passer outre l'identification des publications scientifiques. Pour celles-ci, l'utilisation du Digital Object Identifier (DOI)[7] s'étant déjà généralisée avant que BiblioLabs ait été conçu, cet identifiant est apparu comme une évidence. La première raison de ce choix pour le domaine scientifique est qu'aucune alternative viable ne permet d'identifier séparément les différents articles publiés sur un même support (ISBN[8] et ISSN[9] étant uniquement destinés à identifier les supports, ils ne sont utilisables que lorsqu'ils se rapportent à un contenu scientifique unifié, donc dans une minorité de cas).

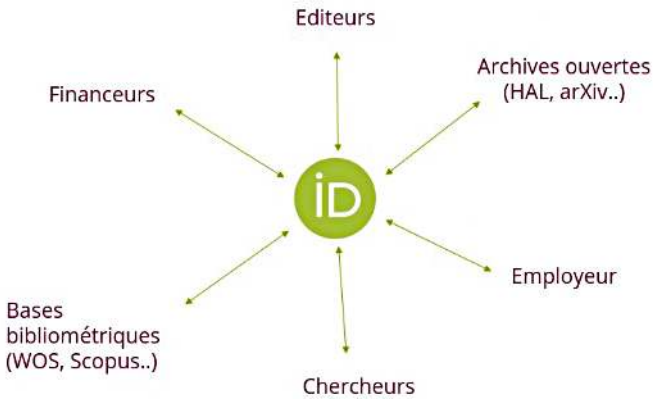
Identifiants auteurs

Moins évidente et plus problématique, l'identification des chercheurs a longtemps été absente de BiblioLabs et l'est encore largement. Puisque la COMUE avait d'abord pour objectif le regroupement et le pilotage consortial et non la gestion individuelle, BiblioLabs se contentera donc, pendant toute la période de création, d'aligner les profils créés algorithmiquement par les plateformes de publications avec les DOI et les identifiants de structures. L'évolution du contexte institutionnel et politique nous a amenés à revoir ce parti pris afin, notamment, de fournir des services aux chercheurs autour de HAL[10]. Pour fournir ces services, l'identifiant Open Researcher and Contributor ID (ORCID) s'est imposé.

Peu à peu, au tournant des années 2010 et 2020, l'identifiant international ORCID[11] s'est affirmé dans le paysage de l'identification des personnes liées à la recherche, grâce à son ouverture, sa couverture de plus en plus étendue et son internationalisation. Il pourrait être considéré comme le « DOI des personnes », mais son principe de fonctionnement, centré sur le chercheur ou le contributeur, sans autorité de contrôle des informations, a ses limites.

7. < https://www.doi.org/ >.
8. < https://www.isbn-international.org/ >.
9. < https://www.issn.org/ >.
10. Comme nous le verrons plus loin, l'UPSaclay ouvre en 2021 un service de versement facilité des publications dans HAL.
11. < https://orcid.org/ >.

Figure 2. Schéma relationnel d'ORCID



ORCID permet à différents acteurs publics et privés dans la recherche de se relier de façon non ambiguë à un chercheur. Ainsi, un chercheur disposant d'un ORCID et le fournissant à son éditeur pour publier un article, ou lors du versement sur une archive ouverte, mais aussi à des organismes financeurs, permet de relier par cet identifiant les travaux publiés, les programmes de recherche, etc.

Il a paru nécessaire d'utiliser une approche différenciée entre le paysage institutionnel, pour lequel notre vision se doit d'être précise, et l'horizon international des collaborations, pour lequel la couverture et la fiabilité peuvent souffrir quelques défauts. Dans le paysage français, nous alignons donc désormais les identifiants ORCID avec les identifiants IDRef[12], tandis que nous nous contentons des premiers pour l'identification de nos partenaires. L'identification des auteurs passe par un système à plusieurs identifiants, qui gagne en complexité quand il s'agit d'identifier les structures.

Identifiants de structures

Le choix de travailler à l'échelon des laboratoires est au départ un parti pris pour une approche territoriale et administrative. Les avantages de la distinction par unité de recherche étaient également évidents dans la finesse de l'analyse statistique, puisque bien plus que les établissements, les laboratoires présentent dans leur organisation une certaine unité thématique, qui permet un découpage disciplinaire au sein de chaque institution, et de l'ensemble, important pour le pilotage.

12. IDref est un identifiant opéré par l'Abes, utilisé pour identifier de façon non ambiguë les acteurs de l'ESR français, chercheurs, établissements, laboratoires, etc. Ils sont attribués par les établissements documentaires de l'ESR, notamment à l'occasion de l'enregistrement d'une thèse de doctorat pour le doctorant mais aussi pour les membres du jury.

L'apport d'unification de la COMUE était donc celui d'une nouvelle tutelle, commune à toutes les signatures du périmètre de recherche, mais qui n'exclurait pas les autres. La difficulté d'un tel travail d'identification est qu'il est particulièrement difficile à saisir pour les organisations étrangères au système français, puisque le même établissement, en l'occurrence l'UPSaclay, était représenté par plusieurs centaines de signatures différentes et valables en même temps. Cette diversité n'était pas prise en compte dans les bases internationales comme le Web of Science, même si elle était visible dans les diverses signatures prises pour base de l'unification dans ces outils. Les identifiants internationaux connus pour les institutions (Ringgold[13], Global Research Identifier Database [GRID][14], puis Research Organization Registry [ROR][15]) n'avaient ni l'universalité d'usage du DOI ni la précision nécessaire au système français, et ne pouvaient donc pas servir de fondement à l'identification des structures dans BiblioLabs. Le répertoire national des structures de recherche (RNSR)[16] permettait théoriquement, en France, de couvrir la totalité des structures existantes, mais ne permettait pas d'analyser les collaborations internationales de l'université. Le système d'identifiants mis en place par l'Agence bibliographique de l'enseignement supérieur (Abes), IDRef[17], présentait le même type d'avantages et d'inconvénients, même s'il proposait un éventail plus large. Chacun de ces identifiants était donc partiellement efficace, et par conséquent utile à obtenir, mais insuffisant pour servir de base à des alignements. L'identifiant choisi au départ comme principal peut donc étonner, mais répond spécifiquement au besoin d'un identifiant capable de traiter les institutions et les laboratoires affiliés, en France comme à l'étranger. Il s'agissait de l'identifiant de structure Scopus, un choix difficile en termes d'ouverture et d'indépendance des données, mais nécessaire et moteur de l'évolution qui pousse désormais BiblioLabs vers un modèle de moins en moins centralisé.

En effet, c'est pour répondre à la question que posent des identifiants de structures à la fois multiples, peu ouverts ou peu internationalisés, que BiblioLabs évolue vers un modèle capable de travailler sans identifiant externe, par un modèle de traçage des sources, de plus en plus compatible avec une approche non tabulaire des données, et suivant les principes du Linked Open Data : le plus important devient la fiabilité des alignements, et non l'unicité des identifiants, et ce principe permet de récupérer toutes les

13. < https://www.ringgold.com/ >.
14. < https://www.grid.ac/ >.
15. < https://ror.org/ >.
16. < https://appliweb.dgri.education.fr/rnsr/ >.
17. < https://www.idref.fr/ >.

informations disponibles dans les sources auxquelles nous avons accès, en gardant la trace de leur origine, pour gagner en rappel grâce au croisement des sources et en précision grâce à la traçabilité.

L'accès aux institutions

L'approche initiale de BiblioLabs par la signature et les structures permet une grande efficacité mais laisse passer des exceptions trop nombreuses. En effet, l'utilisation de la signature normalisée, nous l'avons dit plus haut, est corrélée au niveau d'adhésion à l'institution et son territoire. Certaines cultures professionnelles privilégient des signatures indiquant les fonctions hospitalières ou juridiques et non pas universitaires, limitant la capacité d'identification des publications de la communauté de recherche. Conscients de cette limite, et dans le souci de mieux appréhender l'activité de recherche des structures, une nouvelle étape s'imposait : mieux appréhender les auteurs à travers divers identifiants complémentaires ou alignés, ORCID, IDRef, IdHal, Scopus ID, etc.

Cette nécessaire diversification des identifiants a conduit à un travail important d'alignement, dont la qualité est appelée à devenir centrale pour le bon fonctionnement de BiblioLabs. Nous avons donc mené les chantiers d'alignements en dehors de BiblioLabs afin d'évaluer la qualité des processus de traitement, et de fiabiliser les résultats avant de les intégrer dans BiblioLabs.

Ce cheminement est aussi conduit par l'histoire de l'institution. BiblioLabs, piloté à la COMUE, ne pouvait pas s'appuyer sur les annuaires des établissements, faute d'accès à ceux-ci. Le travail exposé ici a donc commencé d'abord à l'université Paris-Sud dans l'idée à terme d'alimenter et consolider les données présentes dans BiblioLabs. La qualité du résultat et son intérêt en termes de service aux chercheurs ont permis de l'étendre à d'autres établissements volontaires pour la démarche.

Nous avons construit un processus d'alignements d'identifiants chercheurs, à partir d'une extraction d'un nom, d'un prénom et d'une date de naissance sur le périmètre des personnels de l'UPSaclay. Pour ce faire, nous avons utilisé l'outil OpenRefine qui permet de nettoyer, transformer, étendre les données en faisant appel à des services web et à des API.

Nous avons utilisé 3 types de système permettant des alignements et l'enrichissement des données :

- la réconciliation, par les services de réconciliation d'OpenRefine ;
- l'interrogation de *web services* ou API via la reconstitution automatique d'URLs ;
- l'interrogation de *triple stores* grâce à des requêtes en SPARQL.

Nous avons commencé par utiliser le service de réconciliation d'OpenRefine appliqué à l'API de Virtual International Authority File (VIAF, fichier

d'autorité international virtuel)[18]. Ce service permet de récupérer, à partir d'une liste de personnes identifiées par leur nom et prénom, une liste d'identifiants susceptibles de correspondre, chaque association étant complétée d'un indice de confiance qui varie entre 0 et 1. En étudiant les résultats, nous y avons constaté du bruit. Ceci est dû à des variations typographiques, par exemple sur l'usage des majuscules dans les noms et prénoms, ainsi qu'à des homonymies, etc. Il s'est avéré très difficile de vérifier la fiabilité des alignements, même avec le taux de confiance fourni par l'API. Nous avons pu remarquer, en travaillant manuellement sur des échantillons, que le bon alignement n'est pas toujours celui qui a le taux de confiance le plus élevé.

En parallèle, nous avons opéré le même travail de réconciliation avec l'API SolR de l'Abes[19] en reconstituant l'URL avec les différentes variables. Cette URL, construite automatiquement par OpenRefine pour chaque auteur, constitue une requête qui permet de moissonner les données au format JSON ou XML sur l'entrepôt de l'Abes. Mais là aussi, faute de filtres précis, il est difficile, à cause du bruit présent dans les résultats, de récupérer des informations validées, à savoir l'association de la bonne personne avec le bon identifiant. Seule une vérification manuelle, ligne par ligne, aurait permis de valider les alignements proposés. Ce travail sur plus de 10000 lignes n'a pas été retenu en première hypothèse.

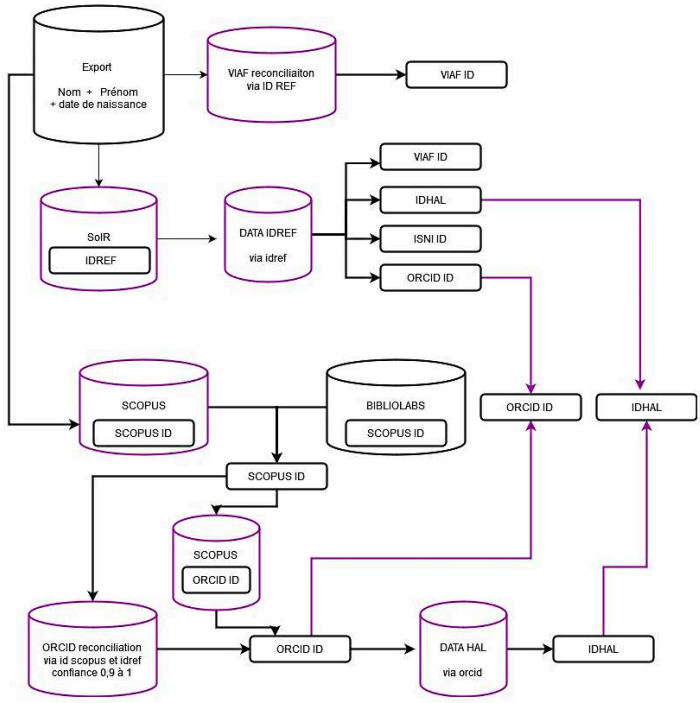
Nous nous sommes donc rapprochés de l'Abes. En effet, les données non accessibles au public dont l'Abes dispose permettent des alignements fondés sur plus de critères et de filtres que ceux ouverts à tous. Ce service a permis d'initier un processus complet d'enrichissement des alignements. En s'appuyant sur les informations de nom, prénom, date de naissance et en cherchant une mention de l'UPSaclay ou Paris-Sud dans certains champs de chaque notice, l'Abes a pu parvenir à un excellent taux de fiabilité pour la réconciliation avec IDRef. Ainsi, sur 6700 personnes à identifier, nous n'avons obtenu qu'environ 100 lignes incohérentes. L'obtention de ce premier identifiant, IDRef, nous a permis ensuite d'interroger data.idref[20] et d'aligner avec chaque personne ses identifiants (ISNI[21], VIAF, IdHal[22], ORCID iD) renseignés dans les données publiques de l'Abes, c'est-à-dire le Sudoc.

Puis nous avons aligné les données obtenues avec l'API de Scopus afin d'ajouter l'ID Scopus aux autres identifiants. Cette jointure a permis de s'assurer de la fiabilité de l'identifiant Scopus et d'enrichir les données alignées

18. < http://viaf.org/ >.
19. < http://documentation.abes.fr/aideidrefdeveloppeur/index.html#UtiliserApiSolr >.
20. < https://data.idref.fr/ >.
21. < https://isni.org/ >.
22. < https://doc.archives-ouvertes.fr/identifiant-auteur-idhal-cv/ >.

grâce à d'autres identifiants ORCID signalés dans cette base. Pour compléter nos alignements, nous avons ajouté une réconciliation via l'API ORCID en prenant comme paramètres les nom, prénom, date de naissance, IDRef et identifiant Scopus, pour optimiser la fiabilité. Ce service de réconciliation proposant également un indice de fiabilité, nous n'avons pris en compte, après quelques tests, que les ORCID alignés avec un indice supérieur à 90 % afin d'éliminer tout risque d'erreurs. Pour finir, nous avons interrogé data.hal[23] via l'identifiant ORCID et ainsi récupéré des alignements avec IdHal, pour les rares cas où les données issues d'IDRef et fournies par l'Abes lors du premier alignement n'étaient pas à jour. L'identifiant Scopus, présent dans chaque enregistrement de la table « auteur » de BiblioLabs, a servi de clé pour intégrer toutes les données alignées dans BiblioLabs, afin d'améliorer la collecte et la fiabilité des traitements réalisés.

Figure 3. Flux de travail d'alignement des identifiants auteurs



Ce schéma synthétise les étapes de traitement détaillées ci-dessus.

23. < https://data.archives-ouvertes.fr/ >.

Ces opérations, qui passent par les services de l'Abes, peuvent être déclinées pour chaque établissement composante ou associé de l'université, et répétées régulièrement au gré des mouvements de personnels. Pour ce faire, il suffit d'avoir un fichier tabulé structuré à l'identique et d'y appliquer le même processus sur OpenRefine. Ainsi, l'intégration d'identifiants chercheurs dans une base de données traduit aussi l'intégration des institutions du territoire dans le système d'information, et permet en retour de préciser la représentation du territoire.

DES DONNÉES EN QUÊTE D'AUTEUR

Cartographier le paysage : laboratoire, Graduate School et institutions

Les représentations visuelles symboliques d'un territoire de données tendent sous les effets combinés de la vogue des cartes heuristiques et des progrès de la visualisation de données à s'abstraire du territoire pour lui préférer une représentation par proximité conceptuelle. Ainsi, dans le cadre de Cartolabe[24], un projet de recherche commun avec un laboratoire de recherche en informatique de l'université (le LRI[25] devenu depuis LISN[26]), les données de recherche de l'université, photographiées à un moment précis, ont pu donner lieu à une représentation en deux dimensions, mêlant personnes, institutions, sujets, mots-clés et publications. Ce projet, expérimental et non destiné à produire une aide à la décision, a permis de mieux conceptualiser la cartographie des données du territoire, en dehors de la notion géographique de territoire.

De manière plus pragmatique, les frontières définies par les unités et par les disciplines, si elles se recoupent partiellement dans la donnée, peuvent être combinées pour former les frontières de nouvelles entités, à la fois administratives et disciplinaires. C'est le cas des Graduate Schools de l'université, entités parfois multidisciplinaires mais toujours thématiques, et dans une moindre mesure, liées aux unités de recherche de l'université. Leur création a donné lieu à une définition alliant plusieurs couches conceptuelles, et permettant une certaine porosité à la fois d'un côté et de l'autre, pour des entités non exclusives entre elles mais cependant bien repérables. Au lieu de travailler uniquement avec le nom des personnes qui se considéreraient comme affiliées à telle ou telle Graduate School, nous avons considéré avec la Direction

24. < https://cartolabe.fr/map/saclay >.
25. < https://scanr.enseignementsup-recherche.gouv.fr/entite/199812948M >.
26. < https://scanr.enseignementsup-recherche.gouv.fr/entite/202123711L >.

de la recherche que, dans un périmètre d'unités potentiellement concernées par la thématique d'une Graduate School, on pouvait définir par leurs mots-clés ou leur appartenance disciplinaire quelles publications relevaient, ou non, de l'entité considérée. Ici, ce n'est plus le paysage qui est mis en données mais des métadonnées qui composent et dessinent le paysage.

Cette approche autorise chaque unité de recherche de participer à un nombre potentiellement illimité d'entités thématiques, tout en permettant de classifier chaque publication dans une ou plusieurs de ces entités, afin de mieux comprendre l'évolution des thématiques de recherche dans l'université. Plus simplement, du point de vue institutionnel, chaque laboratoire ayant pour tutelles un ou plusieurs établissements, le lien entre les identifiants de laboratoires et les identifiants d'établissements permettent des cartographies non exclusives de l'activité scientifique par établissement, prenant en compte les collaborations internes aux unités, entre unités et entre établissements.

Les limites de l'approche territoriale : des publications sans auteur identifiable

Ce territoire appréhendé d'abord par ses structures académiques afin de récolter les publications et de dessiner le paysage de la recherche (domaines couverts, type de publications, titre des revues, éditeurs, *Open Access* des publications, financements etc.) en n'utilisant que des sources publiques a longtemps dû ignorer une dimension essentielle du paysage de la recherche : les auteurs. Ils ne sont pendant longtemps qu'un attribut peu fiable des publications : formes multiples des noms, attributions d'identifiants différents pour une même personne par les services auprès desquels les données étaient collectées, autant de facteurs qui font que la base auteurs n'est en fait qu'un enregistrement d'identifiants reliés à des publications lesquelles sont liées à un laboratoire. Une telle situation est liée en grande partie à un choix pragmatique et à une situation organisationnelle lors du développement de ces outils.

En effet, lier paysage institutionnel et production scientifique de façon pragmatique ne nécessitait pas de consolider les auteurs comme donnée fiable. De plus cette consolidation des données d'auteurs aurait été difficile voire impossible du fait que le lien entre chercheurs et laboratoires restait alors l'apanage de l'établissement et que ces données RH n'étaient pas et ne pouvaient pas être partagées. Cette situation n'a pas pu évoluer avant 2020.

À cette date, la COMUE portant le projet de l'UPSaclay fusionne avec l'université Paris-Sud. L'équipe[27] ayant travaillé sur BiblioLabs a alors accès à des données d'annuaire d'un établissement. Le travail de consolidation de la table auteur, pour tendre à en faire un référentiel fiable peut commencer. Mais il ne s'agit que d'un établissement parmi 14 autres qui restent maîtres de leurs données d'annuaire. Les limites liées aux contextes institutionnels viennent s'ajouter aux limites liées aux identifiants de structures, eux-mêmes limités par le facteur humain qu'est la qualité de la signature.

Une nécessaire diversification des sources et des missions

Dès lors, que penser d'un outil d'alignement qui ne fait qu'ajouter des précisions externes à une vision par définition diminuée, appauvrie ou partielle de la réalité ? La réponse est dans la confrontation de sources multiples, le dialogue organisé au sein d'une base de données entre de multiples bases dont les langages sont à l'occasion réconciliés. Pour BiblioLabs, la question de la multiplication des sources s'est posée avec une particulière acuité lorsque la politique d'ouverture de la science a cherché des indicateurs fiables et complets.

Le Baromètre de la science ouverte (BSO), proposé par ScanR et adapté en 2020 pour une logique locale et institutionnelle à l'initiative de l'université de Lorraine[28], afin que chaque établissement puisse s'en saisir, a été une première occasion de diversifier l'usage de BiblioLabs d'une part et de se confronter aux limites des sources de BiblioLabs d'autre part. En effet, le BSO interroge différentes sources afin d'établir le statut d'une publication à l'égard de la science ouverte (or, argent ou fermé). Parmi ces sources, on retrouve les mêmes que BiblioLabs auxquelles s'ajoutent d'autres sources plus nombreuses. Aussi, afin de produire un BSO le plus fiable et le plus exhaustif, nous l'avons fait tourner sur les références collectées par BiblioLabs, puis nous avons complété avec les outils natifs du baromètre interrogeant Hal, PubMed, Lens, etc. Le BSO ne collecte que des publications avec DOI, de ce fait, la collecte de publications même en diversifiant les sources ne peut pas être exhaustive. Et de fait, les publications collectées en plus par le baromètre étaient assez peu nombreuses en proportion mais suffisante pour souligner la nécessité d'ouvrir les sources de collecte.

Ainsi, un outil initialement réalisé pour traiter de bibliométrie et de pilotage de la recherche, par sa fonction visant à collecter et identifier des publi-

27. À ce jour, l'équipe a été renforcée par une personne ayant développé les compétences nécessaires dans OpenRefine afin de traiter les alignements des auteurs et qui a permis l'enrichissement de BiblioLabs sans travailler spécifiquement sur l'application ou la maintenance.
28. < https://scienceouverte.univ-lorraine.fr/barometre-lorrain-de-la-science-ouverte/ >.

cations sur un périmètre donné, est un outil tout à fait adapté à la production d'indicateurs sur la science ouverte. L'analyse des publications induit une connaissance de leur statut éditorial et donc de leur licence d'utilisation sur laquelle s'appuient les outils de mesure de la science ouverte. Paradoxalement, ce sont les mêmes outils et les mêmes mécanismes qui permettent de produire des indicateurs bibliométriques et des indicateurs d'adhésion à la science ouverte. Dans les deux cas, l'exhaustivité est limitée par un paysage éditorial et disciplinaire qui use diversement des identifiants et en particulier des DOI. Il reste que la capacité que nous avons eue de produire un baromètre à l'UP-Saclay quelques mois après la création de l'établissement a été grandement facilitée par l'existence d'une base bibliométrique ayant dessiné le périmètre d'activité de l'université et collecté les publications associées. Restait à en extraire le statut éditorial de ces publications.

Un second chantier lié au développement de la science ouverte a participé de l'évolution des usages de BiblioLabs et de sa nécessaire ouverture. L'archive ouverte HAL ayant gagné en France un statut incontournable, le choix a été fait d'aider tous les chercheurs de l'institution à y déposer leurs publications (notice et texte intégral). Ce service, qui bénéficie grandement du processus d'alignement des identifiants de personnes que nous avons décrit, profite également du lien entre auteurs et publications apporté par BiblioLabs.

Si les identifiants de structures sont restés plutôt simples à maîtriser grâce au contrôle donné par HAL aux institutions sur les structures dont elles sont tutelles, les identifiants de publications, jusque-là considérés comme faciles à traiter, ont dû gagner en exhaustivité. Beaucoup de publications présentes dans HAL, en particulier en SHS, n'ont pas de DOI car elles s'inscrivent dans une culture éditoriale et disciplinaire différente ou bien n'ont pas encore été validées et identifiées de manière unique.

Le constat fait à partir du service proposé autour de HAL se vérifie à chaque nouveau service : la qualité des données n'est envisageable que dans la pluralité de leurs sources, mais ceci pose pour chaque nouvelle source un nouveau défi technique d'alignement d'identifiants disparates. L'approche pragmatique permet le plus souvent d'obtenir des alignements fiables au détriment de l'exhaustivité, ou bien complets au prix d'interventions manuelles. La pluralité des sources est donc désormais acquise, mais les méthodes de réconciliation diffèrent : si les profils d'auteurs sont alignés automatiquement par le processus décrit au-dessus, les publications restent toujours vérifiées par les personnes qui les manipulent, même si le procédé informatique d'alignement permet de simplifier le travail.

CONCLUSION

De la même manière que la cartographie permet l'appropriation d'un espace géographique, la récolte de données permet la création d'un véritable territoire. Cependant, de la même manière que la cartographie classique représente un territoire géographique, la cartographie par des données représente un certain type de territoire : elle est l'expression d'une subjectivité dans les éléments représentés et dans le point de vue adopté. Les représentations peuvent être trompeuses, surtout hors contexte, et des imperfections sont inévitables.

On pourrait donc considérer comme une évidence que le réel préexiste à la donnée qui le représente : la recherche existait à l'UPSaclay avant BiblioLabs. En revanche, la représentation d'un espace par les données participe à son appropriation. BiblioLabs fait des publications de l'UPSaclay un territoire de données, au sens où ces données sont contrôlées et bornées par des frontières définies et des identifiants normalisés.

Enfin, l'outil évoluant en même temps que les données, la richesse de ces données est un levier important dans la construction de services à la communauté universitaire et aux services supports. Cette richesse est amenée à être partagée et multipliée, en s'appuyant en particulier sur les technologies du Linked Open Data.