ideas belong

Analyse des propositions soumises à l'appel flash ANR science ouverte : pratiques de recherche et données ouvertes

Gilles Dubochet
gilles.dubochet@ideasbelong.org

Cette étude a été financée par le Ministère de l'Enseignement supérieur de la Recherche et de l'Innovation et pilotée par l'Agence Nationale de la Recherche. Elle se fonde sur des données transmises par l'ANR. Son contenu, les choix méthodologiques, analyses, inférences et conclusions, ainsi que tous les graphiques, sont la responsabilité de l'auteur et ne représentent pas la position de l'État ou de l'ANR, ni n'engagent leur responsabilité d'aucune sorte.

Cette analyse porte sur les 100 propositions soumises à l'appel Flash.
Pour des informations sur les 25 projets lauréats, consultez la page dédiée : https://anr.fr/fr/lanr-et-la-recherche/engagements-et-valeurs/la-science-ouverte/projets-laureats-de-lappel-flash-science-ouverte/

18 novembre 2019

Résumé

L'appel à projets « Pratiques de recherche et données ouvertes » de l'ANR est une des premières mesures du Plan national pour la science ouverte. Les propositions qui y ont été soumises sont, en quelque sorte, une représentation de l'offre d'activités que la recherche française est en mesure de proposer, aujourd'hui et avec des moyens relativement limités, pour développer l'ouverture des données. Ce document en fait une analyse, afin de mieux comprendre l'état des forces, mais aussi les limites, d'une approche ascendante de l'ouverture des données.

Les propositions sont issues de façon presque équilibrée entre les trois domaines de la recherche : sciences humaines et sociales, vie et santé, et sciences et technologies. Toutefois, cela cache en réalité une offre fragmentée, portée par quelques communautés particulièrement actives, complémentée par une *longue traîne* d'initiatives individuelles.

Les trois quarts des propositions concernent le développement de plateformes, imaginées en premier lieu comme des « bases de données » pour la saisie, l'accès et l'interopérabilité. L'importance de cette offre résonne d'ailleurs avec le fait que dans la moitié des consortiums, les partenaires se définissent comme producteurs de données avant tout.

Si ces actions autour du développement de plateforme peuvent avoir pour effet de développer la FAIRisation des données, les principes FAIR ne sont toutefois que peu mentionnés dans les propositions et ne semblent par conséquent pas jouer un rôle structurant.

Les actions destinées à renforcer les communautés qui exploitent les données, ou le développement d'un écosystème à valeur ajoutée qui en facilite l'analyse, la curation ou la pérennisation, sont présentes dans l'offre, mais de façon nettement moindre. D'autres aspects, tels que la valorisation (par la société ou l'industrie) des données ouvertes, la protection des données sensibles, ou la facilitation des carrières dans un contexte de données ouvertes, sont pratiquement inexistants.

La moitié des consortiums sont structurés autour d'une collaboration entre des chercheurs dans un domaine scientifique collaborant avec des chercheurs en informatique, ou avec un service technique, ou avec une cellule d'appui à la recherche qui apporte l'expertise technologique. Ce type de collaboration entre des partenaires issus de cultures différentes, s'il permet des propositions innovantes, n'est pas sans difficulté : les modalités de la collaboration et les objectifs des propositions semblent mal définis dans plus de la moitié des cas. L'offre n'aborde que peu cette problématique, et seuls 10 % des propositions se fixent pour but de structurer une communauté, ou de faciliter la collaboration entre partenaires.

Seule une minorité de propositions se positionne par rapport à des initiatives internationales : GoFAIR, RDA, EOSC, etc. Un meilleur ancrage de futurs projets dans ces initiatives semble être une piste d'amélioration qui favorisait le positionnement stratégique d'une offre, par ailleurs assez ambitieuse.

Contexte

Le Plan national pour la science ouverte[1] du Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESRI), annoncé le 4 juillet 2018, a pour ambition de « de faire en sorte que les données produites par la recherche publique française soient progressivement structurées en conformité avec les principes FAIR[2], préservées et, quand cela est possible, ouvertes. »

Pour avancer vers cet objectif ambitieux, le ministère propose plusieurs mesures. D'abord, en rendant « obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics.[3] » Cette mesure impose un cadre règlementaire clair aux chercheurs et pose sans ambigüité l'importance politique de l'ouverture des données. Les chercheurs sont d'ailleurs confrontés à cette même obligation en ce qui concerne les fonds européens.

À cela s'ajoute une série de mesures d'encouragement destinées à « accélérer, coordonner, structurer et organiser ». Elles doivent permettre aux chercheurs de s'approprier l'ouverture des données, évitant ainsi que le Plan national soit perçu comme un simple exercice de conformité administrative. Ces mesures *ascendantes* doivent donc être portées par la communauté scientifique, et leur efficacité dépendra largement de ce que les chercheurs eux-mêmes pourront proposer. C'est dans ce contexte que cette étude s'interroge sur « l'offre » en matière de structuration, de partage et d'ouverture des données qui existe aujourd'hui en France.

L'appel à projets « pratiques de recherche et données ouvertes » de l'Agence National de la Recherche (ANR), publié le 28 mars 2019 et dont les résultats ont été annoncés le 18 juillet, faisait partie des premières mesures incitatives. Les objectifs de cet appel Flash[4] étaient définis d'une façon large qui encourageait « la communauté scientifique elle-même de proposer, domaine par domaine, discipline par discipline, spécialité par spécialité, comment elle peut appliquer les principes de la science ouverte à propos des données de la recherche. » L'appel se concentrait sur des projets pouvant être mis en œuvre rapidement et relativement facilement, du fait du délai très court pour la soumission et d'une enveloppe budgétaire qui ne dépasse pas 100 000 € sur deux ans.

Cent propositions ont été soumises, et vingt-cinq projets ont été financés. La liste des propositions financées est disponible sur le site de l'ANR[5].

Le format ouvert de l'appel ANR posait, en substance, la question suivante aux chercheurs : avec des moyens relativement limités, quelles sont les actions que vous êtes en mesure de proposer, aujourd'hui, pour développer l'ouverture des données ? L'analyse de

[1] http://www.enseignementsup-recherche.gouv.fr/cid132529/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-a-tous-sans-entrave-sans-delai-sans-paiement.html

[2] FAIR : Facile à trouver, Accessible, Interopérable, Réutilisable

[3] Dans les limites du principe « aussi ouvert que possible aussi fermé que nécessaire »

[4] https://anr.fr/fileadmin/aap/2019/aap-data-2019-dossier.docx

[5] https://anr.fr/fileadmin/aap/2019/aap-data-2019-selection.pdf

l'offre composée des propositions soumises peut nous permettre de lire, en filigrane, la réponse à cette question.

L'analyse présentée dans ce rapport est basée sur une étude dans laquelle l'auteur a individuellement analysé le texte de chaque proposition pour le coder dans une grille d'analyse évaluant une cinquantaine de paramètres. La grille inclut des paramètres objectifs (par exemple si la participation à un groupe RDA est mentionnée), mais aussi plus subjectifs (quel rôle semble jouer un partenaire dans le consortium).

Cette méthodologie a permis d'étudier des caractéristiques et de faire émerger des tendances qui n'auraient pas été apparentes par une approche basée sur les données structurées administratives ou des mots-clés. Elle permet également d'obtenir des données comparables à travers un corpus de propositions très diverses.

Cette démarche valorise ce que les porteurs de projets eux-mêmes choisissent de mettre en avant. Il s'agit donc d'une étude de l'offre, mais aussi de la façon dont la communauté présente cette offre.

Quels domaines ?

Des données sont exploitées dans les trois grands domaines de la recherche[6]. Il est donc naturel qu'ils soient tous représentés dans l'offre. Il est néanmoins remarquable que l'appel ANR ait pu mobiliser une offre d'importance comparable dans tous les domaines (Figure 1), démontrant ainsi l'existence d'une volonté à ouvrir les données qui surpasse les barrières disciplinaires.

Si l'on considère les domaines à un niveau plus fin, il ressort toutefois une inégalité importante. Ainsi, plus d'un tiers de l'offre est issue de trois sous-domaines : les sciences du système terre (PE10) ; l'étude du passé humain (SH6) ; et les sciences appliquées de la vie, biotechnologie et ingénierie moléculaire (LS9).

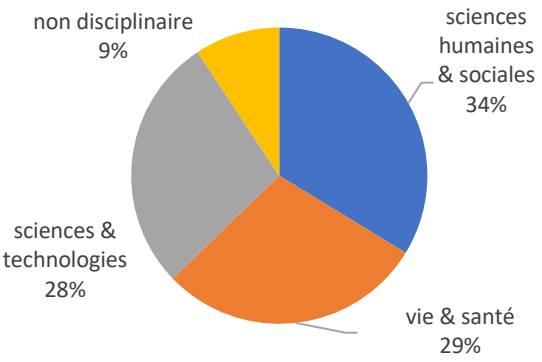

Figure 1: Discipline des données

De plus, la forte représentation de ces sous-domaines ne s'explique pas par une préférence générale pour l'ouverture des données, mais en premier lieu par l'engagement dans ceux-ci de communautés bien identifiables, structurées autour d'une question de recherche particulière et porteuses d'une vision partagée du rôle des données — par exemple, la conservation du patrimoine dans le cas du sous domaine « étude du passé humain ».

Au contraire, 14 des 25 sous-domaines de recherche sont soit absents de l'offre soit portée par seulement une ou deux propositions.

[6] La taxonomie ERC est utilisée pour la définition des domaines et sous-domaines scientifiques : https://erc.europa.eu/sites/default/files/document/file/ERC_Panel_structure_2019.pdf

Notons que quelques communautés pourtant actives dans l'ouverture des données n'ont pas été mobilisées par l'appel ANR. On peut supposer qu'elles disposent d'autres soutiens pour l'ouverture des données — par exemple dans les sciences de l'univers à travers les grandes infrastructures d'observation.

En l'état, on peut dire que l'avancement de pratique de recherche et de données ouvertes est ainsi porté par quelques communautés particulièrement actives, complémentées par une *longue traîne* d'initiatives individuelles. Il est donc utile de considérer si ces communautés peuvent stimuler l'ouverture des données dans des disciplines voisines, par exemple, par la mise en commun des outils et des pratiques de partage de données entre communautés qui observent le système-terre.

Notons également qu'il existe dans l'offre un certain nombre de propositions qui se définissent par rapport à un but général de développement de l'ouverture des données, sans s'attacher à une discipline ou un type de données en particulier. On y trouve des projets pour développer des outils de gestion des données, des plateformes pour faciliter un processus de recherche ouvert, et même quelques projets qui proposent de repenser fondamentalement la pratique scientifique autour du partage des données. Quelle place donner à ces propositions souvent ambitieuses et potentiellement transformatrices pour les pratiques de recherche, mais dont l'impact reste spéculatif ?

Quelles propositions ?

Comment aborder un concept abstrait tel que les données ouvertes ? Dans les trois quarts des cas, les propositions répondent à cette question en développant un outil concret : une *plateforme* de partage des données spécifique.

Parmi ces projets, la majorité (60 %) a pour but d'améliorer ou d'étendre une plateforme existante. Ces propositions ont l'avantage de pouvoir construire sur des moyens et des communautés préexistants. Le reste propose de développer de nouvelles plateformes.

L'importance des propositions qui développent des plateformes spécifiques démontre l'effet structurant fort qu'elles peuvent avoir. En effet, les plateformes rendent les données ouvertes visibles, ainsi que le travail des consortiums. La multiplication des plateformes crée néanmoins un risque de fragmentation des données.

¾

des propositions développent une plateforme spécifique

Le quart restant des propositions abordent les données ouvertes du point de vue, non pas de l'outil, mais des pratiques. On y trouve des propositions pour développer des recommandations, y compris en matière d'utilisation de plateformes existantes, par exemple sous forme d'ontologies[7] ou

[7] Une ontologie est l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, par exemple dans une discipline ou un champ de recherche.

pour accompagner la FAIRisation[8] des plateformes. On y trouve également des propositions qui se placent encore plus en amont dans le processus de structuration, en engageant une communauté dans une réflexion autour des données ouvertes.

Agir pour et avec les données ouvertes

L'appel ANR laissait aux propositions une grande liberté pour « appliquer les principes de la science ouverte à propos des données de la recherche ». Quelles actions ont été proposées en réponse à cette demande ? Une analyse plus détaillée de l'offre peut nous aider à mieux comprendre ce sur quoi les porteurs de projets se sont concentrés pour faire avancer l'ouverture des données.

Pour obtenir une vue synthétique d'une offre composée de propositions très diverses, nous ne pouvons pas nous satisfaire d'une liste à la Prévert. Nous avons donc créé un modèle d'analyse qui propose de distinguer 11 dimensions d'action. Ce modèle est un choix méthodologique, validé en le recoupant avec les actions recommandées par des standards de bonne pratique existants[9]. Ces dimensions ne représentent pas des actions similaires, mais représentent différentes *façons d'appréhender* l'ouverture des données. Certaines dimensions pensent ainsi l'ouverture comme un problème technologique avant tout alors que d'autres l'envisagent comme une question sociale. Certaines dimensions mettent l'accent sur les producteurs de données, d'autres sur les utilisateurs — nous détaillerons ce point plus bas.

Ce modèle permet ainsi de distinguer des actions qui pourraient sembler similaires, mais qui représentent en réalité une façon différente d'appréhender l'ouverture de données. Par exemple, le développement d'une interface de programmation n'a pas le même sens si le but est d'automatiser le téléchargement de données depuis un réseau de capteurs — une perspective qui valorise l'accumulation de données — ou si le but est d'intégrer la plateforme de données avec une plateforme de publication pour faciliter la citation des données — la perspective étant ici la reconnaissance de la contribution des chercheurs qui produisent des données.

L'importance relative dans l'offre de chaque dimension est visualisée dans la Figure 2. Les dimensions sont décrites en dessous.

[8] La FAIRisation est la mise en conformité avec les principes FAIR : des données faciles à trouver, accessibles, interopérables et réutilisables.

[9] En particulier, nous avons référencés les exigences du standard CoreTrustSeal : https://www.coretrustseal.org

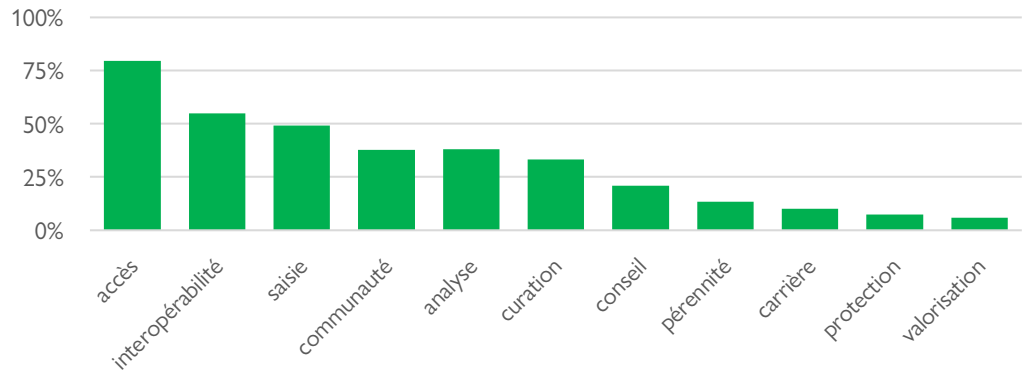

Figure 2 : Actions privilégiées pour l'opérationnalisation des données ouvertes

Accès (présent dans 67 % des propositions) : le projet facilite l'accès, la recherche et la sélection des données. Il collecte ou produit les métadonnées nécessaires. Il propose des méthodes d'accès aux données et métadonnées adaptées aux besoins de recherche des utilisateurs, c'est-à-dire par le téléchargement de fichiers dans un format approprié, par une interface web ou par une interface de programmation.

Cette dimension est la plus représentée dans l'offre. On retrouve en particulier dans un grand nombre de projets un intérêt à identifier et donner accès à des métadonnées.

Interopérabilité (45 %) : le projet facilite la fusion des données par la mise en œuvre des standards d'interopérabilité, ou directement en agrégeant plusieurs sources.

Si cette dimension est souvent présentée comme la mise en œuvre d'un objectif technique, l'offre explore par contre assez peu comment la possibilité de croiser des données s'intègre dans des pratiques de recherche spécifiques.

Saisie (41 %) : le projet facilite la saisie en acceptant des données codées sous forme de fichier dans un format approprié, ou à travers une interface de programmation. Il propose des moyens pour assurer l'intégrité des données et pour enregistrer leur origine. Il peut également proposer des modèles de plans de gestions de données qui facilitent la mise en adéquation des besoins de la saisie de données avec les pratiques de recherche.

Analyse (35 %) : le projet propose des outils d'analyse ou de visualisation spécifiques et adaptés aux compétences techniques et aux besoins de recherche des utilisateurs.

Quoique présent dans un tiers des propositions, le développement de moyens d'analyse des données ouvertes reste néanmoins considéré dans la plupart des cas comme étant de la responsabilité des utilisateurs. On peut s'interroger si cela s'explique par l'existence d'outils déjà satisfaisants, ou s'il s'agit d'une limite à ce qui est perçu comme entrant dans le cadre de l'ouverture des données ?

Communauté (33 %) : le projet s'engage pour structurer les communautés scientifiques qui produisent et exploitent des données. Il promeut activement la convergence des activités de recherche dans des projets de recherche intégrés. Les porteurs de projet se positionnent comme médiateur dans des processus destinés à définir des standards ou des pratiques liées aux données.

En premier lieu, on retrouve ici des propositions visant le développement de standards en collaboration avec la communauté concernée. Cette forme d'engagement permet de lancer une discussion, et joue sans doute un rôle structurant important. Néanmoins, seules quelques propositions mettent en œuvre des approches véritablement participatives pour engager les communautés, la plupart se contentant de questionnaires ou de séminaires.

Curation (30 %) : le projet met en œuvre des moyens pour organiser, évaluer ou améliorer automatiquement la qualité des données préexistantes, en concertation avec les communautés concernées.

Conseil (18 %) : le projet met à disposition des utilisateurs des données un conseil ou des formations individualisées et spécifiques à leur pratique de recherche. Ceux-ci peuvent provenir de l'expertise propre des porteurs du projet, ou par la mise en réseau d'experts issus de la communauté scientifique.

On peut s'interroger si des utilisateurs peuvent s'emparer du résultat de projets qui développent des technologies complexes pour la saisie, la curation, l'accès, l'interopérabilité ou l'analyse des données, sans qu'ils soient aidés pour le faire.

Pérennisation (12 %) :

Il faut ici faire la distinction entre deux aspects liés à la pérennisation. La statistique ci-dessus concerne la conservation des données, c'est-à-dire là où le projet développe un moyen d'archivage et un modèle de financement sur le long terme qui assure une conservation pérenne des données. Si toutes les données ne peuvent être conservées, le projet dispose d'un plan de préservation clair et adapté aux besoins de recherche.

L'autre aspect, lié au précédent, mais néanmoins distinct, concerne la pérennisation des activités de la proposition, c'est-à-dire la possibilité de continuer le projet au-delà des 24 mois financés par l'appel ANR ou, le cas échéant, de garantir que les services développés dans le cadre du projet restent disponibles. L'appel demandait aux porteurs de projets d'identifier des solutions pour pérenniser les activités de la proposition, mais seul environ un quart ont pu identifier une solution potentielle concrète, par exemple par leur intégration dans une plateforme pérenne. Par contre, un quart des propositions ont été conçues pour se terminer dans les 24 mois qui pouvaient être financés, et ne requiert donc pas de stratégie de pérennisation.

La question de la pérennisation, qu'il s'agisse de l'archivage des données ou de faire vivre un projet au-delà de 24 mois, apparait fortement comme une barrière pour les initiatives d'ouverture des données. L'appel Flash, qui n'apporte pas de solution à cette problématique, devra sans doute être complété par d'autres formes de soutien qui restent toutefois à définir.

Carrière (9 %) : le projet s'engage pour que les carrières du personnel qui s'engage pour l'ouverture des données ne soient pas entravées. Il fournit des outils, impose des règles ou favorise des pratiques pour que la contribution de toutes les personnes impliquées dans un écosystème de données ouvertes soit correctement reconnue.

Cela concerne plusieurs catégories de personnel. Des chercheurs, qui doivent être reconnus pour les données qu'ils produisent (citations des données), mais aussi pour les contributions méthodologiques mises en œuvre sur des plateformes de données ouvertes (par exemple avec les humanités numériques ou pour des chercheurs en informatique dans des consortiums interdisciplinaires). Cela concerne aussi les ingénieurs, conservateurs, employés dans des cellules de soutien à la recherche, etc., dont la contribution reste souvent mal comprise, ou peu reconnue par les hiérarchies.

Protection (7 %) : si des données à caractère personnel sont collectées, si leur utilisation est restreinte par des licences, ou si l'éthique scientifique l'exige, le projet met en œuvre des solutions d'exploitation conformes en concertation avec les communautés concernées.

La plupart des propositions soumises dans le cadre de l'appel ne traitent pas de données qui doivent être protégées, et ne semblent en effet pas particulièrement concernées. On peut néanmoins s'interroger si la faiblesse relative de cette dimension ne risque pas de limiter le partage d'expérience avec des communautés où les données sont de nature personnelle, par exemple dans la médecine ou les sciences sociales.

Valorisation (6 %) : dans ce dernier cas, le projet promeut l'exploitation des données par des utilisateurs commerciaux ou sociétaux.

De façon générale, on peut dire que les actions privilégiées pour l'ouverture des données sont celles qui mettent en avant le développement technique de plateformes conçues comme des « bases de données » : la saisie, l'accès et l'interopérabilité. Cela s'explique sans doute, car ces dimensions sont le plus souvent celles qui peuvent être mises en œuvre immédiatement par un petit consortium. Mais le résultat est une offre qui considère l'ouverture des données en tant que telle, plutôt que les pratiques de recherche qui en découlent.

Pour conclure cette section, nous constatons que le modèle proposé pour cette étude donne une vue d'ensemble synthétique des actions pour l'ouverture des données. À postériori, nous aurions pu ajouter les « ontologies » comme une dimension distincte. En effet, qu'il s'agisse de la création d'une taxonomie ou du développement d'un modèle RDF, le souhait de certaines propositions de poser un modèle ontologique pour les données est apparu comme une *façon d'appréhender* l'ouverture des données assez spécifique et distincte d'autres dimensions comme la saisie ou l'accès, notamment.

On peut aussi envisager l'utilité de ce modèle pour structurer le soutien à la science ouverte. En effet, les diverses façons d'agir pour l'ouverture des données impliquent des besoins différents en matière de financement et de gouvernance.

Les principes FAIR

Le Plan national pour la science ouverte recommande l'adoption des principes FAIR pour des données « faciles à trouver, accessibles, interopérables, et réutilisables ». Toutefois, en l'état actuel de l'offre, ces principes ont un effet structurant faible. En effet, moins de 10 % des propositions sont explicitement structurées par un processus de FAIRisation. Et

10 %

des propositions développent une stratégie FAIR explicite

moins de la moitié mentionnent un engagement, même minimal, par rapport aux principes FAIR.

Les principes FAIR ont l'avantage d'offrir des directives claires sur le seuil attendu en matière d'action pour l'ouverture des données. Sans répéter ici les arguments qui sous-tendent ces principes, on peut dire qu'une base de données qui se trouve être hébergée sur un serveur accessible au public, sans que l'opérateur s'engage pour en faciliter l'accès, ne répond pas aux objectifs de la science ouverte. En cela, les principes FAIR créent une ligne de démarcation, imparfaite, mais dont l'avantage est d'exister, entre les propositions qui s'engagent véritablement dans une logique de données ouvertes de celles qui bénéficient de l'image — et des soutiens — de la science ouverte sans véritablement y participer.

Comment expliquer la faible influence des principes FAIR dans l'offre ? On peut imaginer que cela s'explique toujours par un manque d'information des porteurs de projets. Il est toutefois légitime de se demander si les principes sont en adéquation avec les besoins réels de chercheurs ou, au contraire, s'ils jouent correctement leur rôle en signalant une forte proportion de propositions qui ne s'engage pas véritablement dans une logique de science ouverte.

Quels consortiums ?

Si 15 % des propositions sont portées par un seul partenaire, la majorité est portée par des consortiums impliquant plusieurs partenaires[10]. On retrouve 35 % de petits consortiums, composés de 2 ou 3 partenaires. 30 % de consortiums de taille moyenne avec 4 à 6 partenaires. Et environ 20 % de consortiums plus grands.

Les projets seuls ou avec des petits consortiums représentent donc environ la moitié de l'offre. On peut s'interroger si ceux-ci sont adaptés aux objectifs de l'appel, qui met en avant la structuration de la communauté scientifique. Une analyse plus détaillée permet d'y répondre en partie : comme nous l'avons vu en dessus on retrouve dans l'offre beaucoup de propositions avec des objectifs de développement technique spécifiques, qui peuvent s'atteindre plus efficacement dans un consortium restreint. Ces petits consortiums sont donc bien adaptés.

Néanmoins, au-delà de ces cas, il ressort de l'offre une certaine difficulté des porteurs de projets à mobiliser des partenaires pour former leurs consortiums. Par exemple, on retrouve régulièrement dans les propositions des partenaires dont la participation semble importante au succès du projet, mais qui ont seulement « été contactés » ou qui sont « intéressés » sans être formellement intégrés dans le consortium. Cela s'explique sans doute en partie parce que l'appel Flash laisse, par définition, peu de temps. Mais on peut supposer qu'une des raisons qui a permis aux communautés les plus avancées dans l'ouverture des

[10] Nous ne considérons ici pas uniquement les partenaires qui demandent des financements, mais de tous les partenaires mentionnés dans la proposition et qui semblent y jouer un rôle effectif.

données de prendre une place prépondérante dans l'offre est simplement que les personnes intéressées dans ces communautés se connaissaient et ont ainsi pu former des consortiums rapidement. Dans ce cas, une réflexion sur la façon d'aider à se rencontrer les acteurs intéressés par l'ouverture des données au sein des communautés moins avancées, semble particulièrement importante.

Plus de la moitié des consortiums sont le résultat d'une collaboration entre partenaires scientifiques — c'est-à-dire des chercheurs de la communauté qui produisent ou utilisent les données — associés à des partenaires « fournisseurs de technologies ». Parmi ces derniers, on trouve des laboratoires qui conduisent des recherches autour du traitement des données (laboratoires d'informatique, voir groupes d'humanités numériques). Cette forme particulière d'interdisciplinarité entre chercheurs dans un domaine technologique avec des chercheurs d'un autre domaine, est assez commune dans l'offre. Dans d'autres cas, c'est un service informatique, un opérateur d'infrastructure de données ou une cellule d'appui ou d'encadrement de la recherche qui apporte son expertise comme partenaire technologique.

L'importance de ces consortiums mixtes interdisciplinaires démontre le dynamise qui existe au niveau des « fournisseurs de technologies » pour faire avancer l'ouverture des données. Seul un quart des projets sont portés par des consortiums uniquement scientifiques.

Rôle des producteurs

Si les consortiums sont essentiels à la mise en œuvre des propositions, l'impact scientifique ou sociétal que l'ouverture des données peut avoir dépend, en fin de compte, de la façon dont les producteurs et utilisateurs des données — s'emparent des nouvelles capacités développées par les projets.

½

des consortiums proposés sont portés par des « producteurs » de données

Pour analyser cet aspect de l'offre, on peut considérer qu'un chercheur peut jouer trois rôles distincts dans l'écosystème des données ouvertes. Le premier concerne la structuration ou le développement des *capacités* pour l'ouverture des données. Le second est un rôle de producteur, qui se focalise sur la création de données ouvertes à proprement parler. Finalement, on attend du rôle d'utilisateur qu'il exploite ces données pour faire de la recherche ou obtenir un impact sociétal. En pratique, ces rôles se chevauchent souvent ; un producteur est presque toujours aussi un utilisateur. Il n'en reste pas moins que ces rôles permettent de nommer des priorités qu'il est utile de distinguer, pour comprendre la façon dont l'offre se positionne dans l'écosystème des données ouvertes.

Nous considérons donc la façon dont les partenaires eux-mêmes décrivent leur rôle ou leur expertise dans les textes des propositions. Il en ressort que la moitié des consortiums sont uniquement ou très majoritairement composés de partenaires présentés en premier lieu dans un rôle de producteur. Dans un quart des cas, les consortiums se définissent par une dialectique entre producteurs et utilisateurs. Finalement, le quart restant est com-

posé de consortiums « techniques », qui se définissent par leur expertise à fournir des capacités pour l'ouverture des données, sans que les producteurs ou les utilisateurs ne fassent véritablement partie du consortium.

La place prépondérante des producteurs dans l'offre pourrait expliquer l'importance des actions de nature techniques, c'est-à-dire celles liées aux dimensions d'accès, d'interopérabilité, et de saisie. On peut imaginer que ce sont les producteurs qui identifient la « base de données », c'est-à-dire le *lieu* où les données qu'ils produisent sont collectées, comme l'élément structurant de l'ouverture des données.

Maturité de la collaboration dans les consortiums

Un autre aspect que nous avons souhaité analyser est la façon dont les partenaires collaborent au sein des consortiums. S'il est difficile à analyser, ce point nous a semblé important, car les propositions de l'offre se différencient des projets de recherche *classiques* dans leur structure et dans le type de collaboration qui les sous-tendent. Les consortiums composés de partenaires scientifiques et technologiques dépendent bien entendu de leur capacité à mettre en commun leurs compétences et leurs ressources propres, sans que l'un ou l'autre s'approprie le projet, alors même que leur culture ou leur intérêt dans le projet peuvent être assez éloignés. Au-delà de ça, l'ouverture de données est fondée sur la notion de mise en commun des pratiques de recherche, et donc de compromis entre partenaires.

55 %

des propositions ne définissent pas les modalités de collaboration au sein des consortiums

La qualité de la collaboration pendant le projet ne peut, bien entendu, pas être évaluée sur la base de l'offre. Elle dépendra toutefois fondamentalement de l'organisation de la collaboration et de l'existence d'une *bonne volonté* que le consortium aura construite ou négociée dès le départ.

Pour mieux saisir cet aspect complexe, nous considérons donc également comment sont définis les objectifs du projet, et en particulier s'il ressort de ceux-ci que chacun des partenaires du consortium y trouve une place.

Dans 15 % des cas, la proposition semble correspondre aux objectifs d'un seul partenaire sans qu'il ressorte clairement pourquoi les autres partenaires du consortium y participent. Dans environ 40 % des cas, on trouve des objectifs généraux partagés — par exemple de promouvoir l'ouverture des données dans une communauté — sans toutefois qu'il apparaisse comment les compétences ou les ressources de chacun y contribuent. Ce n'est donc qu'une petite moitié des consortiums, soit 45 %, qui s'engagent dans un projet avec une structure de collaboration relativement bien définie.

La problématique de l'ouverture de données, dont les contours restent flous pour une grande partie de la communauté scientifique, semble un défi pour la formation des consortiums. Si l'offre contient quelques projets dont l'objectif premier est la structuration d'une communauté, ceux-ci restent très minoritaires. Les communautés moins expérimentées dans le partage des données seront renforcées en favorisant ce type d'actions. Cela permettra également de mieux comprendre quels processus sont les mieux adaptés pour

engager les différents partenaires de l'ouverture des données dans des modèles effectifs de collaboration.

Renforcer les infrastructures et les réseaux existants

Dans cette dernière section, nous abordons comment les propositions se positionnent face aux principaux réseaux, infrastructures et autres outils de pilotage de la science ouverte en Europe ou dans le monde, que nous décrirons dans cette section par le terme général « d'initiative stratégique ». Les groupes de travail RDA, les réseaux d'implémentations GOFAIR, et l'opérationnalisation d'EOSC (la plateforme européenne de science ouverte) ont été identifiés dans l'appel comme essentiels. On peut y ajouter les infrastructures de données structurées par le pilotage ESFRI ou sous forme d'un ERIC ou d'un réseau COST, ainsi que celles qui s'engagent dans la coalition SCOSS.

Seul un quart des propositions se positionnent dans le texte de leur proposition par rapport à une initiative stratégique, généralement en mentionnant le bénéfice d'un partage d'expérience. Moins de 10 % précisent les modalités d'un engagement concret : participation dans un groupe de travail, contribution à un projet, etc. Il ressort donc de cela que près du trois quarts des propositions ne précisent pas explicitement qu'elles pourraient se positionner dans le contexte d'initiatives stratégiques.

Conclusion

Le paysage que peint cette étude de l'offre en matière de données ouvertes est nuancé. Il en ressort néanmoins plusieurs points forts.

Quelques communautés particulièrement actives portent la plus grande partie de l'activité. Elles proviennent des sciences du système terre, de l'étude du passé humain, et des sciences appliquées de la vie. Notons que quelques autres communautés, par exemple en astrophysique, sont également actives, mais n'ont pas été mobilisées par l'appel. À cela s'ajoute une forte communauté composée d'acteurs intéressés par la *technologie* de l'ouverture des données — chercheurs en informatique, ingénieurs, ou personnel de support à la recherche intéressés engagés dans la science ouverte. Quant aux autres disciplines, on y retrouve quelques individus intéressés, sans qu'ils soient en mesure de s'entourer suffisamment pour créer une masse critique. On peut donc considérer que le soutien à l'ouverture des données existe selon trois axes : en donnant aux communautés les plus avancées les moyens de démontrer qu'une pratique de recherche avec des données ouvertes répond aux exigences de l'excellence scientifique ; en aidant des communautés proches à bénéficier du partage d'expérience pour développer leurs propres pratiques ; en soutenant les individus isolés dans des contextes où la mobilisation de la communauté, plutôt que le développement technologique, est le facteur décisif.

¾

des propositions n'ont pas de positionnement stratégique dans des initiatives internationales

Un autre point fort concerne les actions privilégiées pour l'ouverture des données. Il ressort de cela que l'offre est particulièrement ambitieuse en ce qui concerne le développement de plateformes de données ouvertes envisagées du point de vue de la saisie et de l'accès aux données — des *bases de données* avant tout. Il s'agit là de la mise en place d'une infrastructure fondamentale, puisqu'elle collecte les données produites et est le point de départ pour toutes les activités subséquentes de mise en valeur. D'ailleurs, il ressort aussi de l'étude que les partenaires des consortiums s'identifient avant tout dans un rôle de producteur de données, mettant là aussi l'accent sur la collecte des données. On peut s'interroger si la multiplication des plateformes, portées par des très petits consortiums, est un risque. Néanmoins, la capacité démontrée des communautés, portées par les producteurs des données, à développer l'infrastructure nécessaire renforce l'idée qu'une approche ascendante de l'ouverture des données, comme dans l'appel Flash, peut être envisagée.

Si l'offre est convaincante en matière d'infrastructure, elle est plus faible en ce qui concerne les aspects d'engagement communautaire, de collaboration entre partenaires, et de mise en valeur scientifique ou sociétale des données. Ces aspects plus intangibles sont particulièrement difficiles à structurer dans une proposition. Les communautés les plus actives semblent être celles où, pour des raisons diverses, le partage des données fait déjà partie des pratiques de recherche habituelles. L'ouverture des données est donc simplement l'extension d'une pratique existante, et se fonde sur des technologies, des méthodes et des *socialités* préexistantes dans la communauté.

Au contraire, si l'ouverture des données doit devenir un moteur pour favoriser de nouvelles pratiques de recherche, dans un objectif plus général de science ouverte, la capacité à créer de l'engagement au sein des communautés devient l'élément crucial. Plusieurs propositions essayent de porter ce changement culturel et de pratique, proposant des approches collectives qui essayent de motiver la communauté à s'approprier l'ouverture des données, ou au contraire, en essayant de développer une plateforme suffisamment attractive pour que la communauté s'y engage pour des raisons utilitaires. Dans tous les cas, le succès de ces initiatives dépendra des pratiques de collaboration que ces consortiums pourront mettre en place.

Quelles seront les approches qui permettront de favoriser cette collaboration, sans remettre en cause l'indépendance des chercheurs ? Nous pensons qu'il s'agit là de la question qui sous-tend toute la science ouverte, qui se différencie d'une pratique de recherche classique par la mise en valeur du « génie collectif » issu de la collaboration, plutôt que le « génie individuel ».

À propos de l'auteur

Titulaire d'un doctorat en informatique, Gilles Dubochet a plus de 10 ans d'expérience dans la gouvernance de la recherche. Il a un intérêt particulier pour la science ouverte et l'interfaçage science-société-politique. Il a notamment travaillé sur ces thèmes à Science Europe et à l'EPFL.

Il est le cofondateur de « *Ideas belong* », une société dont le but est de soutenir la collaboration au sein de projets multipartenaires à l'interface science-société-politique. L'approche « *Ideas belong* » aide des partenaires dont les objectifs, la culture, et les motivations diffèrent à mettre en commun leur perspective et leur capacité propres. Partant du constat que les conditions d'une collaboration fructueuses doivent se construire en amont du projet lui-même, elle propose des outils innovants pour soutenir cette première étape souvent négligée. Appuyée sur un modèle distinctif de l'engagement multipartenaire, elle donne du *sens* à la présence de chaque partenaire et augmente l'impact du projet.

https://www.ideasbelong.org/

Pour contacter l'auteur : gilles.dubochet@ideasbelong.org