Les données d'ANPERSANA : un entrepôt pour la valorisation de la recherche à IKER

Aitor Fuentes Zamalloa

Université Paul Valéry – Montpellier III

ITIC

Département Information Documentation

Année 2016 – 2017

Les données d'ANPERSANA : un entrepôt pour la valorisation de la recherche à IKER

Aitor Fuentes Zamalloa

Directrice de mémoire : Céline Paganelli

Master2 Information et documentation : Gestion de l'information et médiation documentaire

REMERCIEMENTS

Je veux d'abord remercier ma directrice de mémoire Céline Paganelli pour sa réactivité dans nos échanges et ses conseils d'experte. Un grand merci aussi à Hans Dillaerts ainsi qu'au service EAD de l'Université Paul Valéry – Montpellier III, notamment à Louis Grand, pour leur flexibilité et humanité quand je me suis vu confronté à des difficultés en dehors du contexte académique.

Le stage dans le centre de documentation du laboratoire de recherche sur la langue et les textes basques, IKER UMR 5478 (CNRS-UBM-UPPA) a été le cadre idéal pour nourrir ma réflexion pendant l'écriture de ce mémoire. Je présente ma gratitude au directeur de la structure Ricardo Etxepare pour m'avoir donné cette opportunité, ainsi qu'à Urtzi Etxeberria, Anne-Marie Benarab et Karim Ait-Alouache pour avoir mobilisé les moyens nécessaires pour mon bien-être pendant les journées d'études, formations et ma présentation de cette dernière semaine.

Je souhaite remercier spécialement mon référent de stage Jean-Philippe Talec, toujours prêt à transmettre son expertise et à échanger autour du thème de la documentation, voire même sur mes perspectives professionnelles.

Mes remerciements vont aussi à toute l'équipe de chercheurs d'IKER qui m'a si bien intégré et accueilli, et qui a favorisé l'avancement de mes recherches en répondant au questionnaire ; merci notamment à ceux qui ont participé aux entretiens me consacrant une bonne partie de leur temps.

Je voudrais également nommer ici les personnes que j'ai rencontrées pendant la durée du master, que je n'ai pas citées avant et qui m'ont aidé, en le sachant ou pas, de façons diverses : María Arana, Elise Dillet, Aines Dufau, Maitena Duhalde, Irantzu Epelde, Corine Haristoy, Michel Jacobson, Beñat Lascano, Sergio Monforte, Nicole Mounier, Lourdes Oñederra, Alain Viaut et Xarles Videgain.

Je souhaite terminer en remerciant ceux qui m'ont soutenu davantage ces derniers mois, et auparavant. A mes parents Ana et Carlos, pour leur soutien dans mes choix ; et à Eluska qui m'a encouragé sans cesse dans les moments de doute, qui a vécu avec moi toutes les phases et pour tout ce qu'on partage.

RESUME

ANPERSANA est la bibliothèque numérique créée par le laboratoire IKER pour valoriser les données produites au cours des projets de recherche de sa communauté scientifique. Ce projet vient s'ajouter aux multiples démarches en faveur du libre accès de cette structure, tel que la revue *Lapurdum* et la plateforme d'auto-archivage Artxiker. Dans un contexte international d'émergence de projets pour l'ouverture des données scientifiques, il convient de définir le rôle de chacun pour éviter le chevauchement des services. Dans ce sens, l'entrepôt de données ANPERSANA se distingue par sa spécialisation disciplinaire dans les études basques. Cette plateforme est confrontée à des choix (types de services, outils intégrés ou procédures de dépôt) qui peuvent inciter l'engagement de la communauté scientifique visée, indispensable à l'évolution du projet. Mais cette implication est aussi influencée plus largement par la conception que les membres de la communauté scientifique ont sur la diffusion de leur travail en accès libre. Le contexte juridique, le temps, la culture professionnelle et les questions de confidentialité des sources influencent en effet les perceptions et pratiques de partage et de diffusion.

Mots clés : Données de la recherche ; Libre accès ; Entrepôts de données ; Plateformes de diffusion ; Valorisation de la recherche ; Bibliothèques spécialisées

Table des matières

INTRODUCTION

Au cours du Master 2 Information et Documentation de l'Université Paul Valéry – Montpellier III, j'ai progressivement développé un intérêt particulier pour les bibliothèques universitaires et les centres de documentation spécialisés. J'avais donc une idée du type de structure dans laquelle je souhaitais réaliser le stage du master.

De retour au Pays Basque, je me suis installé à Bayonne et j'ai tout de suite identifié IKER comme étant l'organisation qui regroupait toutes les caractéristiques de mon lieu de stage idéal. En tant que laboratoire de recherche de la langue et des textes basques, un stage chez IKER était l'occasion de concilier mes centres d'intérêt : la documentation dans l'Enseignement Supérieur et la Recherche (ESR) et l'étude des langues. Je suis en effet titulaire en Philologie anglaise et les disciplines de la linguistique et de la littérature m'intéressent particulièrement.

Suite à une proposition de Monsieur Jean Philipe Talec, Ingénieur d'études du CNRS chez IKER et responsable du centre de documentation, et après concertation avec Madame Céline Paganelli, directrice du master, et directrice de mon mémoire, j'ai décidé de travailler sur les données de la recherche et la bibliothèque numérique d'IKER, ANPERSANA, « recevant et diffusant des sources primaires de la recherche dans le domaine des études basques ».

L'accueil sur mon lieu de stage a été formidable et j'ai pu faire des rencontres très enrichissantes, tant au niveau personnel que professionnel. Sur le plan pédagogique, mes attentes ont également été largement comblées. J'ai pu voir dans le concret des aspects théoriques traités pendant le cursus du parcours Gestion de l'information et médiation documentaire, du master. Surtout, j'ai approfondi le sujet des données de la recherche, que je connaissais peu, pendant le travail quotidien mais aussi lors d'une journée d'études à Pau et une formation avec l'UrfIST de Bordeaux. Les missions périphériques du stage m'ont permis de compléter cet apprentissage par l'acquisition de compétences en lien avec d'autres aspects de la documentation : catalogage et élaboration d'un guide de catalogage, la mise en place d'un système de publipostage pour l'échange de revues, le bulletinage des revues, océrisation de documents image, automatisation des bibliographies, etc.

La mission principale a été de mener une enquête de terrain auprès des chercheurs

du laboratoire en vue de l'enrichissement d'ANPERSANA avec des nouveaux contenus. En effet, pour enrichir ANPERSANA, nous devons d'abord comprendre les implications que cela a pour la communauté scientifique des SHS et des études basques. Quelles sont les incitations et possibles freins à la diffusion des données en ligne ? Comment les chercheurs d'IKER perçoivent le libre accès et l'ouverture des données de la recherche ? Il est également essentiel de définir dans quelle mesure le projet de la bibliothèque numérique répond à ses objectifs et à ceux de l'ouverture des données de la recherche, à savoir rendre visible et valoriser les travaux de recherche de la communauté scientifique du laboratoire et encourager le partage et la diffusion de données au sein de celui-ci. Enfin, on peut s'interroger si les choix techniques et services offerts, contribuent à rendre les données repérables, accessibles, interopérables et réutilisables.

Pour répondre à ce questionnement, nous définirons dans une première partie les données de la recherche ainsi que leur contexte d'existence. Un état de la littérature nous permettra d'aborder les différents aspects de l'ouverture de données (techniques, institutionnels et juridiques) ainsi que d'en présenter les controverses et opportunités pour les SHS et les études basques. La seconde partie sera l'occasion de nous pencher plus spécifiquement sur notre sujet : suite à une brève description d'IKER et ses projets, nous analyserons ANPERSANA en tant qu'entrepôt de données qui fait partie d'un écosystème qui s'agrandit en continu. Nous comparerons ses choix techniques et ses services à ceux d'autres projets similaires. Enfin, dans une troisième et dernière partie nous présenterons les résultats de l'enquête qui cherche à comprendre les pratiques et prédispositions de la communauté scientifique d'IKER au sujet de l'ouverture des données.

PREMIERE PARTIE : L'ouverture des données de la recherche dans les SHS

Après avoir introduit le sujet de ce mémoire, nous en définirons les concepts centraux. Nous aborderons dans un second temps les enjeux du libre accès et des données de la recherche du point de vue de la politique publique, et notamment en questionnant la législation qui l'encadre. Nous approfondirons ensuite la mise en œuvre concrète de l'ouverture des données de la recherche : les infrastructures et les choix techniques sur lesquels elle repose. Enfin, nous nous pencherons sur les spécificités liées au champ de recherche, à savoir les SHS et parmi elles plus spécifiquement la linguistique basque, pour interroger leur impact éventuel sur l'ouverture des données.

1. RAISON D'ETRE DE L'OUVERTURE DES DONNEES DE LA RECHERCHE

Nous allons tout d'abord présenter le contexte dans lequel s'inscrivent l'ouverture des données de la recherche et le libre accès, puis nous définirons les fondements de la culture de l'*open*.

1.1. Internet : contexte propice à l'émergence du mouvement du libre accès

C'est grâce à internet et aux technologies du numérique que la conception démocratique de l'accès à l'information a pu être imaginée. Le web des données constitue le point de rencontre de tous les internautes et l'augmentation de sa fréquentation au cours de cette dernière décennie est incontestable[1]. Communication, recherches d'information ou transactions financières ne sont que des exemples des nombreuses activités qui ont lieu sur le web. Les données, dont les métadonnées, sont des unités d'information qui rendent possibles ces opérations si les conditions (de technologies, architectures et savoir-faire) sont réunies. Le défi aujourd'hui s'inscrit donc autour de notre capacité à nous repérer dans un web de données massif et constamment croissant : trouver l'information dont nous avons besoin et la réutiliser à

[1] Selon le site internetlivestats.com, le pourcentage d'internautes dans le monde est passé de 17,6% en 2006 à 46,1% en 2016. Pour la France, le nombre d'utilisateurs supplémentaires est de presque 40% pour la même période ; 86,4% de la population française était utilisatrice en 2016.

notre convenance. Il s'agit, non seulement d'ouvrir les données en libre accès, mais également de les rendre repérables et réutilisables, et de garantir leur accessibilité permanente et la véracité des sources.

1.2. Les valeurs de la culture de l'*open access*

La culture de l'*open access* réclame le partage et la transparence de l'information scientifique, ainsi que la possibilité d'y accéder gratuitement (Déclaration de Budapest, 2002 ; Déclaration de Berlin, 2003). C'est par la collaboration entre individus et entités que l'on envisage la production d'informations dérivées susceptibles de générer de la connaissance pour le développement sociétal. Issues à l'origine d'initiatives citoyennes, ces valeurs sont progressivement adoptées par les institutions publiques et les organisations privées.

Fondé sur des valeurs libertaires telles que celles que l'on vient de citer, ce mouvement évolue au cours des années 2000, notamment avec les crises financières et économiques conjoncturelles de la fin des années 2000, vers une conception et des objectifs plus libéraux (CHARTRON, 2016)[2]. C'est ainsi que le concept de *knowledge-based economy* (économie de la connaissance), qui émerge dans les années 90 (OECD, 1996), nourrit les débats sur la productivité, la performance et la croissance économique: le partage de l'information et la collaboration scientifique sont ici des conditions pour l'efficience de la science et de la connaissance qui doivent alimenter l'innovation technologique et industrielle au profit de l'économie.

Certains acteurs de l'*open access* (OCDE, Union Européenne) n'opposent pas ces deux approches (libertaire et libérale), mais défendent au contraire la vision qu'elles se complètent. Dans la préface des « Principles and Guidelines for Access to Research Data from Public Funding », l'OECD (2007) énumère ainsi dans un premier temps les avantages sociétaux puis poursuit avec l'affirmation « *access to research data increases the returns from public investment in this area* », qui fait référence à une volonté de rentabiliser le financement public.

[2] Ghislaine Chartron cite l'analyse de Fidelia Ibekwe-SanJuan et Françoise Paquienséguy (2015), ainsi que Le Memorandum de l'OSTP (2013), la recommandation européenne de 2012 et le rapport FINCH de 2012 du gouvernement britannique.

Nous pouvons ainsi conclure que la culture du libre accès se construit et évolue sur des valeurs distinctes : d'une part (i) les fondements d'origine idéalistes qui visent une société plus démocratique et des institutions plus transparentes, et de l'autre (ii) une vision utilitaire qui conçoit la connaissance comme une ressource économique additionnelle, un capital immatériel. Le profit économique et la croissance pouvant contribuer à l'amélioration des conditions de vie, ces deux visions ne sont pas forcément en concurrence.

2. *OPEN RESEARCH DATA* ET *OPEN SCIENCE* : LES NOTIONS DE DONNEES DE LA RECHERCHE ET DE RESULTAT DE RECHERCHE

Après avoir défini les origines et valeurs de la science ouverte et de l'ouverture des données de la recherche, nous approfondirons plus spécifiquement sur les frontières floues qui existent entre les différents types de données de la recherche.

2.1. Science ouverte et ouverture des données de la recherche

Pendant la réalisation de ce mémoire, nous avons eu l'occasion de discuter avec différents acteurs de l'ouverture des données de la recherche : doctorants, chercheurs confirmés et spécialistes de l'Information scientifique et technique (IST). Lors de ces échanges nous constatons que les concepts d'*Open research data* et de l'*Open science* sont souvent interchangeables, comme s'ils étaient synonymes. Il est en effet vrai que les deux notions se réfèrent au libre accès à l'information scientifique, ou *open access*. OpenAire[3] le définit ainsi:

> *The immediate, online, free availability of research outputs without restrictions on use commonly imposed by publisher copyright agreements. Open Access includes the outputs that scholars normally give away for free for publication; it includes peer-reviewed journal articles, conference papers and datasets of various kinds.*

Pour faire la distinction entre ce qu'englobe chacune des notions nous pouvons évoquer que l'*open science* fait référence aux résultats (publications scientifiques de

[3] Initiative européenne à grande échelle qui soutient la recherche ouverte, [En ligne : https://www.openaire.eu/oa-overview]. Consulté le 14 avril 2017.

toutes sortes : articles, thèses, mémoires, etc.), tandis que l'*open research data* se réfère aux données produites au cours du travail de recherche préalable aux résultats. Parmi les définitions des « données de la recherche » que l'on trouve dans la littérature, celle de l'OCDE (2007), inspirée d'après Gaillard (2014) et Delay-Artous (2014) de celle donnée par le Bureau de la gestion et du budget du gouvernement fédéral américain en 1993, semble être l'une des plus couramment citées[4] :

> *Les données de la recherche sont 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 les résultats de la recherche.*

Les auteurs Gaillard (2014) et Cabrera (2015) dégagent des éléments communs aux multiples définitions des données de la recherche. Le premier résume que ce sont « des données numériques, produites au cours d'un processus de recherche et pouvant servir de support à une démonstration scientifique ». Cabrera de son côté établit que « les données de la recherche seraient : les sources principales ou constitutives de la recherche ; des informations ; des enregistrements factuels ». Nous pouvons ainsi établir que les données de la recherche sont des enregistrements factuels, susceptibles de générer de l'information et inhérents au processus scientifique, en tant que supports de démonstration des résultats du travail de recherche. A partir de cette définition, nous revenons à l'idée de proximité entre les notions de science ouverte et d'ouverture des données de la recherche. Le lien entre les données et les résultats est au cœur même de la vérité scientifique, car le libre accès aux unes et aux autres contribue à la transparence des procédés et rend les expériences reproductibles. D'ailleurs, transparence et reproductibilité sont probablement les arguments majeurs en faveur de la diffusion ouverte et conjointe des données et des résultats. Certains projets de diffusion s'en emparent et concrétisent cette indissolubilité en offrant la possibilité de diffuser simultanément production scientifique et données qui l'accompagnent. OpenAire, via la plateforme de son partenaire Zenodo[5], s'inscrit dans cette démarche avec la déclaration :

[4] BSN10, 2016 ; CABRERA, 2015 ; CIRAD, 2016 ; GAILLARD 2014 ; PAIN 2016

[5] Plateforme web de curation, partage et diffusion des données scientifiques, [En ligne : https://zenodo.org/]. Consulté le 14 avril 2017.

« *Truly Open Science requires information about the whole research lifecycle to be open.* »

Nous avons tenté de définir ce que recouvre le concept de « données de la recherche » et une partie de notre définition établit que ce sont « des enregistrements factuels, susceptibles de générer de l'information » (cf. Première partie, 2.1.). Ainsi une donnée, de façon isolée, ne constitue pas une source d'information. Pour qu'elle devienne interprétable, elle doit être mise en relation avec la réalité dont elle est issue ou avec d'autres données. Quand des données sont présentées dans un ensemble intellectuellement cohérent, on dit que c'est un jeu de données (*dataset*). Le contenu d'une base de données, par exemple, est un jeu de données, tout comme une collection thématique de cartes postales.

2.2. Les différents types de données de la recherche : une frontière diffuse

La communauté scientifique utilise toutes sortes de données de la recherche qui peuvent être plus ou moins structurées ou plus ou moins traitées. Selon les niveaux de traitement, les données de la recherche font partie de différentes catégories. Rémi Gaillard (2014) cite l'Australian National Data Service pour synthétiser cette catégorisation :

> *Les données [de la recherche] peuvent être des données brutes, des données non-traitées d'observations de phénomènes particulières. D'autres sont des données traitées, données produites après formatage ou correction des données brutes. D'autres des données dérivées, qui présentent un résumé ou une présentation spécifique des données brutes.*

Un linguiste qui travaille sur un dialecte, dans son travail de terrain, enregistre des locuteurs natifs. Chaque enregistrement est une **donnée brute**, ou **donnée primaire**, et l'ensemble des enregistrements est un **jeu de données** brutes, car les données n'ont pas encore été traitées. Ensuite, le linguiste effectue un tri. Il enlève deux enregistrements qui sont de mauvaise qualité et édite le reste des enregistrements, coupant des parties par exemple. Les enregistrements sont maintenant des **données traitées**, ou **des données secondaires**, aussi que les transcriptions qui les accompagnent. Finalement, le chercheur élabore un document Excel où sont recueillis les aspects du langage en

relation avec son sujet de recherche. Ce document contient des **données dérivées**. Pourtant, la frontière entre donnée brute, donnée traitée et donnée dérivée n'est pas si évidente. On pourrait dire que le document Excel ne comprend que des données traitées, car elles n'ont toujours pas été intégrées dans un corpus plus large ; ou bien que les transcriptions des enregistrements sont des données dérivées, car elles constituent une présentation particulière (textuelle) de la même information, et différente à la présentation d'origine (sonore).

Dans le chapitre « Data Flakes: An Afterword to "Raw Data" Is an Oxymoron », Geoffrey C. Bowker (2013) joue avec le sens d'origine du mot anglais *raw*, « cru », pour démontrer que les données ne sont jamais brutes ou « crues ». Toutes les données de la recherche, même les « données brutes », auraient subies un certain degré de « cuisson ». La métaphore culinaire éclaire que les données de la recherche, de par leur nature numérique, ont forcément subi une transformation qui les met en relation avec un objet de la réalité. Si nous revenons à l'exemple du linguiste, les enregistrements sonores constituent un traitement des faits d'origine, par le biais de l'appareil d'enregistrement utilisé, et entreraient donc dans la catégorie des données traitées.

3. L'EMERGENCE DE L'ANALYSE MASSIVE DES DONNEES : LE BIG DATA

Le *big data* développe les techniques, infrastructures et technologies permettant un accès rapide aux bases de données diverses et de grand volume, comme l'indique le terme *big*. Quand la quantité de données ou de jeux de données est si vaste ou complexe que les techniques traditionnelles de traitement de données deviennent inadéquates, ce sont les algorithmes du *big data* qui entrent en jeu (CABRERA, 2015). Au-delà de l'importance du volume, ces algorithmes sont censés permettre la prédiction de tendances, phénomènes et risques de toutes sortes. Certains domaines, comme la publicité et le marketing, illustrent ce que le *big data* peut entraîner au niveau commercial et en termes de profits économiques. Les projets de sécurité à grande échelle sont aussi particulièrement modulés par ce que l'analyse massive de métadonnées leur apporte. Dans la recherche, ce sont surtout les domaines des sciences techniques et médicales (STM) qui ont suscité l'intérêt du financement public et privé, à cause de leur nature et de leur potentiel d'impact direct dans l'économie et le développement sociétal. Les STM, notamment la médecine et la biologie, sont en effet

des domaines susceptibles d'offrir des apports substantiels à l'avancement de la société au-delà de l'économie, ce qui a conduit les administrations à se pencher sur ces opportunités. Il convient de noter que le secteur privé comme le public fait preuve d'un intérêt manifeste pour le *big data*, intérêt qui se traduit par de grands investissements financiers.

En ce qui concerne la France, Cabrera (2015) évoque la création de la chaire « *Data scientist* » qui cherche à promouvoir ce profil de professionnels, avec la collaboration du Ministère de l'Enseignement supérieur et de la recherche (ESR), de l'école Polytechnique et des organisations privées. Dans la feuille de route du *big data* « La nouvelle France industrielle » (BOURDONCLE et HERMELIN, 2014), il est estimé que les dépenses en logiciels, services et autres dépenses internes des entreprises et des institutions pour le développement du *big data* atteindront la somme de près de 9 milliards d'ici 2020. Au niveau international, les investissements financiers les plus importants viennent des États-Unis et du monde anglo-saxon (GAILLARD, 2014). Pourtant, en Europe aussi, seulement dans le cadre du programme d'innovation et recherche Horizon 2020[6] (H2020), 500 millions d'euros ont été réservés dans un fond pour soutenir la recherche sur le *big data*. L'objectif est d'augmenter ce fond jusqu'à la somme de 2 milliards grâce aux investissements supplémentaires des entreprises privées (Parlement Européen, 2016). L'Europe établit les principaux « défis sociétales » auxquels H2020 se confronte et pour lesquels les *big data* peuvent apporter des solutions : climat, énergie, alimentation, santé, transport, sécurité et les sciences humaines et sociales[7] (SHS).

4. CADRAGE INSTITUTIONNEL ET MODELES ECONOMIQUES

4.1. Juridiquement, les données de la recherche sont-elles des *open data ?*

En droit, l'*open data* ou l'ouverture des données est la notion par laquelle on désigne les démarches d'ouverture de l'information publique, qu'il s'agisse de documents, de données ou de métadonnées. Il s'agit de mettre les données produites par

[6][En ligne : http://www.horizon2020.gouv.fr/]. Consulté le 30/03/2017.

[7] Traduit de la dénomination anglaise « social sciences », qui comporte également les domaines des sciences humaines.

les administrations à la disposition du public qui souhaite les consulter, en libre accès et de façon à ce qu'elles soient réutilisables « sans restriction technique, juridique ou financière injustifiée » (BSN10, 2016). Les textes qui les encadrent sont la Directive 2013/37/UE en Europe, dite PSI, *Public Sector Information*, et la Loi nº78-753 du 17 juillet 1978, dite CADA, codifié depuis le 19 mars 2016 dans le Code des relations entre le public et l'administration. D'un point de vue juridique, les données de la recherche appartiennent à la catégorie des documents administratifs, car elles sont produites par des organisations de recherche et d'enseignement qui dépendent de l'administration et s'inscrivent dans les missions du service public. Cela ne veut pourtant pas dire que le producteur des données de la recherche soit désormais obligé à diffuser toutes ses données. Une distinction est faite entre ce qui est produit par les universités *sur* la recherche, et les données *de* la recherche, pour lesquelles les dispositions sont à ce jour sont encore floues. Mais le fait que l'on ait utilisé la terminologie « donnée de la recherche » dans la Loi Lemaire (article 38), laisse entrevoir qu'une catégorie juridique spécifique sera développée pour ce type de données, pour lesquelles des dispositions légales particulières seront aussi rédigées (MAUREL, 2016).

4.2. Politiques d'État et intérêts éditoriaux en SHS : tensions

La question du dépôt des publications en accès ouvert, surtout quand il y a un contrat avec l'éditeur contraignant pour l'auteur, s'est posée à plusieurs reprises au cours d'une journée d'études dans l'atelier animé par Hélène Skrzypniak (2017). Le débat s'est centré sur le texte législatif qui prévaut dans ces cas-là : le contrat avec l'éditeur ou la loi. La réponse évidente semble être que les clauses d'un contrat ne respectant pas la loi sont invalidées par celle-ci. Cela devient plus compliqué quand il s'agit d'un contrat signé avec un éditeur étranger, dont le pays dispose d'un cadre légal différent : comme la loi d'un pays ne prévaut sur celle d'un autre, ce serait au juge de déterminer. Selon Ghislaine Chartron (2016), « les tensions » sont « croissantes avec les acteurs de l'édition nationale, majoritairement en sciences humaines et sociales » domaine éditorial français très puissant dans toute la francophonie. « L'État est devenu un opérateur majeur et la distorsion introduite par rapport aux autres acteurs ayant investi des moyens pour développer des activités de service dans ce secteur, conduit à des tensions de plus en plus fortes, [dont] le manque de régulation en est la cause majeure ».

La Loi Lemaire, est une tentative pour trouver un équilibre entre les intérêts opposés de l'industrie éditoriale d'une part, et des institutions de l'ESR et auteurs de l'autre. En effet, ces derniers se voyaient souvent dans l'obligation de payer pour avoir accès aux publications, sachant qu'ils avaient déjà payé pour pouvoir les publier ! Cela en est de même pour la publication des données de la recherche, pour lesquelles les éditeurs « ne pourront [plus] se contenter d'utiliser leur position dominante pour en revendre l'accès » (MAUREL, 2016). Les éditeurs qui exigent dans leurs procédures de dépôt d'articles de déposer également les données qui ont été utilisées pendant le travail de recherche, n'en ont plus un droit d'utilisation exclusive. Une fois la publication scientifique rendue publique, après la période d'embargo, les données qui l'accompagnent seront également mises à disposition du public.

En dehors du cadre de l'état, les différentes organisations de financement de la recherche mettent en œuvre leurs propres politiques. H2020 prévoit que l'accès libre aux publications et aux données issues des projets encadrés par le programme soit généralisé. Il n'y a pas d'obligation de publier, mais les bénéficiaires du programme qui choisissent de le faire sont obligés à mettre leurs publications en ligne en accès gratuit, par le moyen de la **voie dorée** (les éditeurs rendent les publications librement accessibles) ou de la **voie verte** (par le moyen de l'auto-archivage, l'auteur publie en accès ouvert à l'issue d'une période d'embargo, 6 mois pour les STM et 12 pour les SHS). En ce qui concerne les données de la recherche, seulement les projets intégrés dans le pilote *Open Research Data* ont l'obligation de les diffuser en accès ouvert.

5. DATA MUST BE FAIR: FINDABLE, ACCESSIBLE, INTEROPERABLE & REUSABLE

L'ouverture des données de la recherche est régie par des principes préconisés par le Data FAIRport Initiative en 2014 lors d'un atelier à Leiden. « Les données doivent être 'justes' » réclame la devise de ce mouvement se servant du terme *fair* en l'anglais, simultanément métaphore et acronyme représentant les qualités des données qui vont être diffusées : *Findable*, *Accessible*, *Interoperable*, et *Reusable*. Largement acceptés par les acteurs de l'ouverture des données de la recherche et les spécialistes du web, ces principes sont également repris par H2020, notamment dans le guide du programme (2016). Mais comment mettre en place un système qui garantisse le respect de ces

principes ? Autrement dit, quelles sont les conditions nécessaires pour que les données de la recherche soient systématiquement rendues repérables, accessibles, interopérables et réutilisables ?

Dans la littérature spécialisée, et notamment sur les sites web du Data FAIRport initiative et du programme H2020, les notions d'infrastructure, de technologie, de communauté d'acteurs, de communauté scientifique, de cycle de vie de l'information, de choix des formats, de choix des techniques et de standard sont constamment évoquées autour de la question du *FAIR data*. Nous avons voulu regrouper ces notions sous deux sections qui représentent les grandes catégories des conditions préalables à la diffusion des données de la recherche FAIR : (i) les infrastructures développées pour le soutien du libre accès dans la science, (ii) le choix des aspects techniques et les bonnes pratiques en gestion de données. Nous introduirons ensuite ces conditions et leurs liens avec les principes FAIR. Ce sera l'occasion d'établir ainsi les éléments au regard desquels nous mènerons, dans la Deuxième partie de ce mémoire, une analyse de la bibliothèque numérique ANPERSANA.

5.1. Infrastructures pour le libre accès dans la science

Au début de ce mémoire, nous avons expliqué que le mouvement du libre accès a surgi avec l'émergence du web des données. Pourtant, internet est tout un ensemble d'infrastructures technologiques, dont l'implantation dépend des investissements financiers des acteurs publics et privés, ainsi que de la capacité économique des particuliers et des organisations qui veulent y avoir accès. Ces infrastructures sont donc composées d'une part des réseaux de communication et serveurs de grande capacité, et de l'autre des postes informatiques ou machines qui se connectent avec les réseaux. Selon Gaillard (2014), le Centre national de la recherche scientifique (CNRS) avertit au début de siècle de la « fragilité des dispositifs supports » en SHS. L'évolution depuis les années 2000 est remarquable. La France « a misé ... sur des infrastructures publiques » (CHARTRON, 2016). La création de la BSN en 2009 marque le tournant de cette évolution, puisque c'est la première initiative qui cherche à regrouper les acteurs de l'IST en organisant 10 groupes de travail ou « segments » autour de sujets stratégiques pour la recherche. Un autre moment clé dans cette évolution est l'aboutissement en 2013 de la Très Grande Infrastructure de Recherche (TGIR) Huma-Num qui coordonne la

majorité des plateformes de diffusion de contenus scientifiques. « Le programme ISTEX est la dernière initiative ; elle vise à constituer un fond national des collections rétrospectives de toutes disciplines, hébergé sur une plateforme publique à laquelle viendraient se greffer des services aux chercheurs. »

Le dossier « Politique nationale de l'IST : des infrastructures en cohérence », publié dans Ar(abes)ques (ABES, février-mars 2017), nous accompagne dans un parcours à travers le paysage français d'infrastructures existantes : Collex-Persée, HAL, Numédif, Isidore et OpenEdition. Le CINES, en tant qu'opérateur de l'archivage numérique pour les données issues de l'ESR, et les nombreuses plateformes institutionnelles d'auto-archivage de publications viennent s'ajouter à cet écosystème. En ce qui concerne les entrepôts de données de la recherche en SHS, une recherche sur le registre d'entrepôts de données re3data[8], avec les filtres « *Humanities and social sciences* » et « *France* », nous donne 16 résultats. Et cela n'inclut pas les entrepôts internationaux qui sont évidemment aussi utilisés en France et les entrepôts généralistes où le dépôt de données en SHS a aussi lieu. En effectuant la même opération avec l'application Open Science Monitor[9] de la Commission Européenne, on obtient 17 résultats. En nous servant de cette même application, nous constatons qu'il y en a 446 en Europe et que la France occupe la 3e place en nombre d'entrepôts de données spécialisés en SHS, derrière le Royaume-Uni et l'Allemagne.[10]

Nous verrons dans la Deuxième partie que chaque acteur apporte son expertise en fonction du rôle qui lui correspond dans un contexte d'ouverture des données quelconque. Concrètement nous essayerons de situer ANPERSANA dans l'écosystème des acteurs (organisations, opérateurs, initiatives, entrepôts des données, etc.) en France, en Europe et à l'international.

5.2. Choix techniques et bonnes pratiques en gestion de données

Il ne suffit pas que les infrastructures d'un pays et d'une organisation soient efficacement implantées. Si les données de la recherche doivent être repérables,

[8] [En ligne : http://www.re3data.org/browse/by-subject/]. Consulté le 27 avril 2017.

[9] [En ligne : http://ec.europa.eu/research/openscience/index.cfm?pg=home&section=monitor]. Consulté le 27 avril 2017.

[10] Recherches effectuées entre le 15 et le 27 avril 2017, dont les résultats sont susceptibles d'évoluer dans l'avenir.

accessibles, interopérables et réutilisables, nous devons également faire les bons choix techniques au niveau de la structure de rattachement du chercheur ainsi qu'au niveau du producteur des données. Cela veut dire choisir les schémas de description, les vocabulaires contrôlés et les formats des fichiers adaptés au domaine de recherche, aux types des données et au moment du projet scientifique. Ce n'est pas la même chose de gérer des enregistrements sonores produits lors des sessions de psychothérapie que de gérer des images collectées auprès d'un archive numérique pour une recherche en histoire de l'art ; ce n'est pas non plus comparable de stocker ses données pendant le projet en cours, que de les stocker en prévision de leur archivage pérenne. Chaque cas de figure nécessite des choix différents pour la gestion des données. Si l'on veut gérer ses données selon les principes FAIR, une gestion appropriée doit être appliquée toute au long de leur cycle de vie.

Le plan de gestion des données (PGD) ou *Data Management Plan* (DMP), désormais requis pour les appels à projets de certains programmes de financement de la recherche comme notamment H2020, est un outil qui permet de prendre conscience des multiples aspects de la gestion des données de la recherche dans les différentes étapes de leur cycle de vie. Cette prise de conscience contribue aux bonnes pratiques de gestion de données (choix de stockage, prise en compte de l'obsolescence des logiciels et formats, de leur interopérabilité, mesures de protection des données, choix des métadonnées, etc.). Le PGD est également un outil de formalisation de ces pratiques. La formalisation des actions menées ou à mener sur les données facilite, le cas échéant, leur mise en ligne sur une plateforme de diffusion ou d'archivage pérenne. Le chercheur aura préalablement pris en compte et matérialisé la majorité des prérequis de la plateforme choisie. Un PGD n'a pas une forme unique ; ses différents constituants (les sections et les champs à remplir) sont conçus selon les besoins du porteur de projet de recherche. Les outils en ligne pour l'aide à la création d'un PGD sont multiples : ceux de l'Université de Monash[11], le DCC[12][13] ou DMP OPIDoR[14] en sont quelques exemples.

11 Téléchargement direct d'une checklist. [En ligne : monash.edu/library/researchdata/file_links/datahdrchecklist_doc.doc]. Consulté le 15 mai 2017.

12 « Checklist for a DMP ». [En ligne : http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf]. Consulté le 15 mai 2017.

13 Logiciel « DPM Online ». [En ligne : https://dmponline.dcc.ac.uk/]. Consulté le 15 mai 2017.

Nous avons aussi joint (cf. Annexe 1, p. 78) le modèle de PGD que le laboratoire IKER a créé pour sa communauté scientifique et que nous avons traduit en basque et en anglais.

6. LES DONNEES DES SHS, DE LA LINGUISTIQUE ET DES ETUDES BASQUES : CONTROVERSES ET OPPORTUNITES

Toute au long de cette Première partie, nous avons tenté de dégager des éléments d'analyse de l'ouverture de la science et des données de la recherche. Dans un contexte d'engagement des acteurs de la recherche, nous avons précédemment constaté que certaines spécialités bénéficient du *big data* au profit des résultats scientifiques inédits et des profits économiques. Dans ce sens, la rentabilité de la production scientifique est de plus en plus présente. Les données, en tant que données numériques, permettent souvent l'analyse automatisée, mais provoque aussi un changement des méthodes scientifiques et bouleverse une partie de la communauté scientifique, plus attachée aux méthodes traditionnelles. D'ailleurs, l'automatisation des procédés, censée notamment contribuer à un gain de temps, est souvent perçue comme contraignante, puisqu'il est nécessaire de traiter spécifiquement et préalablement les données. Les procédures numériques sont aussi à la base de nouvelles évaluations par les pairs très controversées.

Pourtant, la démocratisation (gratuité et accès à distance) de la science, la possibilité de vérifier et de valider les expériences scientifiques, de réutiliser les données et de les réinterpréter, ainsi que le croisement des différents types de données (interdisciplinarité), sont des éléments positifs à mettre dans la balance. Qu'en est-il des SHS au regard des enjeux cités ? Concrètement, quels sont les avantages et les inconvénients de l'ouverture de la science pour les domaines de la linguistique et des études basques ? Dans cette section nous allons essayer de résumer les principales controverses et opportunités autour de ce sujet, notamment du point de vue de la communauté scientifique.

14 Version du logiciel « DPM Online » customisée pour la recherche française, développé par le DCC en partenariat avec le CNRS. [En ligne : https://dmp.opidor.fr/]. Consulté le 15 mai 2017.

6.1. Controverses en SHS, en linguistique et dans les études basques

6.1.1. Un éloignement de l'essence même de la recherche en SHS ?

Christine L. Borgman (2012) explique que la reproductivité des expériences scientifiques peut être problématique dans le sens où le processus scientifique risque d'être réduit à une procédure mécanique. S'imprégner des questions de recherche est une condition nécessaire inséparable de l'interprétation des données. Elle cite Lagoze et Velden (2009) pour évoquer que les chimistes considèrent leur discipline tant comme un art que comme une science. Cela est d'autant plus important pour la recherche en SHS qui, par sa nature moins empirique que la chimie et les autres STM, craint davantage la mécanisation de ses méthodes. Francisca Cabrera (2015) note cet aspect chez les chercheurs interviewés, spécialistes de domaines à approche herméneutique comme la philosophie et la littérature. « Une donnée de la recherche ne va pas de soi car, dans l'essentiel de leur activité, les étapes de réflexion et construction d'un objet scientifique sont inséparables de la construction du texte et visent une forme aboutie du discours à travers la publication. »

6.1.2. Résistances à la diffusion

Cabrera (2015) constate que la confidentialité des sources est la préoccupation la plus couramment exprimée par les interviewés, chercheurs en SHS, quand ils sont questionnés sur les principaux obstacles à la diffusion de leurs données. Les questions éthiques et l'anonymisation des sources sont à l'ordre du jour dans les disciplines comme la linguistique, la plupart des données étant recueillies sur le terrain. Dans le cas de la documentation des langues minoritaires, dont le basque, la question est encore plus présente. Les personnes-sources sont souvent issues des communautés relativement réduites, ce qui les rend plus facilement reconnaissables par leur voix ou le contenu de leur récit (même dans les cas où des informations telles que les noms propres ont été retirées). Le deuxième frein au partage identifié par Cabrera est le manque de temps. Borgman (2012) avait également repéré que « the more handcrafted the data collection and the more labour-intensive the postprocessing for interpretation, the less likely that researchers will share their data. » Nous pouvons interpréter cette citation de deux

façons : soit (i) les données collectées, par leur nature ou leur volume, rendent la phase de traitement préalable à la diffusion trop longue et donc se lancer dans un tel projet semble une perte de temps (ce qui confirme les résultats de Cabrera) ; soit (ii) le chercheur qui investit beaucoup de temps et d'effort dans la collecte et le traitement considère que ce n'est pas juste que d'autres bénéficient de son travail.

En relation avec la deuxième hypothèse, nous dégageons deux profils de chercheurs. D'une part, ceux qui trouvent que leur travail de recherche leur appartient, même s'il est financé par une institution publique (ce qui est majoritairement le cas dans les SHS). D'autre part, ceux qui ont peur de perdre l'avantage scientifique. Il y a aussi certains cas de *scooping*, qui est une forme plus abrupte de perte d'avantage scientifique. En effet, dans ces cas, bien qu'exceptionnels, un chercheur profite de la diffusion précoce des données d'un autre et publie un article avant le producteur des données. Cependant, la pré-diffusion des données peut potentiellement protéger du *scooping*, à partir du moment où le producteur des données devient public et est reconnu par la communauté. (ROSS-HELLAUER, DEPPE et SCHMID, 2017)

6.1.3. Les Peer-review parfois injustes

Les modalités d'évaluation par les pairs sont l'un des aspects les plus controversés de la science ouverte. Certains défendent les traits traditionnels du *peer-review* où (i) l'auteur et/ou le pair sont anonymes ; (ii) l'évaluateur est choisi par l'éditeur ; (iii) l'évaluation et/ou la procédure d'évaluation n'est pas publique. Souvent critiqué, ce modèle encouragerait les évaluations injustes ou arbitraires, à cause de son opacité. L'*Open Peer Review* (OPR) s'approprie les valeurs de la science ouverte pour rendre les évaluations plus transparentes : (i) les identités des auteurs et des pairs sont publiques ; (ii) les évaluations sont publiées ; (iii) le grand public peut participer au processus d'évaluation ; (iv) les discussions sont ouvertes et encouragées ; (v) l'espace d'évaluation est géré par une plateforme de diffusion, indépendante de l'éditeur. (ROSS-HELLAUER et al., 2017) Dans la réalité, la division entre les modalités traditionnelles et celles de l'OPR est diffuse. Chaque éditeur ou plateforme constitue ses propres combinaisons de modalités d'évaluation, essayant de proposer la solution la plus juste. Pour Christine L. Borgman (2012) le problème de l'évaluation par les pairs est surtout lié à l'objet évalué. L'évaluation se base en effet sur l'avis d'experts qui se

servent de l'information donnée par l'auteur, mais rarement sur la reproduction ou la vérification des données car elles ne sont pas toujours accessibles directement. L'ouverture des données de la recherche semble dans ce sens être une solution, au moins partielle.

D'ailleurs, dans leur enquête de 2017, Ross-Hellauer et al. ont trouvé que la communauté scientifique est plus réceptive à l'*Open Access* et à l'*Open Research Data* que à l'OPR. 88% des participants sont pour le libre accès aux publications scientifiques et ils sont 80% concernant les données de la recherche. Ceux qui souhaitent que l'OPR soit la pratique courante ne représentent que 60%. Si on se penche sur les résultats au sujet de l'OPR par discipline, on constate qu'en moyenne les participants issus des SHS (incluant l'économie) sont plus pour ces pratiques (environ 2/3) que ceux des STM (environ 1/3). La seule exception est la discipline qui nous concerne dans ce mémoire, langues et littérature où 1/3 sont pour l'OPR et moins d'un quart des répondants sont contre les modalités traditionnelles de l'évaluation par les pairs. L'aspect de l'OPR qui préoccupe les répondants issus des STM et des langues et littérature est le fait de rendre les identités publiques. La crainte de publier une évaluation négative qui se retournerait contre l'évaluateur à un moment dans sa carrière en serait la cause, surtout dans les disciplines les plus concurrentielles selon les réponses de certains participants.

6.2. Opportunités en linguistique et dans les études basques

6.2.1. Transparence, vérification et réinterprétation

Nous avons signalé que la transparence du processus scientifique et la reproduction des expériences (vérification et validation) sont les principaux arguments des défenseurs de l'ouverture des données de la recherche. D'ailleurs, si nous acceptons la prémisse que données et résultats sont un tout inséparable, nous pourrions nous demander s'il est possible de concevoir une science ouverte sans données ouvertes. En linguistique, la nature de la recherche est souvent descriptive dans la mesure où l'on essaye de définir les structures du langage naturel (CROWFORD, 2015). C'est ainsi également le cas pour les études basques, dont la communauté scientifique pratique davantage la linguistique de corpus, c'est-à-dire l'analyse du langage naturel à partir des corpus de

textes, souvent des transcriptions ou des productions orales collectées sur le terrain. La mise à disposition en accès ouvert de ces corpus permet la vérification des travaux de recherche, mais surtout rend envisageable la réinterprétation et les nouvelles découvertes scientifiques.

Les domaines de la dialectologie et de la documentation des langues minoritaires sont les principaux fournisseurs de ces données, dont le reste des champs de la linguistique bénéficie. Les théoriciens de la linguistique s'en servent pour établir des liens entre langues éloignées ; liens qui justifieraient ou non l'existence d'une grammaire commune (théorie de la Grammaire Universelle[15] de Noam Chomsky), ou qui expliqueraient comment le langage est acquis. La typologie linguistique utilise ces données pour établir des traits communs entre des langues génétiquement éloignées ; la linguistique historique a également besoin des données linguistiques sur lesquelles appliquer les techniques de reconstruction. Enfin, toute autre sous-discipline est également concernée puisqu'il s'agit de comprendre les phénomènes du langage naturel.

6.2.2. Quelques projets en cours

Dans l'étude des langues minoritaires, les données ont un, ou plutôt des, intérêts supplémentaires : la préservation d'un patrimoine immatériel universel et le cas échéant la revitalisation de la langue[16]. Le développement incontestable d'infrastructures, initiatives et entrepôts de données est une opportunité pour documenter ces langues dans de meilleures conditions, mais c'est en même temps un domaine où l'urgence ne cesse de s'accroître, puisque les locuteurs disparaissent progressivement. C'est également le cas pour certains dialectes du basque : le 13 avril dernier, un journal annonce ainsi que le dernier locuteur d'un dialecte d'Arakil, en Navarre, est mort[17]. Les projets en accès libre de documentation et description des langues minoritaires sont multiples dans le monde. L'un des plus remarquables, par son volume de données, est l'entrepôt ELAR, de SOAS (School of Oriental and African Studies, University of

[15] Article « Innateness and Language », Stanford Encyclopedia of Philosophy. [En ligne : https://plato.stanford.edu/entries/innateness-language/]. Consulté le 27 mai 2017.

[16] ELAR, SOAS. [En ligne : https://elar.soas.ac.uk/]. Consulté le 16 mai 2017.

[17] « Urritzolako azken euskalduna hil da », Berria. [En ligne : http://www.berria.eus/albisteak/133750/urritzolako_azken_euskaldun_zaharra_hil_da.htm]. Consulté le 16 mai 2017.

London). ARBRES, projet décrit davantage dans l'article de Mélanie Jouitteau (2012) créatrice de la plateforme, est une expérience *crowdsourcing* de la langue bretonne, qui se caractérise par une approche innovante du processus scientifique même :

> *Le site fournit une grande grammaire du breton de plus de 1700 articles ainsi qu'un centre de ressources pour la recherche en syntaxe formelle … site de recherche "à carnet ouvert". C'est une expérience de science ouverte et participative … Le site est de type wiki, l'accès est public et gratuit et vous pouvez en modifier les pages vous-mêmes. Ce site est participatif et surveillé : je reste responsable de la qualité scientifique.*

Des initiatives concernant la langue basque émergent aussi constamment : Euskalkiak.eus, Ahotsak.org ou dernièrement Mintzoak.eus en sont les plus connues pour la documentation des productions orales. Pourtant, la majorité contiennent des données qui ne sont pas issues de la recherche, même si potentiellement réutilisables dans le cadre d'un projet de recherche. Dans ce sens, la bibliothèque numérique ANPERSANA, nous le verrons dans la Deuxième partie, vient répondre à un besoin.

6.2.3. Le traitement des données en linguistique

Un projet de plateforme de diffusion de données de la recherche a des objectifs différents de ceux d'un projet de bibliothèque numérique pour le grand public. D'une part, les données déposées doivent être issues exclusivement de la recherche. D'autre part, ces données seront formatées et enrichies de façon à ce qu'elles soient traitables et analysables pour de potentielles nouvelles recherches.

Le *data mining* et *text mining* sont des notions à ne pas ignorer dans un projet concernant des données de la recherche, surtout en linguistique. Tandis que le premier recouvre la découverte des données parmi des regroupements de données structurées, comme les bases de données, le deuxième s'occupe de l'extraction d'informations à partir d'un texte quelconque. (CABRERA, 2015) Les opérations de fouille de données se concrétisent au moyen de logiciels de plus en plus sophistiqués qui produisent des classifications automatiques : lemmatiques, syntaxiques, d'entités, statistiques, etc. Pour avoir une idée du nombre de technologies existantes, il suffit de regarder la liste d'outils

de text mining publiée sur le site du NACTEM[18].

Pourtant, comme il est signalé sur le site web de l'un des nombreux projets, Calliope (sans date), « les résultats des analyses automatisées dépendent … de la qualité des pré-traitements linguistiques ». Nous avons précédemment évoqué l'importance de décrire les données et de les enrichir avec des métadonnées qui les rendent FAIR (repérables, accessibles, interopérables et réutilisables), mais une valeur ajoutée supplémentaire d'un entrepôt de données de la recherche est de décrire en ajoutant des informations (des annotations) qui facilitent la fouille automatique de données.

Nous allons maintenant aborder plus en profondeur, dans la Deuxième partie du mémoire, le rôle d'un projet de bibliothèque numérique ou d'entrepôt de données de la recherche. Pour ce faire, nous essayerons d'abord de situer ANPERSANA dans l'écosystème des projets de données de la recherche. Ensuite nous analyserons ses fonctionnalités, choix techniques et procédures au regard d'une sélection d'entrepôts de données.

[18] The National Center for Text Mining. [En ligne : http://www.nactem.ac.uk/software.php]. Consulté le 16 mai 2017.

DEUXIEME PARTIE : La bibliothèque numérique ANPERSANA

« ANPERSANA est une bibliothèque numérique recevant et diffusant des sources primaires de la recherche dans le domaine des études basques (manuscrits, carnets de recherche, images, sons, et autres documents multimédia) », telle que définie sur son site web. Le projet conçu et développé par IKER UMR 5478 (CNRS-UBM-UPPA), existe depuis 2013 sous forme de plateforme web « en anticipant l'obligation de dépôt des données de recherche en archive ouverte pour les projets financés sur fonds publics. » Cette obligation étant aujourd'hui une réalité, le dépôt des contenus sur la bibliothèque numérique est promu auprès des chercheurs, enseignants chercheurs, post-doctorants et doctorants qui composent le public du laboratoire de recherche. Dans un contexte d'émergence de projets pour le soutien de l'ouverture des données de la recherche, il convient cependant de prendre du recul pour comprendre cet écosystème et la place qu'ANPERSANA y occupe.

Dans un premier temps, sont décrits l'origine du projet (contexte de création et objectifs), ainsi que les fonctionnalités, les choix techniques et les procédures de dépôt de la plateforme. Deuxièmement, nous essayons de situer ANPERSANA dans le panorama des projets de données de la recherche en SHS pour ouvrir la discussion sur son rôle et ses objectifs. Une analyse comparative est également menée pour mettre en perspective les fonctionnalités et choix techniques d'ANPERSANA avec une sélection d'entrepôts de données. Nous cherchons ainsi l'identification des points forts de ces entrepôts qui permettront la formulation, dans un troisième point, de recommandations pour ANPERSANA.

1. DESCRIPTION D'ANPERSANA

Après une brève présentation du lieu et des activités réalisées au cours du stage, nous détaillerons les missions de la plateforme ANPERSANA. Ensuite, nous reviendrons sur ses caractéristiques techniques. Enfin, nous présenterons comment s'opèrent le dépôt et la diffusion des données sur cette plateforme web.

1.1. IKER : Contexte de création d'ANPERSANA et lieu de stage

1.1.1. Présentation d'IKER et de son centre de documentation

Le laboratoire de recherche IKER UMR 5478 est créé en 1999 à l'initiative du CNRS. Depuis, la structure fonctionne sous la tutelle du CNRS, de l'Université Bordeaux Montaigne – Bordeaux III (UBM) et de l'Université de Pau et des Pays de l'Adour (UPPA). Les locaux du centre de recherche ainsi que de son centre de documentation se situent à Bayonne, au Chateau-Neuf, dans la même enceinte que la Faculté de Bayonne de l'UPPA et côte-à-côte avec Euskaltzaindia (l'Académie de la langue basque). IKER se spécialise dans le domaine de la langue et les textes basques, c'est ainsi que plusieurs dispositifs de valorisation de la discipline ont été développés ces dernières années. Parmi les plus importants, nous pouvons en signaler deux : (i) la revue scientifique annuelle *Lapurdum* [19], publiée avec le concours de la faculté pluridisciplinaire de Bayonne de l'UPPA par le moyen de Revues.org, un service d'OpenEdition [20] ; (ii) Artxiker [21], la plateforme d'auto-archivage de productions scientifiques sur le basque et les langues typologiquement proches, intégrée dans le réservoir HAL (Hyper Articles en Ligne) du Centre pour la Communication Scientifique Directe – UMS 3668 (INRIA-UdL).

En 2009, un centre de documentation est créé dans les locaux du laboratoire dans une salle de 100 m^2 avec un ameublement adapté (tables, chaises, canapés, fauteuils, tables basses, étagères et un escabeau sécurisé). La salle de lecture est également équipée avec la technologie nécessaire : des postes informatiques avec accès à internet, un service WiFi, un appareil de reproduction multifonction et un vidéoprojecteur. (TALEC, 2012) En ce qui concerne la collection spécialisée, elle se compose d'environ 6000 documents imprimés (évoluant d'une centaine d'ouvrages par an), ainsi que de documents numériques en accès libre ou sécurisé. La base bibliographique de signalement d'entrées est le catalogue KUTXA, développé avec l'application Koha.

[19] Dir. Charles Videgain. Tous les numéros sont disponibles en ligne : [https://lapurdum.revues.org/2367]. Consulté le 8 mai 2017.

[20] « Portail de ressources électroniques en sciences humaines et sociales », [En ligne : https://www.openedition.org/]. Consulté le 27 mai 2017.

[21] Resp. Jean Baptiste Coyos, [En ligne : https://artxiker.ccsd.cnrs.fr/]. Consulté le 27 mai 2017.

1.1.2. Activités pendant le stage au centre de documentation

En plus des services traditionnels de documentation, les projets et activités mis en place au cours de ces dernières années ont contribué aux activités de recherche du laboratoire. Pendant le stage de trois mois, nous avons participé à certains de ces projets réalisant, en parallèle de la mission principale, les activités périphériques suivantes :

- Veille documentaire dans le domaine des études basques : enrichissement des deux sites Netvibes d'IKER, Ikasketak[22] et Ikerlab[23], avec de nouveaux fils RSS et ressources web, réorganisé les sections et modifié la mise en page.
- Gestion des partenariats pour l'échange de revues : chaque numéro de la revue *Lapurdum* est envoyé aux participants de ce partenariat d'échange des revues scientifiques spécialisées dans différents sous-domaines des études basques. Un système de bulletinage avec Excel et de publipostage avec Word et Acces a également été créé.
- Participation à la formation documentaire : assistance de l'ingénieur d'études pendant un TD pour les étudiants de master sur l'automatisation de bibliographies avec JabRef [24] . Plus concrètement, les étudiants ont été accompagnés dans l'exécution des filtres d'export pour obtenir le format bibliographique de sortie désiré.
- Calcul volumétrique de la collection physique de la revue *Gure Herria*, prévoyant sa numérisation pour une éventuelle mise en ligne.
- Catalogage : saisie de documents divers (livres imprimés, ebooks, thèses, mémoires, revues) dans le catalogue KUTXA (Koha) utilisant la description Unimarc. Cela inclut l'océrisation des tables des matières et résumés pour le renseignement des champs 327 (note de contenu) et 330 (résumé ou extrait). En parallèle, un guide de catalogage définissant les conventions de saisie des notices et exemplaires dans KUTXA a été écrit.

Depuis 2013, le projet de bibliothèque numérique ANPERSANA vient s'ajouter à cet ensemble de services et dispositifs. En effet, la mission principale du stage a été de

[22] Veille de ressources pour le Master Etudes basques. [En ligne : http://www.netvibes.com/ikasketak#A_la_une]. Consulté le 26 mai 2017.

[23] Veille de ressources pour la recherche en études basques. [En ligne : http://www.netvibes.com/ikerlab#Aurkezpena]. Consulté le 26 mai 2017.

[24] [En ligne : http://www.jabref.org/]. Consulté le 26 mai 2017.

contribuer au projet ANPERSANA, notamment en identifiant des pistes d'amélioration de la plateforme. Nous décrirons ensuite les composantes de ce projet.

1.2. Objectifs d'ANPERSANA et brève description des collections

Nous remarquons dans la définition précédemment citée d'ANPERSANA que l'un de ses objectifs est la diffusion des « données primaires ». Nous avons également expliqué (cf. Première partie, 2.2.) que la frontière entre les différents types des données (brutes/primaires, traitées/secondaires et dérivées) est diffuse. Dans ce sens, il convient de préciser que les types des données diffusées sur la plateforme comprennent les données de la recherche entendues dans un sens large, au-delà de la terminologie utilisée dans la présentation : l'objectif est de rendre accessibles toutes les données de la recherche issues des études basques, quel que soit leur niveau de traitement. Il n'est d'ailleurs pas évident de classifier les contenus publiés sous un seul type de donnée (primaire, secondaire ou dérivé). Cela est notamment dû au fait que les contenus en ligne sont accessibles sous différentes formes, plus ou moins traitées, publiées simultanément. Ainsi, nous verrons qu'alors que la première des deux collections disponibles sur la plateforme rassemble des contenus plutôt bruts ou primaires, la deuxième témoigne d'une mise à disposition simultanée des données ayant subies des niveaux de traitement différents.

Pour mieux comprendre ces aspects, nous allons présenter plus en détail chacune de ces deux collections.

1.2.1. Collection « Pratique vocale monodique au Pays Basque français (1957-2012), Marie Hirigoyen Bidart de 2003 à 2012 »[25]

Cette collection est composée de 44 contenus ou documents sonores, pour la plupart des entretiens avec des représentants du chant et de la musique basque dans le Pays-Basque nord et le Béarn. Les entretiens sont issus des recherches de terrain dans le cadre de la thèse[26] de Marie Hirigoyen Bidart. Chaque document est composé d'un à

[25] [En ligne : https://anpersana.iker.univ-pau.fr/collections/show/3]. Consulté le 26 mai 2017.

[26] « Le chant basque monodique (1897-1990) : analyse musicologique comparée des sources écrites et musicales », Musique, musicologie et arts de la scène, Université Toulouse Le Mirail - Toulouse II,

trois fichiers OGG, numérisés à partir de 23 minidisques et 11 cassettes DAT. Il s'agit donc de données secondaires dans le sens où elles ont subi une transformation au moment de les migrer d'un support physique (minidisque ou cassette) à un format numérique (OGG) stocké dans un serveur. Les transcriptions associées n'ont pas été inclues dans la collection. Il est prévu qu'une collection supplémentaire regroupant des archives sonores non-éditées de chant basque collectées par Michel Itçaina et Marie Hirigoyen Bidart entre 1957 et 2012 soit prochainement diffusée sur la plateforme.

1.2.2. Collection « Correspondance basque du bateau Le Dauphin (1757) »[27]

Cette collection regroupe 50 lettres écrites en 1757 en basque labourdin de l'époque. Le Dauphin est le nom du bateau censé les transporter à la communauté basque établie à Louisbourg, en Terre-Neuve. Le bateau est saisi par la marine britannique ; les lettres n'arrivent donc jamais à destination. Elles ont récemment été trouvées aux Archives de l'Amirauté Britannique. En tant que correspondance sans visée littéraire et faisant partie de la sphère familiale, la découverte représente une opportunité unique pour l'étude de l'histoire de l'écrit en basque, notamment du dialecte labourdin de la côte. Manuel Padilla-Moyano (2015, pp. 15) l'explique ainsi :

> *Le Dauphin itsasontziko gutuneriak abantaila handi bat du : XVIII. mende erdiko lapurteraren hizkuntza egoera islatzearena. Gutunek, konkretuki, Lapurdiko hamar barietate ordezkatzen dute, eta guztiak 1757ko otsailaren 2tik apirilaren 1era doan tartean izan ziren idatziak. Xede komunikatibo duten idazki pribatuak izateagatik, eta are gehiago jende apalak eginak, ezagutzen dugun beste zein-nahi testuk baino modu fidagarriagoan islatzen dute XVIII. mende erdiko lapurtera mintzatua, 1. atalean azaldu den hurbilaren hizkuntza hura ematen duten idazkiak baitira.*[28]

Soutenue le 25-09-2012. [En ligne : https://tel.archives-ouvertes.fr/tel-00747090/document]. Consulté le 05 juin 2017.

[27] Aurélie Arcocha-Scarcia, Gwendal Denis, Xabier Lamikiz, Jean-Philippe Talec et Charles Videgain. [En ligne : https://anpersana.iker.univ-pau.fr/collections/show/12]. Consulté le 26 mai 2017.

[28] Traduction du Figaro, dans un article de Marie-Amélie Blin, publié les 3 mai 2016 : « Notre connaissance du basque était jusque-là surtout basée sur des œuvres littéraires, les autres types de textes se faisant rares, mais cette découverte comble ce manque. Elle va nous permettre d'étudier la langue telle qu'elle était parlée au milieu du XIIIe siècle, par des gens de la classe moyenne et des milieux modestes. »

Les contenus de cette collection sont publiés sur des fichiers distincts : d'une part l'image de la lettre en PDF et d'autre part les transcriptions du texte en PDF aussi. Deux types de transcriptions sont inclues dans le même fichier : une version A, qui reprend l'orthographe et la grammaire originales, et une version B, dans une colonne en regard, plus proche de l'écriture moderne, adaptant partiellement l'orthographe et la syntaxe. Les données de cette collection, par rapport à la collection « Pratique vocale monodique … », sont soumis à des traitements plus nombreux : un fichier avec les données d'origine (ayant subies seulement une numérisation, ce sont des données brutes) et un fichier contenant simultanément des données ayant subies plusieurs niveaux de traitement (transcription exacte et transcription légèrement adaptée au parler moderne). La publication sur ANPERSANA de ces lettres et de leur transcription fait partie d'un projet plus large, qui vise à inclure la « traduction en français/anglais, l'annotation et l'archivage … tout en prévoyant la quête de documents semblables en d'autres langues [breton, gaëlique, occitan] provenant d'autres bateaux dont la documentation a été localisée à Londres ». (VIDEGAIN et al., 2014, p. 1)

1.3. Fonctionnalités et choix techniques de la plateforme web

Engagé dans les principes du libre accès, le centre de documentation d'IKER a créé ANPERSANA. Cette plateforme vient s'ajouter aux initiatives existantes en faveur de la valorisation et de la visibilité du laboratoire, de sa communauté scientifique et de leur domaine de recherche. Fonctionnalités et choix techniques de la plateforme web doivent alors répondre à ces objectifs, en publiant des données de façon à ce qu'elles soient repérables, accessibles, interopérables et réutilisables (principes FAIR).

Le logiciel choisi pour la création du site web est OMEKA[29], spécialement conçu pour la diffusion de collections numériques, parmi lesquelles celles des bibliothèques. L'un des aspects qui a motivé ce choix des créateurs du projet ANPERSANA est la possibilité de décrire les contenus et les collections selon les éléments Dublin Core (DC). DC est un schéma de métadonnées qui vise la description de contenus sur le web avec une précision suffisante pour permettre l'interprétation du contenu décrit, tout en cherchant un maximum de simplicité pour que les données soient interopérables et

[29] Site web d'OMEKA. [EN ligne : https://omeka.org/]. Consulté le 20 avril 2017.

facilement indexées par les moteurs de recherche. Ainsi, la description avec DC passe par le renseignement des 15 éléments DC définis dans ce schéma de métadonnées.

Pour la description des collections et des contenus sur ANPERSANA, une version étendue de DC a été mise en place, que nous résumons dans le tableau ci-dessous :

Tableau 1 : Eléments Dublin Core pour ANPERSANA

Eléments DUBLIN CORE (DC) dans l'ordre utilisé sur ANPERSANA (Anp.) - 15 éléments (E.) de base - 1 élément supplémentaire (ES) - 16 éléments de raffinement ou *qualifiers* (Q)		
E. DC	**Ordre Anp.**	**Intitulé**
1	1	Titre
Q	1.1	Autre forme de titre
2	2	Créateur
6	3	Contributeur
5	4	Éditeur
7	5	Date
Q	5.1	Date de disponibilité
Q	5.2	Date de création
Q	5.3	Date d'acceptation
Q	5.4	Date du copyright/de droit d'auteur
Q	5.5	Date de soumission
Q	5.6	Date de parution
Q	5.7	Date de modification
Q	5.8	Date de validité
8	6	Type
9	7	Format
Q	7.1	Étendue de la ressource, taille, durée
12	8	Langue
11	9	Source
3	10	Sujet mots clés
4	11	Description
Q	11.1	Résumé
Q	11.2	Table de matières
14	12	Couverture
Q	12.1	Couverture spatiale
Q	12.2	Couverture temporelle
15	13	Droits
Q	13.1	Droit d'accès
Q	13.2	Licence
13	14	Relation
10	15	Identifiant
S	16	Provenance

Les *qualifiers* (Q) sont des éléments de raffinement qui apportent des informations supplémentaires et précises dans les cas où ils sont importants pour la ressource décrite.

L'objectif des éléments supplémentaires (S) est le même. Cependant, au lieu d'apporter une précision sur un élément faisant déjà partie du DC de base, c'est un élément supplémentaire qui est décrit. Dans le cas d'ANPERSANA, cette information supplémentaire est la « provenance », essentielle pour tracer l'origine d'un document et sa succession d'appartenances. Pour le renseignement des différents éléments, le logiciel OMEKA facilite le contrôle du vocabulaire utilisé grâce à un outil de création des listes fermées. C'est ainsi que la personne responsable de la plateforme peut définir au préalable le vocabulaire, limitant les mauvaises pratiques lors de la saisie d'informations. L'avantage est que l'auto-archivage peut ainsi être envisagé sans se soucier davantage de l'utilisation d'une terminologie erronée de la part d'un non-spécialiste de la documentation informatisée.

OMEKA est aussi compatible avec tout un éventail d'extensions libres de droits d'utilisation qui peuvent être incorporées. ANPERSANA en utilise deux qui lui permettent d'offrir des fonctionnalités de géolocalisation et visualisation des contenus. Ces applications améliorent certes la navigation sur le site et son attractivité visuelle, mais ce sont surtout des outils ajoutant de la valeur aux données mises en ligne.

Le plug-in *Geolocation*[30] relie à Google maps l'information de géolocalisation renseignée dans l'élément de raffinement DC « couverture spatiale » et stockée dans la base de données d'OMEKA. Le résultat est la possibilité d'effectuer des recherches par « localisation » dans la plateforme et de parcourir les items affichés sur une carte. Dans ce sens, c'est une disposition graphique et automatisée des données qui facilite une approche d'analyse géographique (importante, par exemple, en Dialectologie).

UniversalViewer[31], également intégré dans ANPERSANA, offre la possibilité de visualiser les documents d'image. Toutes les numérisations des originaux de la collection « Correspondance du Dauphin… » bénéficient de plusieurs options de visualisation : rotation, zoom, plein écran ou réduit et glissement de l'image. Ainsi, l'analyse de l'objet (le type d'écriture, le format de la lettre, la taille), et pas seulement de son contenu, est envisageable à distance.

---

[30] Roy Rosenzweig, Center for History and New Media.
[http://omeka.org/codex/Plugins/Geolocation_2.0]. Consulté le 05 juin 2017.

[31] Développé par la communauté web. [En ligne : https://github.com/UniversalViewer]. Consulté le 05 juin 2017.

Un élément additionnel contribuant à visibiliser les travaux de recherche du laboratoire est la proposition d'un format de citation des données. OMEKA permet l'affichage d'une référence bibliographique pour le contenu consulté, facilitant ainsi la démarche de ceux qui pensent y faire référence ou les réutiliser.

1.4. Procédures de dépôt et diffusion des données

La procédure de dépôt des données concerne le stockage sur le serveur Oparo de l'Université de Pau et des pays de l'Adour (UPPA). En effet, toute diffusion sur la plateforme ANPERSANA est précédée par le stockage des données sur de serveur de l'université et leur description. Pour le stockage, à ce jour, il n'y a pas de conditions très contraignantes. Il suffit que les fichiers de données soient organisés par dossiers dans une arborescence logique et qu'ils ne dépassent pas 100 Go de volume par collection. La répartition des tâches entre le gestionnaire de la plateforme et le producteur des données est pour le moment également flexible : (i) soit les fichiers sont directement déposés par le chercheur ; (ii) soit les fichiers sont envoyés au gestionnaire pour un stockage ultérieur.

Cette flexibilité des conditions et procédures n'est pas pour autant en opposition avec une bonne gestion de ses données. Au contraire, les bonnes pratiques sont présentes tout au long du cycle de vie des données. La prévision de deux jeux de fichiers pour chaque document, par exemple, permettra de choisir le bon format pour la diffusion sur ANPERSANA (évitant les reconversions improvisées). Le bon nommage des fichiers empêchera aussi les problèmes de migration à partir du (ou vers) le serveur. D'ailleurs, le stockage dans le serveur de l'université fait partie des pratiques préconisées de gestion de données. Dans ce sens, le producteur des données aurait intérêt à déposer les données de ses recherches sur Oparo, même si leur diffusion n'est pas envisagée, avec le seul objectif d'un stockage sûr.

Avant la diffusion, nous avons évoqué la procédure de description des données, qui se concrétise par la saisie des éléments étendues DC sélectionnées pour les collections d'ANPERSANA. Comme pour le dépôt des fichiers, ici aussi le producteur a deux options : (i) il peut partiellement déléguer la tâche au gestionnaire, sachant que c'est le producteur qui connaît l'information à saisir dans certains éléments DC et que donc sa collaboration minimale est requise ; ou (ii) décrire les données en autonomie. Un accès

d'administrateur, avec un identifiant et un mot de passe, pour une partie limitée d'ANPERSANA, lui permettra d'accomplir cette tâche. Le gestionnaire collabore ici en tant que spécialiste de la documentation, aidant le producteur à comprendre quelle information précise est attendue dans chaque élément DC, surtout dans ceux qui regroupent des informations en lien avec son sujet de spécialisation (droits, licences, formats de fichiers).

A terme, l'objectif est l'autonomisation du rôle de l'utilisateur, c'est-à-dire de promouvoir l'auto-archivage des données. C'est seulement ainsi que la bibliothèque numérique pourra être enrichie davantage de collections ; le gestionnaire ne pouvant pas prendre en charge le volume de travail que cela supposerait. Pendant le stage au centre de documentation d'IKER, nous avons voulu contribuer à cette démarche d'autonomisation en créant deux documents qui la soutiendraient :

- Un guide de saisie des métadonnées (cf. Annexe 2, p. 84) a été produit. Pour chaque élément ou *qualifier* DC nous avons inclus : la traduction au français du terme DC original (en anglais) ; le terme DC en anglais (si besoin d'effectuer une recherche d'information sur un élément, étant donné que le site du DCMI est en anglais) ; la définition ; au moins un exemple de saisie ; des commentaires additionnels potentiellement utiles lors de la saisie ; le codage (ou type de vocabulaire : libre, contrôlé ou liste fermée) ; et le statut de l'élément (obligatoire, obligatoire si applicable ou optionnel recommandé).
- Pour ceux qui auront rempli au préalable le modèle de PGD d'IKER, nous avons également conçu une table de correspondances (cf. Annexe 3, p. 97) entre le PGD et les éléments de DC étendu intégrés dans le logiciel OMEKA (cf. Tableau 2). L'idée de cette table de correspondances est d'aider à trouver l'information dont on a besoin et que l'on a préalablement recueillie dans le PGD.

Pour conclure, il convient de signaler que tant le dépôt comme la description des données ne sont jamais définitifs. Même quand les contenus ont été publiés sur la plateforme, nous pourrons y revenir pour ajouter, modifier ou supprimer des informations ou des fichiers. A titre d'exemple, la collection « Correspondance du Dauphin… » est en ligne et pour autant le projet est loin d'être complètement abouti, en attente notamment d'être enrichie davantage avec des annotations. D'ailleurs,

l'amélioration de la qualité des données et métadonnées, contribue à leur repérage et à leur visibilité sur le web, ainsi qu'à leur exploitation et réutilisation.

Nous avons décrit les objectifs, fonctionnalités, choix techniques et procédures de dépôt d'ANPERSANA, ainsi que son contexte de création. Les caractéristiques dégagées vont nous servir maintenant pour situer la bibliothèque numérique dans l'écosystème d'entités travaillant pour les données de la recherche et à la comparer à d'autres entrepôts de données.

2. ANALYSE D'ANPERSANA AU REGARD D'AUTRES ENTREPOTS

Cécile Delay-Artous présente en 2015 un essai de cartographie des entités des données de la recherche dans les SHS en France et à l'international. En s'inspirant de son plan cartésien (2015, diapositive 8), nous avons voulu placer ANPERSANA dans cet écosystème pour savoir la place qu'elle occupe. En effet, dans un contexte d'émergence des projets sur l'*Open Access* et l'*Open Research Data*, chaque infrastructure, acteur, projet ou entrepôt des données doit définir son champ d'action par rapport à ce qui a déjà été mis en place, évitant le chevauchement de services et d'activités. Ainsi, nous intégrerons d'abord ANPERSANA dans la cartographie de Delay-Artous. Ensuite, nous comparons ANPERSANA avec une sélection d'entrepôts de données présents dans la cartographie.

2.1. Cartographie des entités des données de la recherche : acteurs et initiatives, structures et entrepôts

2.1.1. Méthodologie de la cartographie

La méthodologie utilisée par Delays-Artous est la suivante :

(i) Définition des critères du plan cartésien : en abscisses le type d'actions menées, avec les outils à l'ouest et les incitations à l'est (les actions de management se situant vers le centre) ; en ordonnées le comportement, avec les acteurs au nord et les projets au sud (les comportements de type réseau se situant vers le centre). Le code couleur est le suivant : **Bleue** = Français, **Jaune** = Européen, **Rouge** = International

(ii) Sélection des entités à placer dans le plan : La sélection n'est pas exhaustive et reprend les « les items rencontrés dans les textes … au cours des lectures ». Parmi les entités « rencontrées », sont inclues les entrepôts de données et les catalogues d'entrepôts de données, ainsi que les acteurs de l'*Open Access* et des données scientifiques ; sont exclues les archives dédiées uniquement aux publications scientifiques, et les acteurs de l'*Open Data* et des données administratives qui ne se soucient pas des données de la recherche ; sont exclues également les « incontournables de la recherche … et du numérique » tel que le CNRS, les ministères, le CNIL, etc., car leurs champs d'action vont au-delà des questions liées aux données de la recherche et de la science ouverte.

(iii) Description des entités, préalable à la représentation graphique : Dans un tableau divisé en trois colonnes, les critères à dégager sont le statut de l'entité, la pérennité, la fonction d'outil ou de service et les objectifs nommés. Dans la 1re colonne, l'information en forme de liste concerne le « qui fait quoi, et pourquoi » et la catégorie. La 2e colonne est renseignée avec des informations sur le nom de la structure, les acronymes, le site web et les sources d'information. La 3e colonne est un copié-collé de la présentation officielle de l'entité, normalement accessible sur son site web.

Pour situer ANPERSANA dans la cartographie de Delay-Artous, nous avons repris le tableau de description d'entités par souci de s'en tenir à la méthodologie initiale. Nous nous sommes aperçus également du fait que certaines entités jusque-là non-incluses méritaient d'y être intégrées : COCOON et OSF. Donc, les tableaux avec la description pour chacune de ces entités ont été ajoutés (cf. Annexe 4, p. 100). Comme dernière remarque avant de continuer, il convient d'expliquer que de nombreuses plateformes de diffusion de contenus en langue basque existent à ce jour. Certaines ont un parcours large et connu (ahotsak.eus ou euskomedia.org), d'autres, plus nouveaux, ont été présentées lors du Colloque [32] inaugural de l'Ethnopôle basque EKE [33] (mintzoak.eus ; Ondarebideak ; BDB Bertsolaritzaren datu-basea ; etc.). Ces

[32] Baionako Euskal Museoa = Musée Basque de Bayonne, article informatif sur le colloque du 19 mai 2017, « Le colloque inaugural de l'Ethnopôle basque a abordé la question des ressources culturelles numériques », 24 mai 2017. [En ligne : http://www.eke.eus/fr/nouvelles/colloque-inaugural-ethnopole-basque-abordera-la-question-des-ressources-culturelles-numeriques-et-de-la-creation]. Consulté le 2 juin 2017.

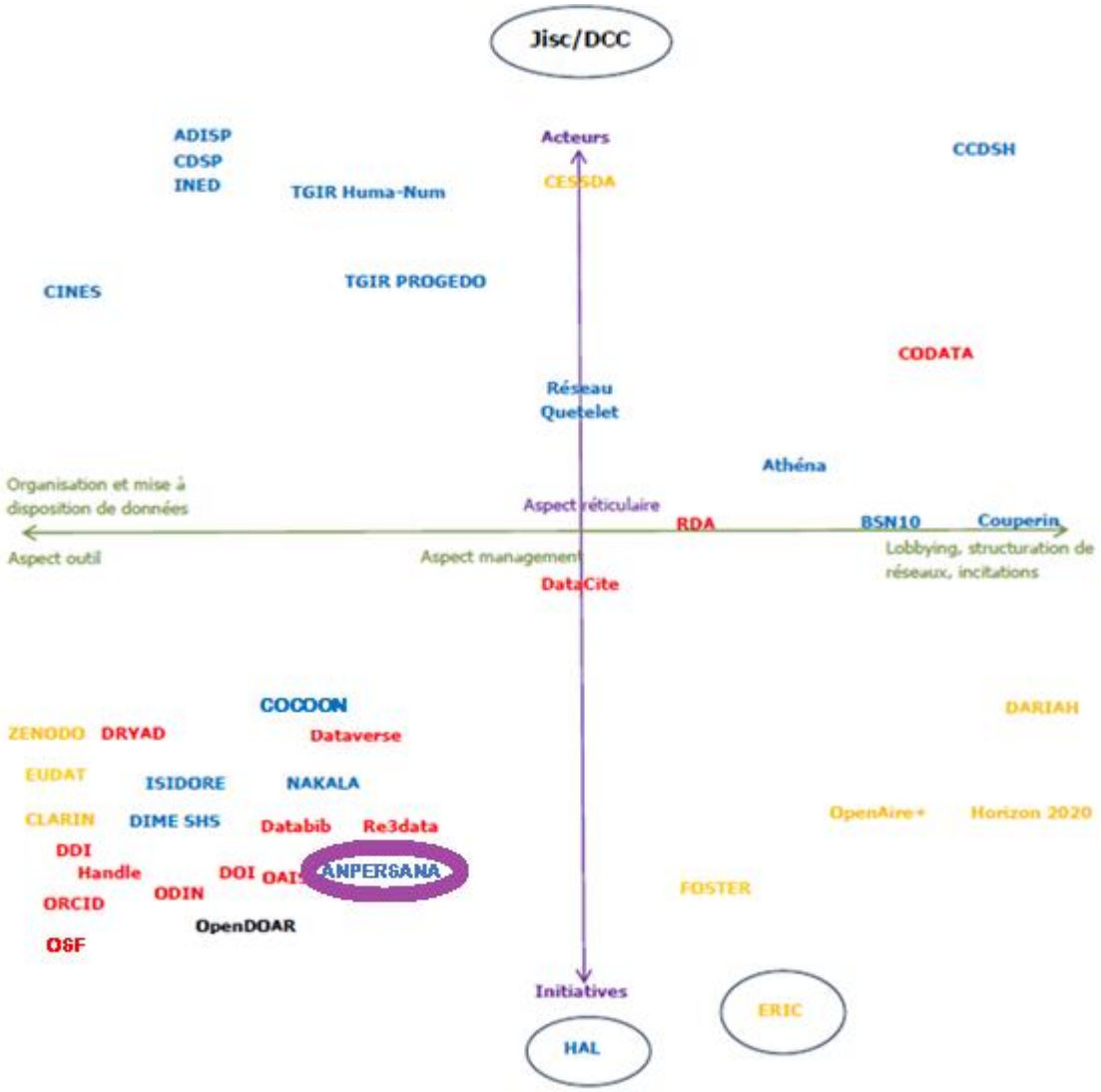
[33] Euskararen Kultur Erakundea = Institut Culturel Basque (ICB). [http://www.eke.eus/fr]. Consulté le 2 juin 2017.

plateformes de ressources numériques rentrent dans le champ du culturel et du patrimonial. Même si elles sont potentiellement utiles à certains projets de recherche, nous les avons exclues de la cartographie car les données diffusées ne sont pas obligatoirement issues de la recherche.

2.1.2. Localisation d'ANPERSANA et description

A partir des tableaux descriptifs (cf. Annexe 4, p. 100), nous avons pu placer les trois nouvelles entités dans le plan cartésien ci-dessous. Elles viennent ainsi s'ajouter à la zone sud-ouest déjà très peuplée :

Graphique 1 : Cartographie de Delay-Artous avec des données additionnelles



COCOON se situe dans une zone légèrement plus au nord qu'ANPERSANA et

OSF. En effet nous avons considéré que COCOON voit sa nature « d'initiative offrant un service » nuancée par le fait qu'il représente aussi une référence en ce qui concerne les données sonores. Pour cette raison, nous avons situé ANPERSANA et OSF dans l'extrême sud de la région sud-ouest. On constate également qu'ANPERSANA et COCOON d'une part, et OSF de l'autre, sont légèrement éloignés par rapport à l'axe horizontal : leur condition d'outil ou d'application de diffusion étant leur caractéristique principale, les outils plus à l'ouest offrent tout un ensemble de services additionnels.

2.1.3. Analyse de la cartographie

Nous savions que l'émergence des projets autour des données de la recherche était d'actualité et nous avons constaté grâce à la cartographie que la région sud-ouest du plan est chargée. Cela illustre qu'il existe un certain nombre d'initiatives qui se concrétisent dans des outils (applications ou plateformes) de diffusion de données des SHS. Affirmer que ce sont trop ou trop peu d'initiatives n'est pourtant pas évident, car il faut être capable de prouver l'existence d'un chevauchement des services, ou au contraire l'existence de besoins et de manques. Si nous avons précédemment utilisé le terme « écosystème », c'est pour illustrer qu'un certain équilibre est nécessaire. Pour cela, il est nécessaire de savoir les services spécifiques offerts par chacun des entrepôts, les fonctionnalités et choix techniques des applications utilisées, le type de données diffusées et le domaine spécifique de recherche d'où sont issues les données ; or le plan cartésien ne prend en compte ces critères. Pourtant, Cécile Delay-Artous identifie que toutes les 17 entités au sud-ouest (19 sur 20 maintenant, avec l'intégration d'OSF et COCOON) sont à caractère multidisciplinaire.

Ce n'est pas suffisant pour établir l'existence des superpositions notables, car parmi les 19 plateformes multidisciplinaires certaines offrent des services différenciés et d'autres traitent des types de données spécifiques. Nous pouvons en revanche confirmer qu'ANPERSANA est l'un des rares entrepôts de données spécialisé dans un domaine. Les frontières du champ d'action du projet ANPERSANA sont ainsi clairement définies : être porteur de la diffusion des données dans les études basques pour soutenir la recherche du laboratoire IKER, qui est la seule structure de recherche sur le basque en France.

2.1.4. Limites de la cartographie

La sélection des entités n'est pas exhaustive et donc les généralisations que nous avons pu évoquer dans l'analyse sont à nuancer. Il se peut que la proportion réelle des outils multidisciplinaires et des outils spécifiques à un domaine se verrait modifiée si l'intégrité des outils existants étaient inclues dans la cartographie. Les entités les plus visibles ont probablement été privilégiés, donc les multidisciplinaires qui sont en général plus larges que les autres.

Nous avons rencontré des difficultés au moment de placer les nouvelles entités dans la cartographie. Le fait qu'elles se situent au sud-ouest du plan a été une question évidente : COCOON, OSF et ANPERSANA sont des initiatives offrant des services spécifiques grâce à leurs fonctionnalités techniques. Ce sont surtout des logiciels. Mais n'ayant pas accès à la méthodologie complète, notamment la description détaillée des critères des axes horizontal et vertical, leur localisation précise au sein de cette région sud-ouest est seulement approximative.

Le type de données diffusées et le domaine spécifique de recherche d'où sont issues les données ne sont pas prises en compte dans le plan cartésien. En effet, cette absence est imposée par le choix d'une représentation par un plan cartésien, qui se caractérise par l'identification de deux critères qui se déclinent à des degrés variés (type d'actions menées / comportement). C'est pour cette raison que nous proposons ensuite d'analyser comparativement les spécificités d'une sélection de plateformes, prenant en compte des critères qui n'ont pas pu été inclus dans la cartographie.

2.2. Comparaison des choix techniques et services des plateformes

Nous interprétons le fait d'être le seul entrepôt de données issues de la recherche dans les études basques comme une opportunité pour le développement de la plateforme et pour la visibilité du laboratoire. Mais c'est aussi une responsabilité puisque que cela entraine une visibilité accrue de ce qui est produit, au niveau des données, dans les études basques. La question à se poser, dans une démarche d'amélioration continue, est celle de savoir comment faire en sorte qu'ANPERSANA soit une vraie valeur ajoutée pour la communauté scientifique d'IKER. Quelles fonctionnalités et services peuvent être développés pour rendre ses collections repérables, accessibles, interopérables et

réutilisables ? La comparaison d'ANPERSANA avec d'autres entrepôts similaires peut nous donner quelques pistes au sujet de ces questions. Nous pourrons nous inspirer de ce qui a été fait ailleurs pour répondre aux problématiques spécifiques à son champ d'action.

2.2.1. Méthodologie pour la comparaison des entrepôts de données

Un tableau comparatif a été conçu au regard de ces questions. Les lignes représentent les différentes composantes des entrepôts de données, que ce soit du point de vue de leur domaine d'action ou des services offerts. Les colonnes renseignent les points d'entrée de ces informations pour chacune des plateformes sélectionnées. Nous insistons sur le fait qu'une composante, pour nous, est tout service, fonctionnalité, choix technique et procédure, ou même le domaine d'action et les types de données recouverts par l'entrepôt. Les composantes susceptibles d'être intégrées sont issues d'une collecte continue effectuée au cours de la rédaction du mémoire et des missions du stage.

En ce qui concerne les colonnes, apparaissent ANPERSANA ainsi que trois autres entrepôts de données : Zenodo, OSF et COCOON. Cette sélection cherche à intégrer des entrepôts confirmés et largement utilisées par la communauté scientifique à laquelle ils rendent service. Ils représentant aussi une certaine diversité de composantes. Nous avons choisi également un nombre d'entrepôts réduit pour que le tableau reste visuellement accessible.

2.2.2. Tableau comparatif et analyse

D'après la méthodologie décrite, voici ci-dessous le tableau comparatif réalisé :

Tableau 2 : Comparateur d'entrepôts de données

	ANPERSANA	OSF	Zenodo	Cocoon
Types de données	Tous	Tous	Tous	Sonores
Domaines	Etudes basques : linguistique, littérature, didactique	Multidisciplinaire	Multidisciplinaire, SHS	Multidisciplinaire, SHS
Langue des données linguistiques	Principalement le basque	Toutes les langues	Toutes les langues	Toutes les langues
Langue de l'outil	Français (traductions au basque, à l'anglais et à l'espagnol en cours)	Anglais	Anglais	Français
Langue de description de données	Français (traductions au basque, en cours)	Au choix du producteur	Au choix du producteur	Au choix du producteur
Gestion de projets	NON	OUI	OUI	NON
Outil collaboratif (fonctionnalités de réseau social)	NON	OUI	Partiel (intégration de projets dans communautés autogérées)	NON
Autoarchivage	Sur demande (évolution vers l'autonomisation)	OUI	OUI	Sur demande
Chronologie de diffusion de données	Les données et les métadonnées peuvent être ajoutées et modifiées au fur et à mesure	Les données et les métadonnées peuvent être ajoutées et modifiées au fur et à mesure	Les données et les métadonnées peuvent être ajoutées et modifiées au fur et à mesure	Les données et les métadonnées peuvent être ajoutées et modifiées au fur et à mesure
Diffusion des données ou seulement stockage	Choix de l'usager et selon droits	Choix de l'usager et selon droits	Choix de l'usager et selon droits	Choix de l'usager et selon droits
Intégration d'outils	NON	OUI	NON	NON

Serveur stockage	Université (France)	Serveur d'*OSF storage* : inconnu Serveurs des outils de stockage externe : selon outil	CERN *document server* (Europe)	Selon grille d'HumaNum (France)
Pérennisation	NON	Partiel (control anti-obsolescence des formats automatisé)	OUI, CERN	OUI, CINES
Présentation des données	Par projet, organisés en collections de type bibliothèque numérique	Par projet, organisation hiérarchique de dossiers et fichiers	Par projet, organisation hiérarchique de dossiers et fichiers	Par projet, organisés en collections moins visuelles que sur une bibliothèque numérique classique
Diffusion simultanée de publications	NON	NON (seulement *preprints*)	OUI	NON
Citation de données proposée	OUI, un seul format	OUI, éventail très large de formats de sortie possibles	OUI, éventail très large de formats de sortie possibles	OUI, un seul format
GUID, Identificateur global unique	En cours (ARK ou DOI envisagés)	DOI	DOI	HDL et OAI
Description des ressources : schémas de métadonnées	Dublin Core étendu	(information introuvable)	JSON Schema (options d'exportation vers : MARCXML, Dublin Core, DataCite Metadata Schema)	Dublin Core et OLAC RDF (actuellement en usage interne uniquement)
Protocole d'échange de métadonnées	NON	Via SHARE pour les *preprints* (inconnu pour le reste de données)	REST Api ; OAI-PMH	OAI-PMH
Annotations de texte	Au choix du producteur de données ; Projet d'annotations des contenus d'une collection en TEI en cours	Au choix du producteur de données	Au choix du producteur de données	Au choix du producteur de données
Géolocalisation	OUI, sur carte, *Geolocalisation* relié à *Google maps*	NON	NON	OUI, sur carte, *OpenStreetMap* relié à *ArcGIS Maps for Office*
Visualisalion	OUI, *UniversalViewer* intégré pour les images (zoom, rotation, navigation)	Pas de service particulier	Pas de service particulier	OUI, application inconnue pour visualisation de la transcription synchronisée de l'audio

Nous constatons que les approches de développent d'un outil de diffusion de données peuvent être diverses. Dans ce sens, les projets spécialisés dans un domaine ou un type de donnée semblent développer des services spécialisés en relation avec leur champ d'action. Ainsi COCOON intègre une application de lecture synchronisée des transcriptions pour accompagner l'audio de ses données sonores. Le schéma de description de métadonnées est également adapté aux exigences de description du type des données décrits : sonores dans le cas de COCOON qui utilise des éléments OLAC[34] en plus de DC. ANPERSANA aussi, concernant les données linguistiques, envisage le développement d'annotations de texte en TEI.

En revanche, les plateformes pluridisciplinaires offrent un éventail de services plus large. La nature pluridisciplinaire de la plateforme peut en quelque sorte justifier l'existence de quelques-uns de ces services, dans le sens où il faut répondre aux besoins différents de chaque discipline. Par exemple, de plus nombreux domaines entrainent une plus grande variété de formats de références bibliographiques, conformes aux exigences spécifiques de l'édition scientifique de référence en la matière. D'autres services offerts, comme la gestion de projets de recherche ou la diffusion en parallèle des publications scientifiques associées aux données, semblent être liés aux objectifs de départ qui visent à permettre une gestion de données intégrale. Ces plateformes tendent à être un espace où l'on peut tout faire.

Nous présenterons dans la partie suivante les composantes susceptibles d'apporter de nouvelles opportunités pour ANPERSANA, la communauté scientifique d'IKER et les études basques.

[34] Open Language Archives Community. [En ligne : http://www.language-archives.org/OLAC/metadata.html]. Consulté le 2 juin 2017.

3. ANPERSANA DANS L'AVENIR

Nous allons à présent décrire les actions stratégiques prévues à court terme. Puis, dans la partie Recommandations, nous verrons les composantes identifiées grâce au tableau comparatif (cf. Tableau 3) qui viendraient s'ajouter à la réflexion sur leur possible inclusion dans l'outil ANPERSANA.

3.1. Actions prévues

3.1.1. Annotations de texte en TEI

Nous avons signalé que les lettres de la collection du Dauphin seront annotées. Cela permettra « le traitement de données (noms de personnes, lieux, évènements, à intérêt historique et linguistique, comme les diverses formes d'un nom propre), … des analyses linguistiques et structures de traits et … la gestion d'en-têtes et de métadonnées. » (VIDEGAIN et al., 2014, p. 13) Ces annotations se feront en *Text Encoding Initiative* (TEI), largement accepté comme étant une vraie valeur ajoutée pour l'analyse linguistique. Nous avons pour autant constaté que le niveau de précision des annotations en TEI peut être très variable. En effet, la description des données linguistiques contenues dans un document s'effectue à plusieurs niveaux. Nous pouvons distinguer des regroupements de données linguistiques plus larges comme la phrase et la proposition grammaticale, ou réaliser des annotations plus minutieuses sur les syntagmes, les mots, voir les composants morphologiques de ces derniers. Du type d'annotation dépendra la possibilité de réaliser des recherches plus ou moins précises dans le texte ou le type de traitement que l'on pourra réaliser à l'aide d'un logiciel de *text-mining*. Ces aspects seront certainement pris en compte pendant le déroulement du projet du Dauphin. D'ailleurs, les contributions d'un linguiste ainsi que d'un spécialiste en TEI sont prévues, pour d'une part aborder les aspects linguistiques des annotations et de l'autre concrétiser les choix dans un code adapté et correct.

3.1.2. L'archivage pérenne de données

L'archivage pérenne des données est aussi l'un des projets d'ANPERSANA à court terme. Assurer la pérennité des données déposées signifie garantir que les données produites au laboratoire soient accessibles et réutilisables dans l'avenir. Pour ceux qui ont une conscience de l'obsolescence technologique et des questions d'interopérabilité, cela suppose un service pour le moins attirant. Pour les autres, ce sera l'occasion de leur transmettre l'importance de ces aspects et de l'intérêt de déposer ses données sur un entrepôt qui assure la pérennité (au-delà des démarches de diffusion). L'option la plus envisageable à ce jour est le travail en partenariat avec un opérateur qui a les moyens pour assurer l'archivage pérenne. Le CINES est le choix évident, en tant qu'opérateur spécifiquement développé pour les données issues des structures de l'ESR. Pourtant, une fois la faisabilité du projet confirmée et les procédures de dépôt accordées, la migration vers leurs serveurs s'accomplit sous certaines préparations préalables qui demandent la collaboration des producteurs de données. Le CINES publie ces conditions dans leur site web et accompagne la structure déposante sur demande.

Une collaboration avec le CINES serait en plus l'occasion d'attribuer un identifiant pérenne aux ressources numériques d'ANPERSANA. Actuellement, l'identification de contenus s'achève par un code d'utilisation interne qui sert à inventorier les différents contenus qui sont accessibles via le protocole http d'un URL. Dans certains cas, quand l'URL n'est plus disponible (parce que l'architecture du site a changé) ou quand l'URL renvoie à une autre ressource (parce qu'une nouvelle ressource remplace l'originale), nous ne pouvons plus accéder au contenu souhaité. Le but d'un identifiant pérenne est d'éviter ces problèmes et de garantir que le contenu soit toujours accessible. Le CINES utilise des identifiants ARK (*Archival Ressources Key*) attribués par la California Digital Library.

3.1.3. Données moissonnables - Protocole OAI-PMH

Les contenus d'ANPERSANA, nous l'avons dit, ont été décrits avec le schéma de métadonnées DC pour que les moteurs de recherche puissent les indexer. L'étape suivante serait de faciliter l'échange des métadonnées avec les fournisseurs de services qui seraient potentiellement intéressés par les données publiées sur ANPERSANA. Cela

permettrait de multiplier la visibilité de la plateforme grâce aux référencements sur les sites des institutions qui appliquent le protocole du côté du *harvester* ou « moissionneur », comme Isidore[35] par exemple. L'OAI-PMH (*Open Archives Initiative - Protocol for Metadata Harvesting*) est un protocole d'échange de métadonnées entre le fournisseur de données et le moissonneur. L'application OMEKA offre la possibilité de télécharger un plugin[36] qui permet d'exposer les métadonnées de son entrepôt selon ce protocole, basé sur l'information des éléments DC. ANPERSANA ayant adopté et appliqué la description avec DC, il suffirait d'accomplir la mise en place technique de ce plugin et de signaler aux fournisseurs de services la disponibilité des données.

Maintenant que nous avons balayé les pistes d'amélioration de la plateforme envisagées à court terme, nous allons présenter quelques recommandations que nous considérons intéressantes pour la suite.

3.2. Recommandations

Ces recommandations incluent les composantes que nous avons identifié lors de la comparaison d'entrepôts (cf. Tableau 3) ou au cours de notre travail pendant le stage. L'objectif de les regrouper ici est de nourrir la réflexion sur les stratégies d'amélioration de la plateforme pour diffuser des données FAIR, répondre aux besoins de la communauté scientifique d'IKER et favoriser la recherche dans les études basques.

3.2.1. Description Dublin Core multilingue

Il est évident que les utilisateurs d'ANPERSANA sont pour la plupart des locuteurs de la langue basque. Pourtant, même si les données diffusées sont issues d'un centre de recherche du CNRS, la communauté scientifique qui compose le laboratoire n'est pas toujours bilingue basque-français, certains sont bilingues basque-espagnol. Il est ainsi important de prendre en compte les utilisateurs de la plateforme souvent spécialistes de la langue basque qui ne parlent pas le français : soit parce qu'ils sont originaires du Pays

[35] [En ligne : https://www.rechercheisidore.fr/]. Consulté le 2 juin 2017.

[36] John Flatnes. Dernière version : 2.1.1, 2016. [En ligne : https://omeka.org/add-ons/plugins/oai-pmh-repository/]. Consulté le 2 juin 2017.

Basque sud, soit parce qu'ils font partie de la communauté internationale des chercheurs. Cette réalité oblige à réfléchir sur l'aspect linguistique de l'offre des services. Dans ces sens, nous avons signalé qu'une démarche de mise en ligne en basque de la plateforme est en cours. Cependant, l'utilisateur qui voudrait faire une recherche par mots clés en basque avec le moteur de recherche de la plateforme n'obtiendrait surement pas de résultats ; il en va de même pour les recherches avec un moteur de recherche du web, qui n'aurait indexé que les informations que l'on a renseignées, en français, dans les éléments DC.

Comment concrétiser alors dans l'avenir le référencement DC des contenus de façon multilingue ? La réponse n'est pas évidente, car il y a plusieurs options dont la faisabilité sur le logiciel OMEKA ou la validité selon les préconisations de DCMI seraient à tester :

- Option 1 – Ajouter la traduction au basque ou au français précédée du symbole « = » : Nous nous inspirons du catalogage en Unimarc pour proposer ce format de saisie de métadonnées multilingues. Par exemple, dans la collection « Correspondance du Dauphin… », l'élément dc.title serait renseigné ainsi : Correspondance du bateau le Dauphin = Le Dauphin itsasontziko gutunak. Le problème ici repose sur le fait que DCMI préconise l'ajout d'un nouveau champ d'élément, ou *element iteration*, pour chaque information.
- Option 2 – L'itération d'éléments DC : en tant qu'option recommandée par DCMI, et réalisable sur OMEKA, elle semble être l'option à favoriser. L'élément dc.title ressemblerait à l'exemple suivant :

  Title : Correspondance du bateau le Dauphin

  Title : Le Dauphin itsasontziko gutunak
- Option 3 – Traductions partiellement automatisées : Si nous proposons une troisième option, c'est parce que les deux précédentes exigent un travail excessivement long. La troisième option consisterait à prévoir les traductions à renseigner en synchronisant le site web en basque (en cours d'élaboration) et le site en français. Le déposant ou le gestionnaire de la plateforme pourraient ainsi saisir les informations dans la langue de leur choix. Au-delà du fait que cela implique la création au préalable de listes fermées avec des vocabulaires contrôlés dans les deux langues, nous ne savons pas si cette synchronisation est techniquement

faisable avec le logiciel OMEKA. Sheila Brennan du personnel d'OMEKA, dans le forum du site, donne une réponse ambiguë à un utilisateur qui se pose la même question que nous : « you can … have two mirror installations that are connected by the theme's navigational elements but contain separate items, exhibits, and pages in their own language ». Donc, il faudrait installer OMEKA deux fois, et synchroniser ensuite les deux plateformes. En revanche, Brennan considère cette option « *challenging* » d'un point de vue technique et favorise notre option 2 en évoquant la possibilité d'ajouter des « *fields for each metadata element so that it is available in more than one language*. ». (2013) Quatre années sont passées depuis cette réponse et un plugin[37] plus récent a été développé pour proposer le site en plusieurs langues, mais nous ne l'avons pas testé au niveau de la synchronisation multilingue des éléments DC.

3.2.2. Description OLAC pour les données sonores

COCOON, entrepôt des données sonores issues de la recherche en SHS, utilise des éléments OLAC en plus des éléments DC. Nous avons déjà évoqué que l'avantage de DC réside dans sa simplicité, car la contrainte du nombre réduit d'informations rend les métadonnées plus interopérables. Donc, par principe, nous émettons des craintes sur l'ajout d'éléments supplémentaires. En revanche, le fait de savoir que COCOON est une plateforme confirmée nous encourage dans l'idée d'utiliser ce type de métadonnées pour les contenus sonores d'ANPERSANA. D'ailleurs, le Guide de bonnes pratiques numériques d'HumaNum (2015), le TGIR parent du projet COCOON, recommande leur utilisation. Cinq attributs ou éléments de raffinement seraient intégrés dans le schéma DC étendu d'ANPERSANA pour être en concordance avec les consignes d'HumaNum et les pratiques des COCOON :

(1) « language » sous l'élément DC « subject » : dc.language est déjà un élément de DC. L'ajouter en tant qu'attribut de dc.subject suppose de considérer la langue de l'enregistrement comme sujet de recherche, au lieu de le limiter à une référence à la langue du contenu.

(2) « linguistic-field » sous l'élément DC « subject » : Une liste fermée avec les

[37] Omeka_Plugin_SwitchLanguage, dernière version du 8 décembre 2016. [En ligne : https://gitlab.com/TIME_LAS/Omeka_Plugin_SwitchLang]. Consulté le 31 mai 2017.

sous-domaines de la linguistique (morphologie, dialectologie, syntaxe, etc.) est préconisée.

(3) « discourse-type » sous les éléments DC « type » et « subject » : Une liste fermée avec les divers types de discours (oratoire, dramatique, narratif, chant, etc.) est préconisée.

(4) « linguistic-data-type » sous l'élément DC « type » : Une liste fermée avec les divers types de données linguistiques de la locution (texte_primaire, lexique, description_linguistique, etc.) est préconisée.

(5) « role » sous les éléments « contributor » et « creator ». Une liste fermée avec les types de contribution (transcripteur, traducteur, éditeur, chercheur, etc.) est préconisée. Sur ANPERSANA la contribution est ajoutée entre parenthèses suivant le nom de l'auteur ou le contributeur.

3.2.3. Visualisation synchronisée des transcriptions

L'entrepôt de données sonores COCOON nous sert de nouveau d'exemple, cette fois-ci en ce qui concerne la visualisation de données. Les fichiers audios qui sont accompagnés de fichiers de texte contenant la transcription peuvent être simultanément écoutés et lus de façon synchronisée. Une flèche ainsi que du surlignage de texte indiquent l'extrait qui représente la partie de l'enregistrement que l'on écoute. Le bénéfice évident de cet outil est de pouvoir suivre le locuteur facilement, même dans les cas où on ne parle pas sa langue. Mais ce serait également un atout pour la phonétique et la phonologie, la transcription orthographique permettant de distinguer le début et la fin des mots, voire de morphèmes, représentant les phones qui sont souvent fusionnés, modifiés ou élidés dû au contact avec d'autres phones. Cette application peut certainement être utile pour ANPERSANA. Dans les deux collections publiées sur le site jusqu'à présent, il n'y en a aucune contenant à la fois du son et des transcriptions d'enregistrements. Cependant, ce type des données seront potentiellement les plus fréquemment déposées dans l'avenir, car il s'agit du type de données que l'on produit davantage à IKER (cf. Troisième partie, 1.4.).

Pour évaluer l'intégrabilité de cet outil dans le logiciel OMEKA, nous avons

contacté directement Michel Jacobson du Laboratoire Ligérien de Linguistique[38], responsable de la conception du site COCOON. Suite à sa réponse, nous avons compris qu'il s'agit en effet de pas un seul, mais deux outils, qu'il faudrait développer parallèlement. D'abord, il faut un outil qui permette le placement des *timecodes* dans les annotations de la transcription (Elan, Clan, Praat ou Transcriber, par exemple). Ensuite, il faudrait un outil de visualisation de ces annotations. Dans le cas de COCOON, c'est Jacobson même qui intègre cette capacité de visualisation en développent le code en *javascript* qui le permet. Mais, il y a des logiciels aussi qui peuvent être utilisés pour ce type d'exploitation, notamment des annotations en format SMIL (*Synchronized Multimedia Integration Language*), basé sur XML.

Si SMIL est évoqué, c'est parce qu'ANPERSANA prévoit les annotations de texte en TEI. En effet, SMIL est le format recommandé par le W3C (TEI, 2016), organisation de référence développant des standards pour le web, pour son intégration dans TEI. C'est encore un exemple de ce qui peut être envisagé avec ce format. Au-delà des annotations contenant de l'information linguistique précédemment évoquées, nous voyons maintenant toute une autre utilité, la synchronisation de la transcription et de l'audio.

3.2.4. Remarques sur les services non-inclus dans les recommandations

Pour conclure, nous voulons évoquer le fait que de ces trois recommandations, aucune n'est finalement issue des entrepôts multidisciplinaires étudiés. La première a été repérée lors de la constatation pendant le stage d'un besoin de répondre à la réalité linguistique d'une communauté. Les deux suivantes nous les avons dégagées à partir des composantes de l'entrepôt COCOON qui, comme ANPERSANA, opère dans un champ d'action plus réduit que celui des plateformes pluridisciplinaires. Dans notre idée de départ, nous croyions pouvoir nous inspirer du large éventail de services offerts par les entrepôts visant l'intégralité des types de données et des domaines de la recherche. OSF et Zenodo permettent la gestion des projets de recherche du début à la fin et ont des fonctionnalités de réseaux social. En revanche, nous avons trouvé que ces services dépassent les objectifs d'ANPERSANA et que leur mise en place ne serait pas réaliste à

[38] LLL - UMR7270 - Universités d'Orléans et de Tours, BnF, CNRS. [En ligne : http://www.lll.cnrs.fr/]. Consulté le 12 juin 2017.

court ou moyen terme. La possibilité qu'offre Zenodo de diffuser en simultanée la publication scientifique associée aux données ne nous paraît pas, non plus, utile pour ANPERSANA, les contenus étant déjà associés par le moyen d'hyperliens aux travaux publiés sur la revue *Lapurdum* ou sur la plateforme d'autoarchivage Artxiker.

Certaines composantes ont été exclues parce qu'ils dépassent nos compétences et donc la possibilité d'évaluer leur pertinence. En ce qui concerne la possibilité de proposer des formats de référence bibliographiques multiples, facilitant la citation des ressources, nos explorations n'ont pas suffi à trouver une solution qui puisse s'intégrer dans le logiciel OMEKA. Nous avons également écarté une éventuelle discussion sur la syntaxe RDF et l'opportunité de s'en servir pour exposer les hyperliens aux moteurs de recherche.

Après avoir décrit les fonctionnalités d'ANPERSANA, situé la plateforme dans l'écosystème d'entités de données de la recherche et formulé des recommandations, la Troisième partie du mémoire sera l'occasion de présenter l'enquête menée au sein de la communauté scientifique d'IKER, au sujet des pratiques et perceptions concernant les données de la recherche.

TROISIEME PARTIE : Enquête sur les données de la recherche auprès de la communauté scientifique d'IKER

Nous avons dressé un panorama de la littérature concernant les grandes problématiques des données de la recherche et conduit une analyse de l'outil ANPERSANA. Pour compléter ces explorations, il nous paraît essentiel d'aborder le sujet du point de vue de la communauté concernée. Notre idée de départ a été d'élaborer un questionnaire exploratoire avec l'objectif d'obtenir un aperçu des pratiques et perceptions des chercheurs du laboratoire IKER concernant les données de la recherche en générale et le projet ANPERSANA en particulier. Suite à l'analyse des résultats du questionnaire, de nouvelles questions ont émergé et nous avons souhaité approfondir certains de ces aspects. C'est ainsi que notre enquête se décline en deux temps : le questionnaire exploratoire d'abord, et des entretiens semi-directifs ensuite.

1. QUESTIONNAIRE

1.1. Objectifs

L'objectif principal du questionnaire est de faire un état des lieux, ou bilan des faits : d'une part, pour comprendre les enjeux de l'ouverture des données de la recherche dans le cas concret du laboratoire IKER et, dans la mesure du possible, dans les études basques ; et d'autre part, identifier et définir des profils types de chercheurs du laboratoire. Concrètement, nous cherchons à connaître les pratiques de gestion des données au sein de la communauté scientifique d'IKER et à savoir quelles sont leurs perceptions au sujet de l'ouverture des données de la recherche. Nous souhaitons également identifier les chercheurs qui sont intéressées par le projet de bibliothèque numérique ANPERSANA et qui voudraient y diffuser leurs données.

Parallèlement, les résultats du questionnaire seront potentiellement utilisés par le centre de documentation d'IKER pour créer des formations ou accompagnements adaptés aux besoins des chercheurs.

1.2. Méthodologie

Les objectifs principaux étant la réalisation d'un état des lieux et la définition des

profils, le questionnaire est à visée exploratoire. Nous avons entrepris une approche ouverte et généraliste pour limiter les biais liés à nos préconceptions. C'est ainsi que dans cette première partie de l'enquête, aucune hypothèse n'est formulée et les questions du questionnaire prétendent toucher les grands sujets en relation avec l'ouverture des données de la recherche. Dans cette perspective, le résultat est un questionnaire (cf. Annexe 5, p. 103) composé de 51 questions regroupées en 6 groupes. Les deux premiers groupes de questions ont été conçus pour définir le statut du répondant et ses projets de recherche. Les 3 groupes suivants sont divisés en sujets concernant la gestion des données : pratiques, droits / aspects éthiques et collaboration scientifique. Le 6e et dernier groupe est composé d'une série de questions qui visent à connaître l'avis des participants à propos de l'ouverture des données de la recherche et de la science. Le logiciel utilisé pour la création et la diffusion du questionnaire, ainsi que pour la réception de réponses, est LimeSurvey[39]. Nous avons intégré une version française et une version basque des questions, équivalentes, pour répondre à la réalité linguistique de tous les participants. La communication avec les participants a eu lieu également dans la langue de leur choix.

1.3. Echantillon

Le public visé est toute la communauté scientifique du laboratoire IKER, qui est composée de chercheurs du CNRS, d'enseignants chercheurs d'universités, de post-docs, de doctorants et d'ITA[40], tous spécialisés dans le domaine des études basques. Ce premier groupe de participants potentiels est composé d'une trentaine de personnes. Nous avons voulu faire parvenir le questionnaire à un groupe plus ample, avec l'intention d'avoir une vision suffisamment proche de la réalité des données de la recherche dans les études basques (sans pour autant avoir la prétention de mener une enquête représentative de la communauté scientifique plus large). Pour cela, nous avons enrichi le nombre de participants potentiels y incluant également des chercheurs qui sont de différentes façons proches du laboratoire : des chercheurs qui étaient dans le passé rattachés à celui-ci et d'autres qui sont rattachés à des organisations qui travaillent

[39] Disponible sur [URL : https://www.limesurvey.org/], consulté le 21 avril 2017.

[40] Ingénieur, technique et administratif du CNRS

conjointement avec IKER dans certains projets sur des langues minoritaires autre que le basque, comme le gascon (occitan)[41] et le breton. 51 personnes ont été finalement visées. Les contacts ont été ajoutés à la base de données de l'application LimeSurvey et ont été contactés via courriel électronique. Ceux qui n'ont pas répondu au questionnaire lors du premier message, nous les avons recontactés 8 jours plus tard par le moyen d'un message de rappel. A la date limite établie pour recevoir les réponses, 26 personnes ont complété le questionnaire dans sa totalité : 6 d'entre elles sont des chercheurs du CNRS, 5 des enseignants-chercheurs d'universités, 5 des post-docs, 9 des doctorants avec contrat et un doctorant sans contrat.

1.4. Résultats

Statut des participants

67,74 % des participants sont rattachés à des établissements dépendant du ESR, tandis que 22,58 ont des contrats avec l'Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), sous l'administration espagnole. Les 9,68 restants font partie de l'Académie de la langue basque (Euskaltzaindia, qui a le statut d'association), d'une université allemande ou n'ont pas de rattachement à une organisation. En ce qui concerne l'âge des participants, il y a une représentation plus grande des plus jeunes : trois quarts d'entre eux ont en effet moins de 46 ans. De la même façon, les années d'expérience des participants dans la recherche sont presque équivalentes aux données sur l'âge : 65% ont moins de 10 ans d'expérience et 35% en ont plus de 10.

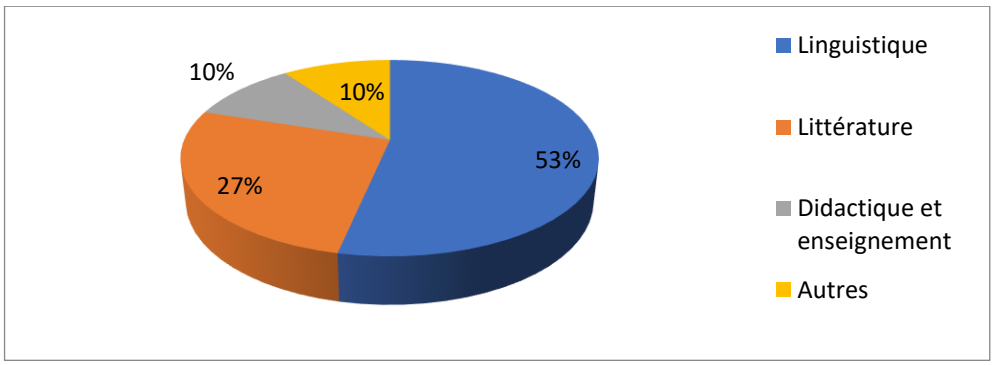
Spécialités et types de données produits au laboratoire

La spécialité la plus représentée est la linguistique, 16 sur 26 participants ayant choisi cette option. Il faut signaler que certains chercheurs sont spécialistes dans plus d'un domaine ou mènent des travaux interdisciplinaires. Donc, les 26 répondants ont donné 30 réponses à la question « quelle est votre spécialité », 22 ayant coché une seule spécialité et 4 ayant choisi 2 options chacun. Finalement, un peu plus de la moitié des

[41] La majorité des spécialistes considèrent que le gascon est un dialecte de l'occitan ; d'autres le situent comme une langue appart, dû à son système vocalique différencié : Wikipédia, « Gascon », dernière actualisation de l'article le 14 avril 2017. [En ligne : https://fr.wikipedia.org/wiki/Gascon]. Consulté le 5 mai 2017.

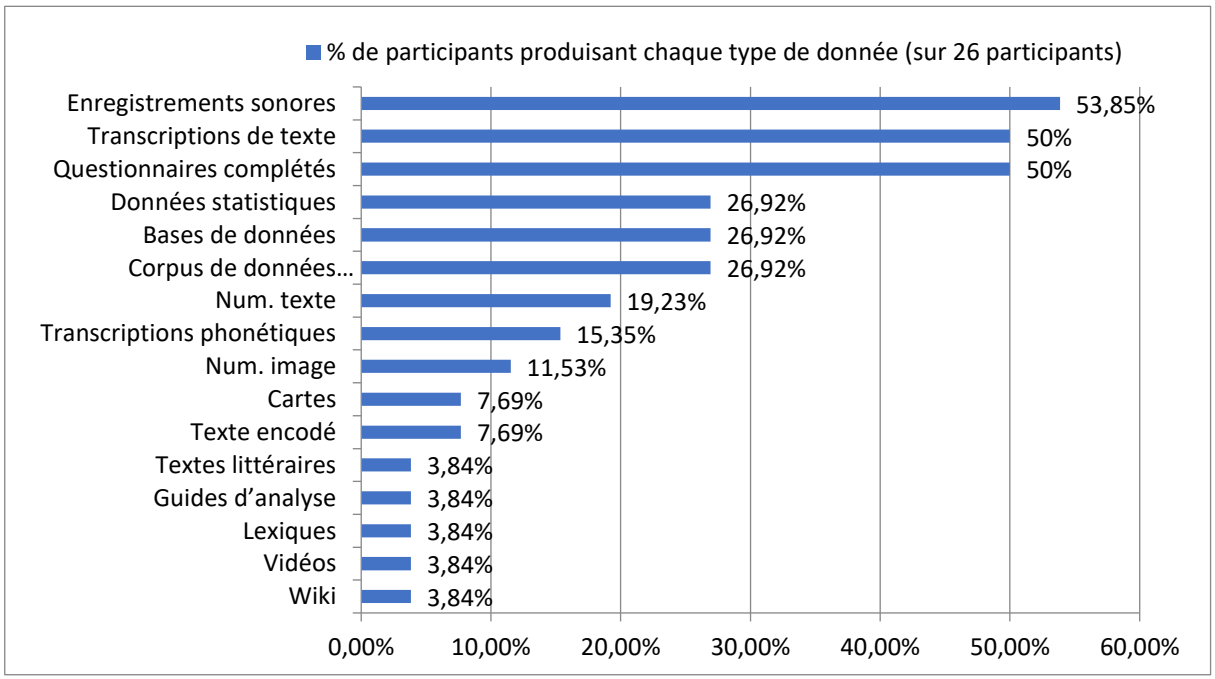
30 réponses correspondent à la linguistique, 8 à la littérature, seulement 3 pour la didactique et l'enseignement, et enfin un pour chacune des disciplines suivantes : musicologie, les sciences politiques et l'histoire.

Graphique 2 : Domaines de spécialisation



Nous avons également pu établir quels sont les types des données collectées et produites dans le laboratoire. Il y a une diversité manifeste de celles-ci, mais les plus courantes sont des enregistrements sonores, des transcriptions de texte et des questionnaires complétés : environ la moitié des participants en produisant. Dans un deuxième échelon, nous trouvons les données statistiques, les bases de données et les corpus de données linguistiques, produits par un peu plus d'un quart des répondants.

Graphique 3 : Types de données



Pratiques de stockage des données

Pour regrouper la diversité de réponses sur les pratiques de stockage des participants, nous avons utilisé une grille qui permet de classifier les pratiques de stockage par niveau de risque lié à l'emplacement des fichiers. Cette grille nous l'avons conçue dans le cadre de l'élaboration d'un Guide des bonnes pratiques en gestion de données dirigée à la communauté scientifique d'IKER ; nous ajoutons ci-dessous une version incluant une colonne supplémentaire recensant le nombre de participants qui déclarent utiliser tel ou tel type d'emplacement.

Tableau 3 : Pratiques de stockage des participants et niveau de risque

DEGRÉ DE RISQUE	EMPLACEMENT	PARTICIPANTS
Pratique à haut risque	1 emplacement : normalement sur le poste informatique utilisé pour la création du fichier	1/26
Pratique à risque modéré	2 emplacements	14/26
Pratique recommandée	3 emplacements ou plus, dont au moins 2 sur site et au moins 1 en ligne[42]	11/26

Pratiques de partage et de diffusion des données

La majorité des participants, 17 sur 26, partagent leurs données avec d'autres chercheurs et 16 sur ces 17 sont des spécialistes en linguistique. En effet, la totalité des linguistes (16/16) déclarent partager leurs données. Il y a par ailleurs 7 participants qui ont des inquiétudes par rapport au partage de leurs données. En ce qui concerne la diffusion en ligne, 12 participants sont prédisposés à le faire dans l'avenir, 13 s'interrogent et seulement 1 donne une réponse négative.

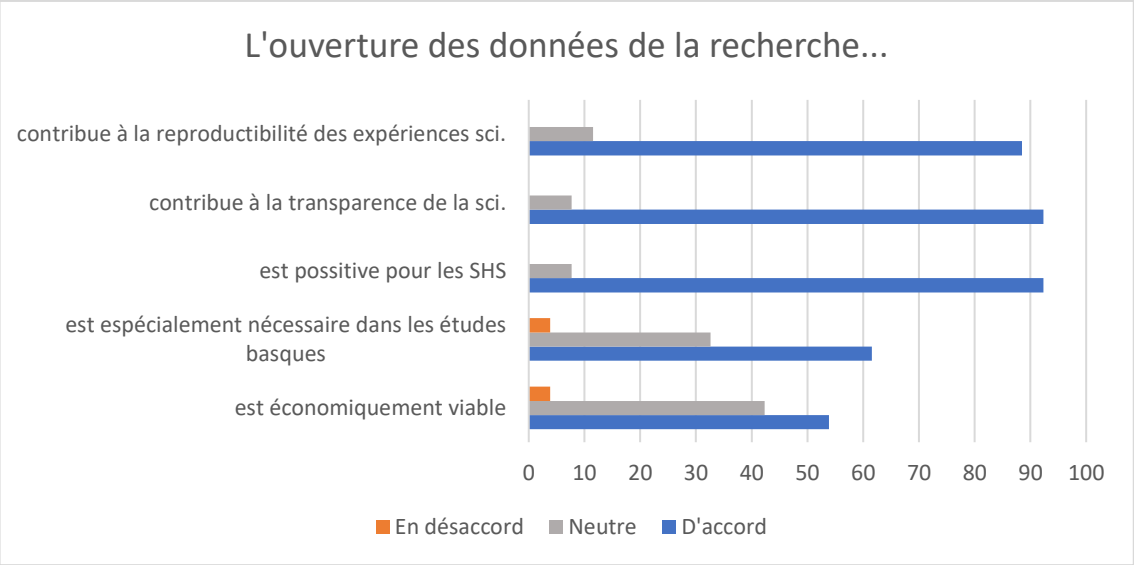
Vision de l'ouverture des données de la recherche et sur ANPERSANA

Dans le questionnaire, nous avons mis 5 degrés progressifs d'adhésion pour la dizaine de propositions : « complètement d'accord », « d'accord », « ni d'accord, ni pas d'accord », « pas d'accord » et « pas du tout d'accord ». Pour la présentation des

[42] Le niveau de sureté des emplacements en ligne est variable, selon la localisation des serveurs du fournisseur de services et le protocole de transfert de données appliqué.

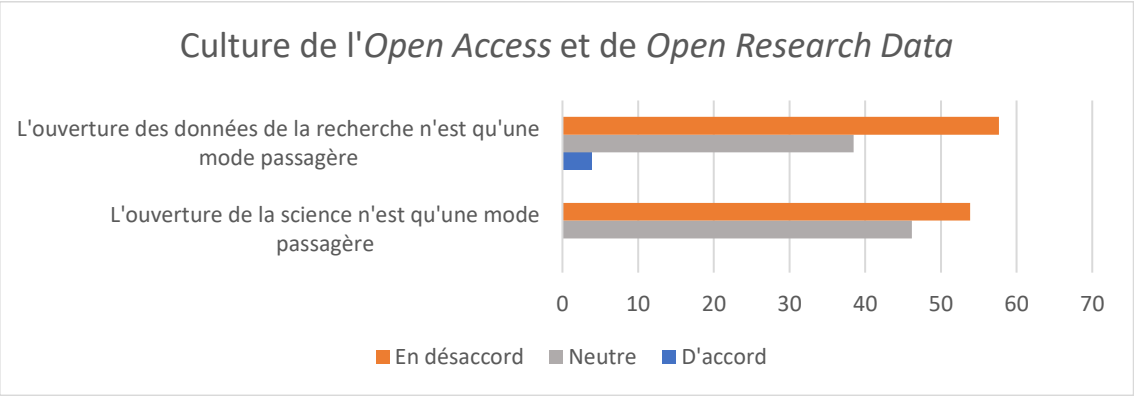
résultats nous avons voulu les regroupés en 3 catégories : « d'accord », « neutre » et « en désaccord ». En général, l'avis sur l'ouverture des données de la recherche est positif ou très positif. Environ 90% des répondants considèrent que l'ouverture des données de la recherche est positive pour les SHS et contribue à la transparence de la science ainsi qu'à la reproductibilité des expériences scientifiques. Nous avons cependant vu précédemment que les participants sont plus indécis quant à la viabilité économique de l'ouverture des données de la recherche : presque 45% ne se positionnent pas et 1 participant déclare qu'elle n'est pas viable. Un peu plus de 60% évoquent que l'ouverture des données en études basques est particulièrement nécessaire, par rapport à d'autre domaines. Un tiers ne sait pas et seulement une personne est en désaccord avec cette proposition.

Graphique 4 : Opinions sur l'ouverture des données de la recherche (1)



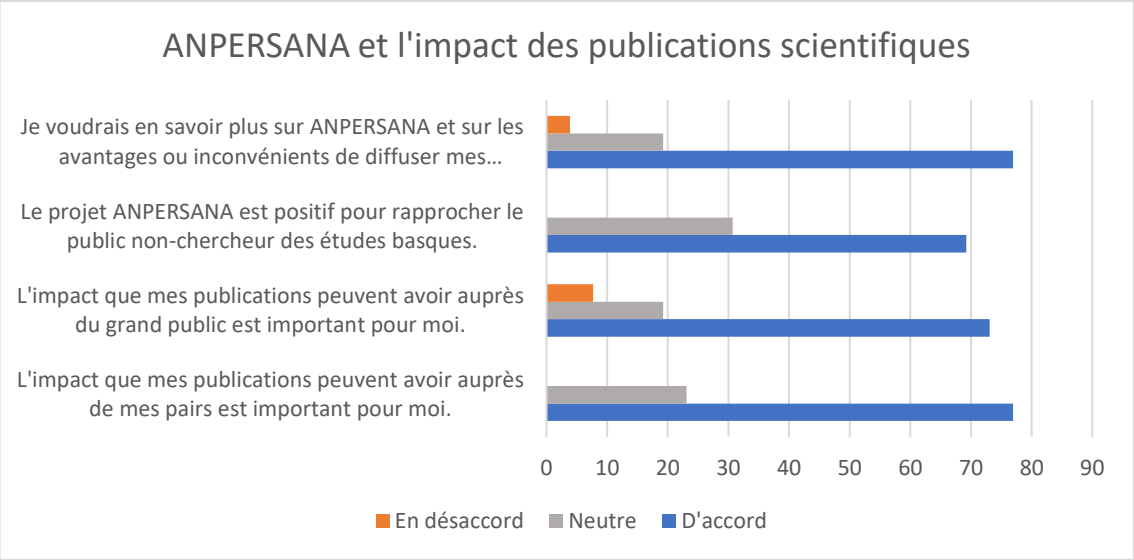
L'indécision est présente également concernant la possibilité que la culture de l'*open* ne soit qu'une mode passagère. Dans ce sens, presque 40% des réponses sont neutres pour les données de la recherche, environ 45% pour les publications scientifiques. Néanmoins, dans le cas des deux propositions, plus de la moitié de la communauté scientifique du laboratoire ne croit pas qu'il s'agisse d'une mode passagère.

Graphique 5 : Opinions sur l'ouverture des données de la recherche (2)



Les répondants sont plus de 70% à donner de l'importance à l'impact qu'ont leurs publications auprès du grand public ; ils sont 80% concernant l'impact auprès des pairs. Au sujet d'ANPERSANA, presque 70% considèrent qu'il s'agit d'un projet pouvant contribuer à rapprocher le public non-chercheur des études basques ; et presque 80% voudrait en savoir plus sur ce projet ainsi que sur les avantages ou inconvénients de diffuser leurs données par ce moyen.

Graphique 6 : Opinions sur ANPERSANA et sur l'impact des publications sci.

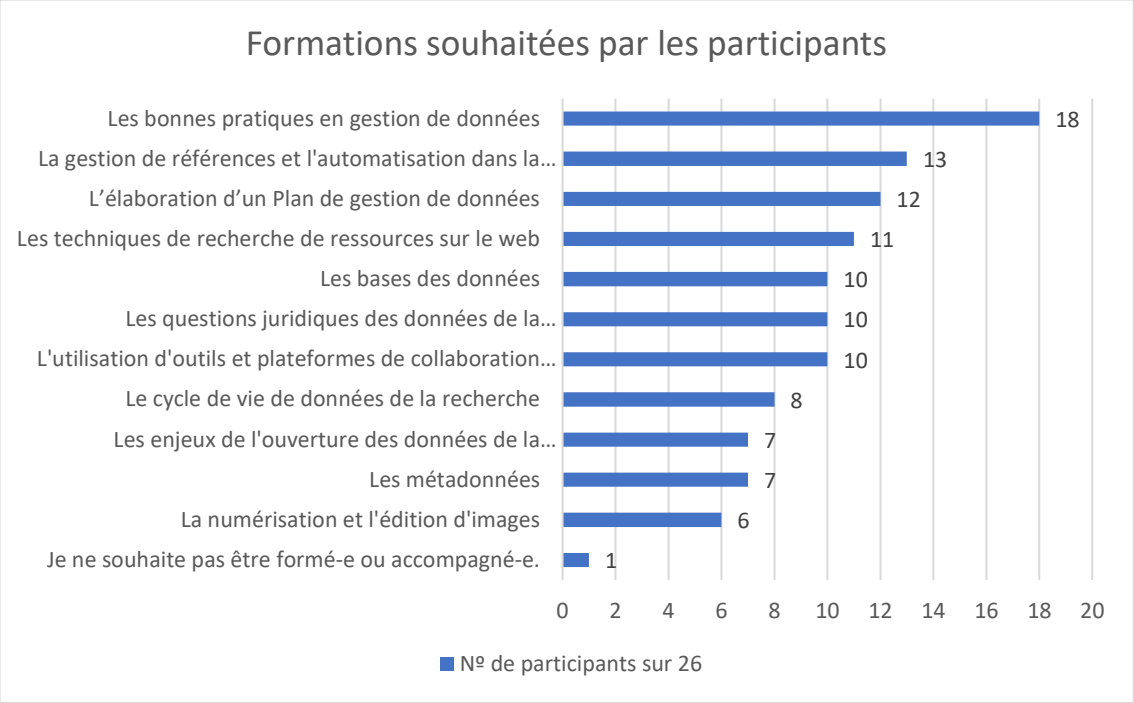


11 des 21 répondants ont déclaré que s'ils diffuseraient leurs données, ils le feraient sur ANPERSANA ; 4 le feraient sur d'autres plateformes et les 6 restants ne prennent pas partie dans ce sens.

Formations sur les données de la recherche

Seulement une personne a suivi des formations sur les données de la recherche. Dans l'éventualité d'une proposition de formations ou accompagnements au laboratoire, le thème qui intéresse le plus est celui des bonnes pratiques en gestion de données (18 sur 26) ; 13 voudraient être formés sur l'automatisation de bibliographies, 12 sur l'élaboration d'un plan de gestion de données et 11 sur les techniques de recherche de ressources sur le web. Ensuite 10 participants citent les thèmes des bases de données, questions juridiques et plateformes de collaboration scientifique.

Graphique 7 : Thèmes que les participants aimeraient approfondir



Les résultats décrits nous donnent un aperçu général des comportements et représentations autour des données de la recherche chez IKER. Nous allons maintenant présenter dans l'analyse les principales tendances que nous avons dégagé du croisement de ces résultats.

1.5. Analyse

Suite aux multiples croisements de résultats que nous avons envisagé (cf. Annexe 6, p. 115), nous allons présenter les principales correspondances qui ont été trouvées. Nous

pouvons avancer qu'aucun lien n'a été repéré entre le statut des participants (l'âge, les années d'expérience, le type de contrat, le type de projet, la structure de rattachement) et des tendances en termes de pratiques ou prédispositions particulières. Les principaux facteurs semblent plutôt être l'accès à la connaissance sur le sujet des données de la recherche et le type de données produites.

1.5.1. Connaissances en matière de données de la recherche

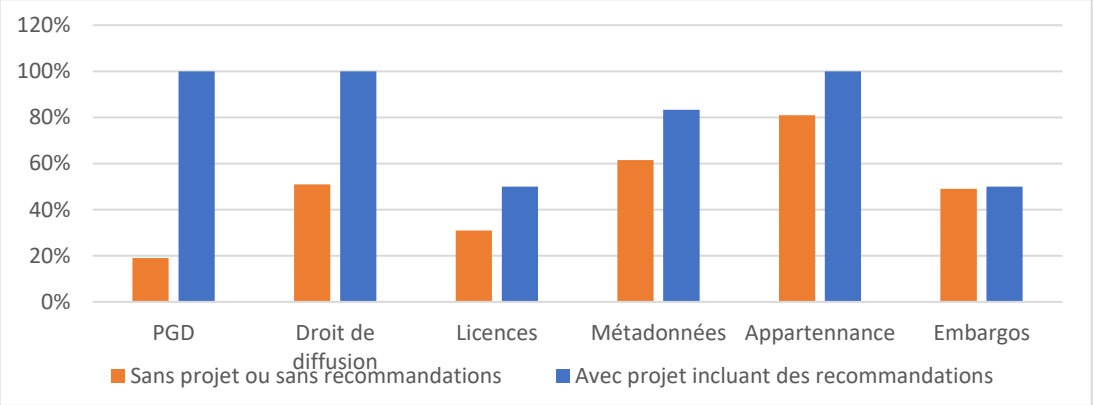
En ce qui concerne les questions juridiques, les participants ont connaissance de la notion de propriété des données. Plus des trois quarts savent si les données qu'ils ont produites leur appartiennent ou pas ; en revanche, les 19% restants ne le sait pas. On peut faire l'hypothèse que ces derniers s'interrogent en raison de la visibilité des notions de « bien commun » ou « bien public » dans le débat public (notamment autour de la loi Lemaire). En revanche, la question de la disposition des données est plus floue pour les participants : presque 40% ne savent pas s'ils ont le droit de diffusion et un peu plus de la moitié ne savent pas s'il y a une période d'embargo associée à leurs données. Le pourcentage s'élève à presque 70% pour ceux qui déclarent ne pas savoir que l'on peut protéger ses données avec la licence de son choix quand on dispose du droit d'auteur.

Concernant les métadonnées, les connaissances à ce sujet sont aussi disparates : 6 participants (23,08%) connaissent bien le concept et 10 (38,46%) le connaissent globalement, les 10 restants déclarent ne pas savoir qu'est-ce que sont les métadonnées. La question sur laquelle les répondants ont montré avoir le moins de compétences est le PGD : seul un répondant maitrise bien le sujet et 4 le connaissent à peu près. En revanche 80 % ne sait pas ce qu'est un PGD. Ce n'est pas étonnant, étant donné que seulement 3 d'entre eux en ont un pour leur projet. Nous avons avancé que le statut des participants n'était pas un facteur déterminant. En relation avec le type de connaissances que l'on vient d'aborder et le fait que certaines notions ne soient pas excessivement claires (notamment le PGD et les périodes d'embargo, et une moindre mesure le droit d'auteur, les licences et les métadonnées), nous avons identifié deux facteurs qui peuvent expliquer cette situation :

(1) La formation : seulement une personne, sur 26, a suivi une formation sur les données de la recherche. Il semble évident que se former davantage contribuerait à l'augmentation des connaissances.

(2) Les précisions des projets de recherche : Seulement 6 des 13 projets de recherche financés comportent des recommandations sur les différents aspects des données de la recherche. Ainsi les 4 participants qui savent ce qu'est un PGD, ont un projet qui inclut ce type de préconisation. Le lien entre les « préconisations dans le projet » et les connaissances se confirme à plusieurs reprises. Nous constatons dans le graphique suivant que le pourcentage de ceux qui maitrisent ces notions augmente dans les cas où ils sont porteurs d'un projet qui inclut des recommandations dans les précisions ou le contrat :

Graphique 8 : Rapport entre connaissances et recommandations dans le projet



Les différences sont plus significatives dans le cas des connaissances sur le PGD, le droit de diffusion et les licences. Pour l'appartenance des données et les métadonnées, l'impact des recommandations est reste marqué mais dans une moindre mesure. Il n'y a en revanche pas de différences significatives dans le cas des périodes d'embargo.

Avant d'aborder d'autres aspects de ces résultats, nous voulons brièvement aborder à nouveau l'interchangeabilité des notions *Open Access* et *Open Research Data* (cf. Première partie, 2.1.), car il y a un exemple de ce type de confusion dans l'enquête. 11 des 21 participants ont déclaré que dans le cas d'une diffusion de leurs données ils le feraient sur ANPERSANA, 10 sur une autre plateforme. Parmi ceux qui ont choisi une autre plateforme, dans le champ de texte libre, certains ont évoqué les entités suivantes : Artxiker et HumaNum. La première est une plateforme d'auto-archivage de publications scientifiques, tandis que la seconde est une TGIR qui est constituée, certes du projet NAKALA (qui s'occupe de la diffusion de données), mais aussi d'autres projets divers autour du libre accès et de l'IST en SHS.

1.5.2. Les données confidentielles

Les résultats montrent que la communauté scientifique d'IKER est en général ouverte au partage et à la diffusion des données. La collaboration scientifique est une réalité et la perception au sujet du libre accès et de l'ouverture des données de la recherche est positive. Nous avons déjà dit que 12 participants souhaiteraient diffuser certaines de leurs données » et 13 ne se prononcent pas mais ne sont pas fermés à cette option. Il convient de souligner que l'on parle ici de diffusion en ligne et pas seulement d'un partage au niveau d'un projet ou entre chercheurs. D'ailleurs, si l'on regarde les données sur le partage, les résultats sont encore plus parlants, car 16 des 26 le font déjà et seulement 7 ont des inquiétudes à ce sujet. Encore une fois, avoir des inquiétudes ne veut pas dire qu'ils s'y opposeraient.

Les résultats les plus significatifs sont issus du croissement des données sur la spécialité avec les données sur les pratiques de partage. Les 16 spécialistes du domaine de la linguistique partagent leurs données. Nous avons trouvé ces résultats étonnants sachant que la linguistique est un domaine où on produit des données pour lesquelles les questions d'anonymat se posent davantage (notamment les enregistrements sonores de type entretien). Pourtant, au laboratoire, parmi les 14 participants produisant des enregistrements sonores, seulement 7 sont des linguistes. Cela, ajouté au fait que tous les linguistes du laboratoire partagent leurs données, confirme que, dans le cas d'IKER, le type de données produites n'est pas associé aux pratiques de partage. Au contraire, nous pouvons confirmer le lien entre les types de données et la prédisposition à la diffusion. En effet, nous trouvons que parmi les 8 linguistes qui ne sont pas surs de vouloir diffuser leurs données, 7 sont ceux qui produisent des enregistrements sonores. Dans ce sens, Cabrera constate dans son étude (2015) qu'en abordant les « principales difficultés ou obstacles » à la diffusion, les interviewés évoquent les « questions de confidentialité ». Nous croyons que cela est d'autant plus vrai pour les études basques. On estime en effet à 800.000 le nombre de locuteurs de la langue basque. A cela s'ajoute une extension territoriale relativement réduite, ce qui rend les participants des enquêtes moins anonymes.

1.6. Limites

Au vu du nombre réduit de public visé et avec un taux de réponse de 50%, l'analyse du questionnaire se base sur les réponses des 26 participants qui composent l'échantillon. Cela a été suffisant pour réussir à faire un état de lieux du laboratoire et identifier certaines tendances au sein de celui-ci. Nous ne pouvons pas pour autant généraliser ces enjeux. En ce qui concerne la formulation de nos questions, nous avons constaté deux difficultés. La question optionnelle « avez-vous un PGD ? » a reçu particulièrement peu de réponses, 13 sur 26 participants. Nous interprétons que la question ait pu être perçue comme un contrôle du travail quotidien des participants. Bien au contraire, la question visait à comprendre les pratiques en gestion de données pour éventuellement apporter un outil pour faciliter leur prise en compte (comme le modèle de PGD d'IKER, par exemple, en Annexe 1, p. 78). La question « Avez-vous des inquiétudes concernant le partage de données ? » visait la notion de partage, dont celle de diffusion aussi, mais pas seulement. Nous trouvons que la question « Avez-vous des inquiétudes concernant la diffusion en ligne de vos données ? » aurait été mieux adaptée pour l'objectif de l'enquête, vu que le type de partage d'ANPERSANA n'inclût que la diffusion. Le taux de réponses affirmatives quant aux inquiétudes aurait été, nous croyons, plus élevé.

Dans la suite de notre enquête, à travers les entretiens, nous essayerons de dépasser certaines de ces limites ainsi d'approfondir les tendances identifiées.

2. ENTRETIENS

2.1. Objectifs

L'objectif principal de cette deuxième partie d'enquête est de connaitre davantage les pratiques et perceptions de la communauté scientifique d'IKER au sujet des données de la recherche. Nous cherchons donc à nuancer et approfondir avec nos interlocuteurs les tendances identifiées lors de l'analyse du questionnaire : (i) le lien entre les recommandations dans les précisions du projet et l'acquisition de connaissances sur les données de la recherche ; (ii) la relation entre le type de données produites et leur

diffusion. Les notions de « partage » et « diffusion » seront également abordées pour clarifier nos questions au sujet de l'une des limites du questionnaire en lien avec cet aspect.

2.2. Méthodologie

Pour répondre à ces objectifs, le choix a été de mener des entretiens semi-directifs : suffisamment encadrés pour aborder les thématiques qui nous intéressent, mais flexibles pour laisser la place à la libre réflexion des interviewés qui pourraient apporter des regards nouveaux. Nous avons donc élaboré une grille d'entretien reprenant les thèmes choisis. La grille a été utilisée pendant les entretiens pour diriger la conversation dans la direction souhaitée, notamment pour parler des aspects qui ne sont pas spontanément abordés. Les entretiens se déroulent entre le 27 avril 2017 et le 3 mai 2017, et sont enregistrés à l'aide d'un téléphone portable. Les fichiers sonores OGG sont ensuite stockés ailleurs selon les pratiques recommandées, supprimés du téléphone et migrés au format WAVE. Le volume des 3 fichiers est de 36 MB et une durée totale de 39 minutes et 21 secondes. Les grilles d'entretien ont été remplies avec les notes prises (cf. Annexe 7, p. 120) lors de l'écoute postérieure aux enregistrements. Nous avons pris les notes en français à partir des entretiens qui ont eu lieu en basque. Les seules transcriptions littérales que nous rendons publiques sont les notions (traduites du basque) que nous avons voulu inclure dans l'analyse des entretiens.

2.3. Echantillon

Nous avons voulu obtenir un échantillon composé de participants qui connaissent bien la vie du laboratoire et qui peuvent apporter des réponses nourries. Trois personnes ont accepté notre proposition de participation à un entretien. Les profils des participants, un PhD chercheur, un PhD récent et un doctorant, sont variés, également en ce qui concerne les années d'expérience dans la recherche. Les disciplines des études basques des informants incluent la dialectologie, la géolinguistique, la sociolinguistique, l'acquisition des langues et l'analyse discursive, ce sont donc tous des linguistes. L'un des participants est également expert en méthodologies scientifiques.

2.4. Analyse

2.4.1. Le respect des sources et le partage de la connaissance : faire converger ces valeurs

D'abord, nous avons voulu clarifier l'interprétation de la question « avez-vous des inquiétudes concernant le partage de données ? » du questionnaire en abordant la notion de « partage » avec les 3 interviewés. Pour savoir si pour eux le partage inclût pour eux la notion de « diffusion en ligne », nous leur avons demandé ce que représentaient pour eux le partage de données. Les 3 ont dédié une partie de l'explication à la notion de « diffusion en ligne ». Donc, il semble que l'aspect « diffusion » est généralement intégré dans la représentation du « partage de données ». Nous pouvons partiellement confirmer alors que les inquiétudes par rapport au partage et à la diffusion des données chez IKER restent peu élevées. Parmi les 3 interviewés, l'un d'entre eux déclare avoir des inquiétudes au niveau personnel quant aux conséquences juridiques de la diffusion ses données en dehors du cadre légal. La méconnaissance de ce que l'on a le droit de faire avec ses données est à l'origine de cette crainte. En parallèle, les 3 informants coïncident sur le fait d'avoir des préoccupations éthiques liées à la diffusion. Ainsi, lorsque les interviewés produisent des enregistrements de productions orales, ils prennent des « précautions » pour « respecter » les parties « compromettantes », « personnelles », « confidentielles » et « intimes » des entretiens, ainsi que « l'anonymat » de leurs sources. Ces valeurs de respect de leurs sources font partie de la déontologie scientifique.

D'un autre côté, les interviewés expriment parallèlement leur « envie » de diffuser ses données en ligne, avec plus ou moins d'effusion. Celui qui l'exprime avec plus de retenue déclare comprendre qu'il y a des « avantages » à la diffusion mais priorise la préservation des conséquences étiques et juridiques. Un autre participant attend la permission des personnes concernées[43] et pense les relancer pour pouvoir diffuser ses

[43] Il a l'autorisation officielle mais souhaite avoir une confirmation de leur part sur la convenance de rendre les données publiques. Sa démarche se justifie par la nature de ses données, dont nous ne pouvons pas donner plus de précisions pour garantir l'anonymat du participant.

données sur ANPERSANA. Le troisième interviewé a l'habitude de déposer ses données sur des plateformes de diffusion autres qu'ANPERSANA. Cette troisième personne conduit une analyse intéressante autour du conflit d'intérêt auquel elle est confrontée. Pour éviter d'avoir l'impression d'être en face d'une décision salomonique (mettre ses données à disposition des autres pour contribuer à l'avancement de la science ou garder les données pour soi, et protéger ainsi ses sources), elle nuance sa posture en fonction des différents types de données. En ce qui concerne les entretiens, les « libres » et les « directifs » sont distingués, les premiers étant « éthiquement plus compromettants ». Qu'il s'agisse des uns ou des autres, le chercheur « choisit de diffuser des données plus ou moins traitées ». Au lieu de publier les « données brutes », on peut publier les « données secondaires » sur lesquelles des « filtres » seront appliqués pour s'accorder aux droits des personnes sources. Ces filtres peuvent être ajoutés à l'aide d'un logiciel, comme « Praat » et incluent « l'édition de contenu compromettant » ainsi que « la suppression de contenu qui permet l'identification ».

Nous concluons avec une brève référence au sujet du partage des données du point de vu de celui qui voudrait s'en servir, que ce soit pour la vérification ou la réinterprétation. Aucun des interviewés n'a senti de frustration dans les situations où les données d'une publication ne sont pas accessibles. Les manières de le gérer sont cependant diverses : il y en a qui préfèrent « faire confiance à l'auteur » et « utiliser les données des résultats ». L'autre option est celle de demander à l'auteur la possibilité de récupérer ses données « mais avec beaucoup de précaution ». Il faut penser que les collègues chercheurs ont des inquiétudes similaires aux nôtres quant au partage. « On essaye de traiter les données des autres tel qu'on aimerait que l'on traite les nôtres. »

2.4.2. Les précisions du projet et le PGD : mieux connaître ses données

Quand nous avons analysé la relation entre les précisions sur les données dans le projet de recherche et la connaissance sur les données de la recherche, nous l'avons intuitivement analysé comme un impact des recommandations du financeur (cf. Troisième partie, 1.4.). Pendant les entretiens, le participant le plus expérimenté se réfère aux précisions d'un projet en parlant de son rédacteur (cf. Annexe 7, nº1, p. 120). La présence de précision sur la gestion des données serait donc du fait du chercheur lui-même. Cette perspective ne change pas notre constatation de la correspondance entre

« précisions sur les données » et « augmentation de connaissances sur les données ». Si les autres chercheurs du groupe ont inclus ces questions dans leur projet de leur propre initiative, et non sous l'impulsion d'un financeur, cela dénote d'une prise de conscience autour des enjeux liés aux données de la recherche : quels sont les sources des données, quelles données on produit, qui les produit, qui y aura accès, quelle est la loi qui encadre les données, comment veut-on les partager et est-ce possible, etc. Certes, le financeur joue souvent un rôle (H2020 notamment investit davantage dans la communication sur la gestion des données), mais l'acquisition des connaissances dépend aussi de l'approche du producteur de données. D'ailleurs, le même interviewé signale qu'« au bout d'un moment on est forcément amené à se poser ces questions lors de la constatation du volume des données qu'on a entre les mains et la responsabilité que cela suppose ». Ce qui est perçu parfois comme une perte de temps, comme l'a documenté Cabrera dans à partir des 50 entretiens de son mémoire (2015) et l'avons constaté nous même dans l'entretien nº2 (cf. Annexe 7, p. 121), peut favoriser au contraire l'efficience du travail de recherche. La gestion des données tout au long de leur cycle de vie limite les problèmes issus d'une mauvaise gestion, ou de l'absence de celle-ci. Et les solutions à postériori exigées par ces potentiels problèmes sont souvent couteuses et demandent parfois l'intervention de spécialistes externes, en informatique par exemple.

Si la prise de conscience de ce que l'on veut et l'on peut faire avec ses données a une relation avec l'inclusion d'une partie dédiée au sujet dans la rédaction du projet de recherche, on est forcément amenés à établir un parallèle avec le rôle du PGD. Dans l'état de la littérature de la première partie (cf. 5.2.) nous avons indiqué qu'un PGD « est un outil de prise de conscience … et de formalisation » des questions autour de la gestion de données. Les deux démarches, rédaction des précisions et constitution d'un PGD ont donc un même objectif. Cependant, tandis que ce deuxième est présent et actualisé tout au long du projet, la rédaction de précisions sur la gestion des données suppose la prévision et la déclaration d'intentions aux prémices du projet, lors de la réponse à un appel d'offre par exemple. Les précisions sont aussi l'occasion de présenter la résolution d'intégrer un PGD qui accompagnera la vie du projet.

CONCLUSION

Le laboratoire de recherche sur la langue et les textes basques IKER du CNRS, de l'UBM et de l'UPPA, a développé tout un éventail de services pour valoriser les travaux de recherche de sa communauté scientifique. Parmi les plus importants, il y a la revue en accès libre *Lapurdum* et la plateforme d'autoarchivage des publications Artxiker. La bibliothèque numérique ANPERSANA vient s'ajouter à ces projets en tant qu'entrepôt de diffusion des données produites par les doctorants et les chercheurs d'IKER.

Dans un contexte d'émergence des projets sur les données de la recherche, nous trouvons qu'il est essentiel de savoir qui fait quoi pour éviter le chevauchement de services. Dans la deuxième partie du mémoire, nous avons situé ANPERSANA dans la cartographie de Delay-Artous pour savoir la place que l'entrepôt occupe dans cet écosystème. ANPERSANA se différencie, en tant qu'entrepôt spécialisé dans le domaine des études basques. La cartographie a également été l'occasion de regrouper une sélection de plateformes, dont les aspects techniques et les services ont été comparés à ceux d'ANPERSANA. Nous avons pu ainsi confirmer qu'ANPERSANA compte à ce jour avec des services contribuant à la repérabilité, accessibilité, interopérabilité et réutilisation des données du domaine, ainsi qu'à la visibilité et à la valorisation du laboratoire IKER, au sein duquel elles sont produites. Cette approche comparative nous a permis de nous inspirer des services mis en œuvre par d'autres entrepôts pour formuler des recommandations qui cherchent à nourrir la réflexion sur les outils susceptibles d'être intégrés dans l'application à moyen terme.

Lors de l'enquête menée au laboratoire, présentée en troisième partie, nous avons trouvé que la bibliothèque numérique, développée pour valoriser les données de la recherche de la communauté d'IKER, est perçue positivement au sein du laboratoire. Il y a aussi une bonne prédisposition à la diffusion de données et la culture de partage prévaut. Le respect de la confidentialité des sources est en même temps une préoccupation majeure, surtout pour ceux qui produisent des enregistrements oraux. Il y a également une considération notable pour le travail des pairs, concernant la réutilisation de leurs données. Nous constatons pour autant que, parfois, ces valeurs de partage de la connaissance et de respect des sources sont perçues comme opposées. Nous avons observé également que les chercheurs du laboratoire porteurs d'un projet comportant des précisions ou recommandations sur les données de la recherche ont plus

des connaissances sur les aspects techniques et juridiques des données. La littérature spécialisée montre que le PGD contribue aussi à la prise de conscience sur ces aspects et permet de les formaliser. D'ailleurs, l'inclusion dans le projet d'une section dédiée à la gestion des données de la recherche, ainsi que l'intégration d'un PGD, sont désormais des conditions requises dans la plupart des appels d'offre des financeurs publics.

La réflexion au préalable sur les données, à partir du moment même de la rédaction du projet, ainsi que l'appropriation d'un outil comme le PGD, sembles donc primordiales. Les chercheurs sont ainsi confrontés à leurs propres données et se posent la question de ce qu'ils veulent et peuvent en faire. Il nous semble donc déterminant d'intégrer ces aspects dès le départ pour éviter une remédiation ultérieure avec des solutions techniques couteuses. Cette prise de conscience sur la gestion des données dès le début permet également d'entrevoir une compatibilité des valeurs de partage et de respect. Le chercheur peut choisir de diffuser des données dites secondaires ou dérivées, pour éviter les problèmes de confidentialité souvent liés aux données brutes ou moins traitées. Savoir que l'on peut faire converger ces valeurs joue un rôle incitatif vers la diffusion, favorise les bonnes pratiques et facilite donc les procédures de dépôt.

En somme, les conditions mises en place par IKER avec la création des divers outils de valorisation de la recherche au sein du laboratoire démontrent son engagement avec le libre accès. Sa communauté scientifique est en effet imprégnée de cette culture de partage et perçoit ANPERSANA comme un outil apportant une vraie valeur ajoutée à leurs travaux. Les résultats de nos enquêtes montrent que les perspectives d'évolution dans ce sens et l'enrichissement de la bibliothèque numérique avec de nouvelles collections passent par la poursuite de cette implication de la part de la structure et de la communauté scientifique. La réflexion sur les services offerts et sur les possibilités d'amélioration, ainsi que l'appropriation des pratiques et outils qui contribuent à la prise de conscience de ses droits sur ces données, sont les principaux éléments incitatifs à l'ouverture des données de la recherche chez IKER.

BIBLIOGRAPHIE

Généralités

« Berlin Declaration on Open Access to Knowledge », Berlin, 22 octobre 2003. [En ligne : https://openaccess.mpg.de/67605/berlin_declaration_engl.pdf]. Consulté le 22 avril 2017.

BORGMAN, Christine L., « The conundrum of sharing research data. », *Journal of the ASIST*, 63, 2012, p. 1059-1078. [En ligne : http://onlinelibrary.wiley.com/doi/10.1002/asi.22634/epdf]. Consulté le 30 avril 2017.

BOWKER, Geoffrey C. 2013. « Data Flakes : An Afterword to 'Raw Data' is an Oxymoron », in *"Raw Data" is an Oxymoron*, éd. Lisa Gitelman, 167–171. Cambridge, Massachussets, 2013. [En ligne : http://raley.english.ucsb.edu/wp-content/Engl800/RawData-excerpts.pdf]. Consulté le 23 avril 2017.

« Budapest Open Access Initiative », Budapest, 14 février 2002. [En ligne : http://www.budapestopenaccessinitiative.org/read]. Consulté le 22 avril 2017.

CABRERA, Francisca, « Les données de la recherche en Sciences humaines et sociales : enjeux et pratiques. Enquête exploratoire », Mémoire, Information et documentation, CNAM, soutenu le 10 décembre 2014. [En ligne : https://memsic.ccsd.cnrs.fr/mem_01117375/document]. Consulté le 26 mars 2017

DELAY-ARTOUS, Cécile, « Open, or not Open, Research Data », Mémoire, Information et documentation, CNAM, soutenu le 10 décembre 2014. [En ligne : https://memsic.ccsd.cnrs.fr/mem_01128833/document]. Consulté le 30 avril 2017.

DELAY-ARTOUS, Cécile, « Où sont les données de la recherche ; essai de cartographie », Présentation au colloque ISKO France, 2015. [En ligne : isko-france.asso.fr/isko2015/presentation2015/Delay-Artous.pptx]. Consulté le 2 juin 2017.

PAIN, Marilou, « Les données de la recherche et leurs entrepôts, de la documentation à la réutilisation : l'étude de cas pour l'archive HAL », Mémoire de stage, Information et communication, Enssib, Université de Lyon, soutenu en septembre 2016. [En ligne : https://memsic.ccsd.cnrs.fr/mem_01374509/document]. Consulté le 30 avril 2017.

ROSS-HELLAUER, Tony, DEPPE, Arvid et SCHMID, Birgit, « OpenAIRE Survey on Open Peer Review Attitudes and experience amongst editors, authors and reviewers », [preprint], 2017. [En ligne : https://zenodo.org/record/570864]. Consulté le 27 mai 2017.

Politiques et modèles

ABES, Association bibliographique de l'enseignement supérieur, « Politique nationale de l'IST : des infrastructures en cohérence », *Ar(abes)ques*, 84, Montpellier, février-mars

2017. [En ligne : www.abes.fr/content/download/3671/15314/version/1/file/Arabesques-84-web.pdf]. Consulté le 5 avril 2017.

BOURDONCLE, François et HERMELIN, Paul, « La nouvelle France industrielle : big data, feuille de route », 2014. [En ligne : http://www.economie.gouv.fr/files/files/PDF/Feuille-de-route_big-data151214.pdf]. Consulté le 22 avril 2017.

CHARTRON, Ghislaine, « Stratégie, politique et reformulation de l'*open access* », *Revue française des sciences de l'information et de la communication*, 8, 2016. [En ligne : https://rfsic.revues.org/1836]. Consulté le 12 mars 2017.

GAILLARD, Rémi. « De l'*Open data* à l'*Open research data* : quelle(s) politique(s) pour les données de recherche ? », Mémoire d'étude, Information et documentation, Enssib, soutenu en janvier 2014. [En ligne : http://www.enssib.fr/bibliotheque-numerique/documents/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche.pdf]. Consulté le 26 mars 2017.

OECD, Organisation for Economic Co-operation and Development, « Knowledge-based economy », Paris, 1996. [En ligne : https://www.oecd.org/sti/sci-tech/1913021.pdf]. Consulté le 22mars 2017.

OECD, « OECD principles and guidelines for access to research data », [sans lieu], 2007. [En ligne : https://www.oecd.org/sti/sci-tech/38500813.pdf]. Consulté le 22 avril 2017.

Aspects techniques des données

BRENNAN, Sheila, « Forums : Omeka multilanguage sites », Items and Archives, OMEKA, 2013. [En ligne : http://omeka.org/forums-legacy/topic/omeka-multilanguage-site#post-54227]. Consulté le 25 avril 2017.

Calliope team, « Calliope Text Mining Software, "fouille de textes" et Analyse de Tendances autour du logiciel Calliope », Paris Ouest et CNRS, [sans date]. [En ligne : https://www.calliope-textmining.com/textmining-fr.html]. Consulté le 20 mai 2017.

HumaNum, « Guide de bonnes pratiques », 2015. [En ligne : http://www.huma-num.fr/ressources/guide-des-bonnes-pratiques-numeriques]. Consulté le 9 juin 2017.

JACOBSON, Michel, « Domestiquez vos données sources », formation du 16/05/2017 au 17/05/2017 à l'Université de Bordeaux, UrfIST de Bordeaux.

Cadre juridique des données

BSN10, « Ouverture des données de la recherche – Guide d'analyse du cadre juridique en France », version1, 2016. [En ligne : http://www.bibliothequescientifiquenumerique.fr/wp-

content/uploads/2017/01/Guide_analyse_Cadre_Juridique_Ouverture_donness_Recherche_V1.pdf]. Consulté le 26 avril 2017.

Directive 2013/37/UE du Parlement européen et du Conseil du 26 juin 2013 modifiant la directive 2003/98/CE (Directive PSI). [En ligne : http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:FULL:FR:PDF]. Consulté le 17 mai 2017.

Loi n° 78-753 du 17 juillet 1978 (Loi CADA)portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal, dernière version 9 octobre 2016. [En ligne : https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000339241]. Consulté le 17 mai 2017.

Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique (Loi LEMAIRE). [En ligne : https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&dateTexte=20170621]. Consulté le 17 mai 2017.

MAUREL, Lionel (Calimaq), « Quel statut pour les données de la recherche après la loi numérique ? », S.I.Lex, 2016. [En ligne : https://scinfolex.com/2016/11/03/quel-statut-pour-les-donnees-de-la-recherche-apres-la-loi-numerique/]. Consulté le 27 mai 2017.

SKRZYPNIAK, Hélène, « Journée d'études sur les "Données de recherche" », Atelier juridique, IUT STID, Pau, 13 avril 2017.

IKER, ANPERSANA et données linguistiques

CRAWFORD W.J., *Doing corpus linguistics*, Routledge, [sans lieu], 2015.

JOUITTEAU, Mélanie, « La linguistique comme science ouverte », *Lapurdum*, 16, 2012, p. 93-115. [En ligne : http://lapurdum.revues.org/2357]. Consulté le 25 avril 2017.

PADILLA-MOYANO, Manuel, « Le Dauphin itsasontziko gutuneria (1757) edo euskararen historia behetik », *Lapurdum*, Numéro spécial 2, 2015, pp 15 [En ligne : http://lapurdum.revues.org/2553]. Consulté le 05 juin 2017.

TALEC, Jean-Philippe, « Apport d'un centre de documentation dans la vie d'un laboratoire », *Lapurdum*, 2012. [En ligne : https://lapurdum.revues.org/2367]. Consulté le mai 2017.

VIDEGAIN, Charles, LAMIKIZ, Xabier, DENIS, Gwendal, ARCOCHA-SCARCIA, Aurélie, TALEC, Jean-Philippe, « La correspondance en langue basque du Dauphin, 1757. », Maison des Sciences de l'Homme Aquitaine, MSHA, 2014. [En ligne : https://www.msha.fr/msha/quinquenal2016-2020.php?menu=5]. Consulté le 20 mai 2017.

LISTE DE TABLEAUX

LISTE DE GRAPHIQUES

ANNEXES

ANNEXE 1 : Modèle de PGD d'IKER (version réduite) – IKER CNRS 5478

Organisation	CNRS IKER (UMR 5478)	
Adresse		
Nature		
Coordinateur du PGD		
Rédaction		
Diffusion		
Titre complet projet		
Date début projet		
Date fin projet		
Début rédaction PDG		
Version		
Référence		
Site Web		
Version	Modification	Date
01	création document	
02	mise à jour après obtention du projet	
03	correction interne	
04	version actuelle	
Participants	Affiliation	Contact

1. Informations sur le projet	
1. Informations on the project [en] / 1. Proiektuari buruzko informazioa [eus]	
Cette section a pour vocation de renseigner administrativement sur le projet auquel ce PGD est lié, de le présenter et de le décrire succinctement.	
This section aims to provide administrative information about the project to which this DMP is linked and to present and briefly describe it.	
Atal honen helburua DKP honekin erlazionatutako proiektuari buruzko informazio administratiboa eskaintzea da eta baita, proiektua aurkeztu eta laburki deskribatzea.	
Référence de la convention de financement / *Grant agreement number* / *Finantzaketa-akordioaren erreferentzia*	
Programme de recherche / *Research program* / *Ikerketa programa*	
Acronyme du projet / *Project acronym* / *Proiektuaren akronimoa*	

Titre du projet / *Project title* / *Proiektuaren izenburua*	
Objectifs du projet / *Goals of the project* / *Proiektuaren helburuak*	
Mots-clefs du projet / *Keywords* / *Hitz-gakoak*	
Coordinateur-Bénéficiaire / *Coordinator-Recipient* / *Koordinatzailea-Hartzailea*	
Responsable scientifique du projet / *Project leader* / *Proiektuaren arduradun zientifikoa*	
Affiliation et unité de rattachement du responsable scientifique / *Administrative affiliation of the project leader* / *Arduradun zientifikoaren afiliazioa eta lotura administratiboa*	
Document référence / *Reference document* / Erreferentzia-dokumentua	

2. Responsabilité des données

2. Responsibility for the data [en] / 2. Datuen errespontsabilitatea [eus]

Cette section vise à identifier la ou les personne(s) qui seront en charge de la mise en œuvre et de la mise à jour du PGD tout au long du projet, ainsi que la propriété intellectuelle des données liées au projet de recherche.

This section provides information about the person(s) who will be responsible for implementing and updating the DMP throughout the project, and the intellectual property of them.

Atal honek, DKP abiatu eta, proiektuak iraugo duen artean, DKP eguneratuko duen arduraduna, edo arduradunak, izendatzen ditu. Aldi berean, ikerketa proiektuari lotutako datuen jabetza intelektuala ezarriko da.

Nom du responsable de la gestion des données au cours du projet de recherche / *Name of the person in charge of data management during the project* / *Proiektu osoan zehar datuen kudeaketaz arduratuko den pertsonaren izena*	
Propriété des données / *Data property*/ *Datuen jabetza*	

3. Ressources nécessaires à la mise en œuvre du PGD

3. Resources needed to implement the DMP [en] / 3. DKP abiatzeko behar diren baliabideak [eus]

Estimez les compétences nécessaires à la mise en œuvre du PGD : gestion, curation (sélection, nettoyage, normalisation et enrichissement des données), conservation à long terme et les coûts associés.

Estimate the skills needed to implement the DMP: management, curation (selection, cleaning, normalization and data enrichment), long-term conservation and associated costs.

Estima itzazu DKP abiatzeko behar diren gaitasunak : kudeaketa, kurazioa (datuen aukeratze, garbitze, normalizatze et aberastea), epe-luzeko kontserbazioa eta lotutako kostuak.

Matériel / *Hardware* / *Materiala (euskarri fisikoak)*	
Personnel / *Staff* / *Lan-taldea*	
Formations / *Training* / *Formakuntza*	
Montant financier / *Costs* / *Kostu ekonomikoa*	

4. Jeux de données

4. Datasets [en] / 4. Datu-multzoak [eus]

Selon le projet, un jeu de données peut recouvrir des réalités différentes. En effet, un ou plusieurs jeu(x) de données peuvent être lié(s) au projet de recherche, et désigner : i) un lot techniquement homogène, ou ii) un lot intellectuellement cohérent même si celui-ci est composé de lots techniquement hétérogènes.

Depending of the project, a dataset may cover different realities. In fact, a unique or multiple datasets may be linked to the research project and represent a technically uniform batch of data or an intellectually consistent one, potentially made of technically heterogeneous samples.

Datu-multzo batek biltzen dituen errealitateak, proiektuaren araberakoak dira. Alegia, datu-multzo bat edo gehiago erlazionatu dakizkioke ikerketa proiektuari. Gainera, i) datu-multzoek bilduma teknikoki homogeneoak osa ditzakete ii) edo baita teknikoki heterogeneoak direnak, nahiz eta intelektualki koherenteak izan.

Nombre de jeu(x) de données/ *Number of datasets* / *Datu-multzo kopurua*	

4.1. Description des données – Jeu de données

4.1. Data description – Dataset [en] / 4.1. Datuen deskribapena – Datu-multzoak [eus]

Cette section a pour vocation de présenter le jeu de données qui sera produit et/ou reçu dans le cadre du projet.

This section aims to generally present the type of data that will be produced and/or received in connection with the project.

Atal honetan proiektuaren mugetan ekoitzi edota bilduko den datu-multzoa aurkeztuko da.

Identifiant et nom du jeu de données / *Reference and name of data set* / *Datu-multzoaren identifikatzailea eta izena*	
Nature des données / *Nature of data* / *Datuen izaera edo datu motak*	
Réutilisation de données existantes/ *Reuse of existing data* / *Egun dauden datuen berrerabilera*	
Méthode de production des données/ *Method of production of data* / *Datuen ekoizte metodoa*	
Formats des données / *Data standard* / Datuen formatuak	

4.2. Au cours du projet - Stockage, accès et sécurité des données

4.2. During the project - Storage, access and security [en] / 4.2. Proiektu osoan zehar – Datuen biltzea, atzigarritasuna eta segurantza [eus]

Cette section définit les modalités d'hébergement, de sauvegarde et d'accès aux données pendant la phase active du projet.

This section defines hosting, backup and data access during the active phase of the project.

Atal honek datuen hosting-a, segurtasun kopiak eta atzigarritasuna definitzen ditu, proiektuaren fase aktibok irauten duen artean.

Stockage et enregistrement des données / *Storage and recording of data / Datuen biltze eta erregistratzea*

Support des données / *Medium of data* / *Datuen euskarria*	
Volumétrie prévisionnelle / *Projected volume* / *Aurreikusitako bolumena*	
Stockage et enregistrement des données / *Storage and recording / Datuen biltze eta erregistratzea*	
Type d'hébergement / *Data hosting* / *Datuen hosting-a*	
Sécurité des données / *Data security*	
Risques ou menaces sur les données / *Risks or threats to data / Datuekiko arrisku edo mehatxuak*	
Garantie de confidentialité des données / *Data privacy* / *Datuen konfidentzialtasun bermea*	
Garantie d'intégrité et de traçabilité / *Data integrity and traceability*/ Datuen integritate eta trazabilitate bermea	
Accès aux données / *Access to data*	
Lecture des données / *Data reading* / *Datuen irakurketa*	
Garantie de disponibilité des données / *Data availability/ Datuen eskuragarritasun bermea*	
Gestion des accès / *Access procedures* / *Sarbideen kudeaketa*	
Echanges et partage / *Data sharing* / *Datu trukatze eta partekatzea*	

4.3. Métadonnées : documentation et organisation des données

4.3. Metadata: documentation and data organization [en] / 4.3. Metadatuak: datuen dokumentatzea eta antolakuntza [eus]

Cette section précise la manière dont seront décrites et organisées les données produites ou reçues au cours du projet.

This section details how the data generated or received during the project will be described and organized.

Atal honetan proiektua gauzatu ahala jaso edo ekoitzitako datuak nola deskribatu eta antolatu zehazten da.

Standards et formats disciplinaires des métadonnées / *Standards and metadata* / *Estandarrak eta metadatuak*	
Mode de production et responsabilité des métadonnées / *Method of production and metadata responsibility* / *Metadatuen ekoizmolde eta erantzukizuna*	
Arborescence de classement / *Tree classification* / *Sailkapenaren antolakuntza*	
Règles de nommage des jeux de données / *Rules for naming data sets* / *Datuak izendatzeko arauak*	
Documentation associée / *Relevant documentation* / *Lotutako dokumentazioa*	

4.4. A l'issue du projet – Dissémination

4.4. At the end of the project – Dissemination [en] / 4.4. Proiektua bukatzerakoan – Hedapena [eus]

Cette section précise la manière dont seront diffusées les données après le projet.
This section details how the data generated or received during the project will be disseminated after the project.
Atal honetan proiektua bukatzerakoan datuak nola hedatuko diren zehazten da.
Partage, diffusion et réutilisation des données / *Data sharing, diffusion and reuse* / *Datuen partekatzea, hedapena eta berrerabilera*
Cette section précise les modalités et les éventuelles précautions éthiques, juridiques et techniques selon lesquelles seront diffusées les données.
This section describes the procedures and specifies any ethical, legal and technical safeguards under which the data will be released.
Atal honetan datuak hedatzeko prozedurak eta hedatu aurretik kontuan hartu beharreko auzi etiko, juridiko eta teknikoak zehazten dira.

Principe général de diffusion / *General principle of diffusion* / *Hedapenaren printzipio orokorra*	
Type de licence / *Type of license/ Lizentzia mota*	
Potentiel de réutilisation / Potentiel for reuse / Izan dezaketen berrerabilera	
Existence de publications associées aux données / Existing publications related to the data / Datu hauen harira egun dauden argitalpenak	
Dépôt et dissémination des données / Data repository and access / Datuen gordailua eta datuen hedapena	

Protection des données sensibles / *Protection of sensitive data*
Pour diverses raisons éthiques, juridiques, financières ou encore techniques, certaines données peuvent nécessiter une protection spécifique et à ce titre échapper aux principes de diffusion. Cette section a donc pour but d'identifier et de définir les critères de protection des données sensibles susceptibles d'être produites ou collectées dans le cadre du projet.
For a variety of ethical, legal, financial or technical reasons, data may require specific protection and as such, escape the rules of distribution. This section identifies and defines criteria for protection of sensitive data that can be produced or collected as part of the project.
Arrazoi etiko, juridiko, ekonomiko edota tekniko ezberdinak direla eta, zenbait datuk babes berezi bat eska dezakete eta hala, hedapen arauetatik at gelditu. Atal honetan, beraz, proiektuan ekoitzi edo bildu litezkeen datu sentsibleak babesteko printzipioak definitzen dira.

Identification des jeux de données sensibles / *Identification of sensitive data sets* / *Datu sentsibleen identifikazioa*	
Justification du principe d'exception aux conditions générales de diffusion / *Justification for the exception to the general principles of diffusion* / Hedapenaren arau orokorren *salbuespenerako justifikazioa*	
Mesures de protection / *Precautionary mesures* / *Babeserako neurriak*	
Embargo / *Embargo period* / *Enbargo epea*	

5. Sélection et archivage des données

5. Data selection and long term preservation [en] / 5. Datuen hautaketa eta artxibaketa [eus]

Cette section concerne l'ensemble des données produites ou collectées au cours du projet, qu'elles aient été diffusées ou non. Il est fortement recommandé de prendre contact avec l'archiviste de votre établissement lors de la rédaction de cette section.

This section applies to all data generated or collected during the project, whether distributed or not. It is strongly recommended to contact the archivist of your establishment to draft this section.

Atal hau proiektuan ekoitzi eta bildutako datu guztiei dagokie, hedatuak izan direnentz kontuan hartu gabe. Beraz, ondoko atala zure-zuen erakundeko artxibozainarekin batera osa dezazun gomendatzen da.

Sort des données à l'issue du projet / *Fate of data at the end of the project* / *Zer izango den datuez proiektua bukatzean*	
Sélection des données / *Data selection* / *Datuen hautaketa*	
Volume final des données / *Final volume of data* / *Datuen bolumena bukaeran*	
Durées de conservation préconisée / *Recommended lifetime* / *Gomendatutako kontserbazio iraupena*	
Plateforme d'archivage / *Long term preservation plateform* / *Artxibaketa plataforma*	

ANNEXE 2 : Guide de saisie de métadonnées DC sur ANPERSANA

Conçu pour ANPERSANA telle qu'elle est aujourd'hui, le 02 juin 2017 ; à actualiser systématiquement si les vocabulaires contrôlés, les éléments et les attributs sont modifiés ou suite à toute autre action qui suppose un changement des informations à saisir.

L'objectif de ce guide est d'accompagner le producteur de données dans la procédure de dépôt des données sur ANPERSANA avec le logiciel OMEKA. Concrètement, c'est un recueil de recommandations pour la saisie des métadonnées associées aux données et jeux des données issues des différents projets de recherche du laboratoire IKER. Les préconisations ont été conçues de sorte que les métadonnées soient compatibles avec les éléments de description de métadonnées Dublin Core.

Chaque champ à renseigner est listé et décrit ci-dessous dans le même ordre d'apparition que sur votre espace privé ANPERSANA - OMEKA. Le jeu de champs des métadonnées Dublin Core est commun à tous les enregistrements, que ce soit des contenus, des fichiers ou des collections.

Si vous avez un Plan de Gestion de Données (PGD), le modèle d'IKER, vous pouvez utiliser également le tableau des correspondances que nous avons élaboré pour vous accompagner dans le transfert d'information entre votre PGD et Dublin Core.

1. TITRE	
Étiquette (anglais)	dc.Title
Définition	Ici on écrit l'intitulé du jeu des données.
Exemple	Lettre de Martin de Arrunde [père] à Martin de Arrunde [fils] Extraits du vocabulaire de …
Commentaire	Le choix de l'intitulé doit représenter le jeu de données décrit. Typiquement, le nom par lequel la ressource est officiellement connue.
Codage	Texte libre
Statut	Obligatoire

1.1 AUTRE FORME DE TITRE	
Étiquette (anglais)	dc.AlternativeTitle
Définition	Traduction du titre.
Exemple	
Commentaire	Traduire le titre au basque OU au français. Traduire au basque ET au français, si « 1. TITRE » en une autre langue que le basque ou le français.
Codage	Texte libre
Statut	Obligatoire

2 CREATEUR	
Étiquette (anglais)	dc.Creator
Définition	Entité (personne ou organisme) responsable de la création du contenu du document.
Exemple	Zuazo, Koldo
Commentaire	Le nom et prénom doivent apparaître sous leur forme développée. Le format de saisie à utiliser, déjà formaté dans la liste déroulante, est le suivant : « Nom(s), Prénom », comme dans l'exemple. Si vous avez créé un corpus linguistique, vous renseignez votre nom et prénom. Pareillement, si vous avez effectué un entretien ou si vous enregistrez un cours dans une classe, c'est vous le créateur de ces enregistrements (même s'il y a d'autres participants et vous n'avez pas le droit d'auteur). Pourtant, si vous avez collecté des manuscrits, images, sons, ou toute autre type de document pour vos recherches, qui n'a pas été produit par vous, c'est le responsable de la création qu'il faut renseigner (souvent l'auteur).
Codage	Liste déroulante fermée : choisir parmi les options de la liste ; si le créateur n'apparaît pas sur la liste, consultez le documentaliste pour l'inclure.
Statut	Obligatoire

3 CONTRIBUTEUR	
Étiquette (anglais)	dc.Contributor
Définition	Entité (personne ou organisme) ayant contribué à la création du contenu ou la forme du document ; tous ceux qui ont participé à la production du jeu des données, sans être les auteurs principaux, sont des contributeurs.
Exemple	Smith, John (transcriptions) Société Xxxx (photographie) (numérisation) La Maison des Sciences de l'Homme d'Aquitaine appuie ce travail dans le cadre de la quinquennale recherche 2016-2020.
Commentaire	Le nom doit apparaître sous sa forme développée. Le format de saisie à utiliser normalement est le suivant : « Nom, Prénom » suivis du type de contribution entre parenthèses. On peut aussi insérer une phrase pour expliquer la contribution.
Codage	Si la liste déroulante fermée est active : choisir contributeur et type de contribution. Si la liste déroulante n'est pas active, suivre les instructions sous Commentaire et Exemple.
Statut	Obligatoire si applicable

4 ÉDITEUR	
Étiquette (anglais)	dc.Publisher
Définition	Entité (personne ou organisme) responsable de de la mise à disposition du document
Exemple	CNRS IKER (UMR 5478)
Commentaire	En tant que gestionnaire d'ANPERSANA, CNRS IKER (UMR 5478) sera normalement l'éditeur.
Liste déroulante fermée	Liste déroulante fermée ; si l'éditeur n'apparaît pas sur la liste, consultez le documentaliste pour l'inclure.
Statut	Obligatoire

5 DATE	
Étiquette (anglais)	dc.Date
Définition	Un point ou une période dans le temps associés à un événement dans le cycle de vie de la ressource
Exemple	Année: AAAA (1997) Année et mois: MM.AAAA (07.2014) Date complète: JJ. MM.AAAA (16.07.2014) Date complète avec heures et minutes: JJ. MM.AAAA Thh:mmTZD (16.07.2014T19:20+01:00) Entre deux dates : 1990/1993
Codage	Vocabulaire contrôlé - format des exemples
Commentaire	Normalement on saisit ici la date de création des originaux, avant la numérisation. Pourtant, il y a l'option de préciser des dates importantes supplémentaires dans les sous-éléments de 5 DATE (de 5.1 à 5.8) Utiliser le format de date des exemples. (Le codage risque d'évoluer dans l'avenir pour être conforme avec la norme ISO 8601 ou les recommandations *Date and Time Formats* du W3C.) La date peut être un temps défini entre deux dates. La date la plus précise est la date préférable : 1824/1896 est préférable à 19e siècle.
Statut	Obligatoire

5.1 DATE DE DISPONIBILITE	
Étiquette (anglais)	dc.DateAvailable
Définition	Date à laquelle la ressource décrite sera disponible en ligne et qui aura préalablement été définie en coordination avec le documentaliste.
Exemple	
Commentaire	
Codage	Vocabulaire contrôlé - format des dates des exemples de 5 DATE.
Statut	Obligatoire

5.2 DATE DE CREATION	
Étiquette (anglais)	dc.Creation
Définition	Date à laquelle le ou les fichiers de la ressource sont téléchargés à la plateforme (l'étape précédant à la publication sur ANPERSANA)
Exemple	
Commentaire	
Codage	Vocabulaire contrôlé - format des dates des exemples de 5 DATE.
Statut	Obligatoire

5.3 DATE D'ACCEPTATION	
Étiquette (anglais)	dc.AcceptanceDate
Définition	Date d'acceptation de la ressource.
Exemple	
Commentaire	On va rarement saisir la date d'acceptation sur ANPERSANA, car les ressources pour lesquelles la date d'acceptation peut être importante incluent les thèses (acceptées par le département d'une université) et les articles scientifiques (acceptés par un journal). Ces types de documents ne sont pas des données de la recherche couramment produites par la communauté scientifique d'IKER.
Codage	Vocabulaire contrôlé - format des dates des exemples de 5 DATE.
Statut	Obligatoire si applicable

5.4 DATE DU COPYRIGHT/DROIT D'AUTEUR	
Étiquette (anglais)	dc.CopyrightDate
Définition	Date à partir de laquelle vous avez les droits d'auteur des données. Le jour où vous avez assigné une licence quelconque à vos données.
Exemple	
Commentaire	Le concept *copyright* n'existe pas en France, même si le terme est souvent utilisé. Le droit d'auteur est le concept qui se rapproche le plus du concept désigné par la terminologie anglo-saxonne.
Codage	Vocabulaire contrôlé - format des dates des exemples de 5 DATE.
Statut	Optionnel, recommandé

5.5 DATE DE SOUMISSION	
Étiquette (anglais)	dc.SubmissionDate
Définition	La première date à laquelle le ou les fichiers de la ressource sont téléchargés à la bibliothèque numérique ANPERSANA, sans qu'ils soient forcement publiés en ligne (voir 5.1 DATE DISPONIBILITE).
Exemple	
Commentaire	On va rarement saisir la date de soumission sur ANPERSANA, car les ressources pour lesquelles la date de soumission peut être importante incluent les thèses (soumisses au département d'une université) et les articles scientifiques (soumis à un journal).
Codage	Vocabulaire contrôlé - format des dates des exemples de 5 DATE.
Statut	Obligatoire si applicable

5.6 DATE DE PARUTION	
Étiquette (anglais)	dc.FormalIssuance
Définition	Date d'apparition des données dans une publication.
Exemple	
Commentaire	
Codage	Vocabulaire contrôlé - format des dates des exemples de 5 DATE.
Statut	Recommandé si applicable

5.7 DATE DE MODIFICATION	
Étiquette (anglais)	dc.ModificationDate
Définition	Date à laquelle la ressource a été modifiée.
Exemple	
Commentaire	Ne pas renseigner si la ressource est publiée sur ANPERSANA pour la première fois. D'ailleurs, ce champ ne peut concerner que les modifications effectuées ultérieurement.
Codage	Vocabulaire contrôlé - format des dates des exemples de 5 DATE.
Statut	Obligatoire si applicable

5.8 DATE DE VALIDITE	
Étiquette (anglais)	dc.ValidityDate
Définition	Date ou période pendant laquelle la ressource est valable.
Exemple	
Commentaire	Applicable seulement si les données sont susceptibles de périmer ou devenir obsolètes, ou si elles doivent être retirées à une date connue pour une raison quelconque.
Codage	Vocabulaire contrôlé - format des dates des exemples de 5 DATE.
Statut	Obligatoire si applicable

6 TYPE	
Étiquette (anglais)	dc.Type
Définition	Le « type » est la dénomination qui décrit le document contenant les données ou le jeu des données. Cette dénomination renvoie plus ou moins précisément à la nature des données : la source (producteur, mode de production ou contexte de production), le genre (littéraire, historique, journalistique, etc.) et/ou la destinée (destiné à être vu, lu, entendu).
Exemple	- collection - jeu de données (pour les tableaux ou bases de données) - évènement - image - ressource interactive - vidéo - objet - software - enregistrement sonore - manuscrits - ressources textuelles - musique notée - musique notée manuscrite - ressource cartographique - ressource cartographique manuscrite
Commentaire	
Codage	Il est recommandé de choisir parmi les termes signalés dans l'exemple. Pour répondre aux besoins d'ANPERSANA, nous avons fait une sélection de types de documents en combinant certains des termes définis par la DCMI (traduits) et certains des recommandés par la BnF.
Statut	Obligatoire

7 FORMAT	
Étiquette (anglais)	dc.Format
Définition	Le format est le type de fichier numérique contenant toute l'information sur les données. Quand on nomme un fichier, ce nom est suivi d'une extension. Cette extension décrit comment l'information est organisée dans le fichier et permet de choisir l'application qui va interpréter l'information et la rendre compréhensible aux humains.
Exemple	EXTENSION — FORMAT .pdf — PDF .jpg — JPEG .avi — AVI .wmv — WMV .wav — WAV .mp3 — MP3 .png — PNG .csv — CSV .xml — XML .sql — SQL
Commentaire	
Codage	Liste déroulante fermée. (La liste est susceptible d'évoluer visant plus de précision. Par exemple : PDF/A, PDF/X, etc. au lieu de PDF) La liste des formats reprend les termes du référentiel MIME. Si le format du fichier n'est pas sur la liste, consulter le documentaliste.
Statut	Obligatoire

7.1 ÉTENDUE DE LA RESSOURCE, TAILLE, DUREE	
Étiquette (anglais)	dc.SizeorDuration
Définition	L'information concernant l'étendue, la taille et la durée de la ressource sont à renseigner ici.
Exemple	« Enquêtes de terrain effectuées au Pays Basque et au Béarn par Marie Hirigoyen Bidart de 2003 à 2012 » : **47:21:32 ; 44 contenus (documents)** « Entretien avec Ximun Haran, le 21 janvier 2012 » : **01:28:19** « Lettre de Francha Barrere à Pierre Hayete » : **2 fichiers (fac-similé 7 p., 7667 Ko et transcription 5 p., 330 Ko)**
Commentaire	Les spécificités à saisir dépendront du type de document et du format de fichier. Par exemple, on se réfère à l'étendue d'un manuscrit en nombre de pages, mais pour l'étendue d'une base des données ce sont les tables, les lignes, les colonnes, les éléments, etc. qu'on prend en compte. Egalement, pour les enregistrements on parle de durée plutôt que d'étendue. La limite de taille totale des fichiers par collection est de 100 Go. Des quotas plus élevés peuvent être demandés et accordés au cas par cas.
Codage	Texte libre.
Statut	Obligatoire

7.2 SUPPORT	
Étiquette (anglais)	dc.Medium
Définition	Le support est le matériel portant la ressource.
Exemple	Papier, argile, parchemin, huile sur toile, cd, dvd, casettes, bois, pierre, etc.
Commentaire	Ne pas renseigner les documents d'origine numérique, car ils n'ont pas de support à décrire. Ils peuvent être transportés d'un support à un autre et les fichiers sont invisibles aux humains sans une machine qui les interpète.
Codage	Texte libre. Avant la saisie, on recommande la vérification des termes sur TermScience.
Statut	Recommandé si applicable

8 LANGUE	
Étiquette (anglais)	dc.Language
Définition	La langue du contenu intellectuel de la ressource.
Exemple	eus = basque fra = français esp = espagnol eng = anglais
Commentaire	S'il y a plusieurs langues, on inclût toutes, et s'il n'y a pas de langue utilisée (comme peut être le cas de certaines images ou vidéos) on ne saisit pas ce champ.
Codage	Pour le moment, on choisit la langue de la liste fermée. Dans l'avenir, l'inclusion des codes à 3 caractères d'ISO 639-2 est prévue, comme dans l'exemple. Ces valeurs sont disponibles sur le catalogue Ethnologue, disponible sur : www.ethnologue.com
Statut	Obligatoire si applicable

9 SOURCE	
Étiquette (anglais)	dc.Source
Définition	Une ressource liée de laquelle dérive la ressource décrite ; information sur l'origine.
Exemple	« Parmi les 3 exemplaires connus, celui qui se trouve dans les Archives de X, URL : X » « Enregistrement réalisé à l'aide d'un microphone X et un téléphone X » « Base de données élaboré avec logiciel X et information extraite avec l'application X »
Commentaire	L'information doit être succincte mais suffisamment complète. Attention : à ne pas confondre avec « Provenance ». La provenance se réfère à l'appartenance et à la garde de la ressource, tandis que la source fait référence à l'objet même de la ressource d'origine, précédant la numérisation (voir « 16 PROVENANCE »).
Codage	Texte libre. Ajouter la date associée, si connue, à chaque entrée de source : « le JJ.MM.AAAA »
Statut	Obligatoire

10 SUJET, MOTS CLES	
Étiquette (anglais)	dc.Subject
Définition	Les sujets de la ressource.
Exemple	Pour le titre *Guero*, les mots clés pourrait être : Axular, Pedro (1556-1664) christianisme ascétisme littérature basque (XVIe siècle) langue basque (XVIe siècle)
Commentaire	Le sujet sera représenté à l'aide de mots clés, phrases clés ou de codes de classification. On peut s'aider des mots clés utilisés pour une ressource similaire dans le catalogue en ligne d'une bibliothèque reconnue.
Codage	Texte libre. On recommande l'utilisation des termes TermScience.
Statut	Obligatoire

11 DESCRIPTION	
Étiquette (anglais)	dc.Description
Définition	Une présentation du contenu de la ressource.
Exemple	
Commentaire	Sous 11 DESCRIPTION on peut inclure : un résumé, une table des matières, une référence à une représentation graphique du contenu ou un exposé du contenu.
Codage	Texte libre.
Statut	Parmi « 11 DESCRIPTION », « 11.1 RESUME » et « 11.2 TABLE DE MATIERES » remplir un élément minimum obligatoirement.

11.1 RESUME	
Étiquette (anglais)	dc.Resume
Définition	Une présentation du contenu de la ressource par le moyen d'un résumé ou d'un exposé du contenu en texte libre.
Exemple	*In a world in which lives are shaped by irrevocable choices and by fortuitous events, a world in which everything occurs but once, existence seems to lose its substance, its weight. Hence, we feel "the unbearable lightness of being" not only as the consequence of our pristine actions but also in the public sphere, and the two inevitably intertwine.*
Commentaire	
Codage	Texte libre.
Statut	Parmi « 11 DESCRIPTION », « 11.1 RESUME » et « 11.2 TABLE DE MATIERES » remplir un élément minimum obligatoirement.

11.2 TABLE DE MATIERES	
Étiquette (anglais)	dc.TableofContents
Définition	Une table des matières
Exemple	AURKIBIDEA 1 Emazteak literaturan 1.1 Hastapenak 1.2 Sarako eskola a) xxxxx b) xxxxx 1.3 Xxxxxxx 2.
Commentaire	
Codage	Texte libre.
Statut	Parmi « 11 DESCRIPTION », « 11.1 RESUME » et « 11.2 TABLE DE MATIERES » remplir un élément minimum obligatoirement.

12.1 COUVERTURE SPATIALE	
Étiquette (anglais)	dc.Description
Définition	Les caractéristiques spatiales de la ressource. La ou les localisations en lien avec la ressource (pas en lien avec le contenu).
Exemple	Allemagne Vénice Marrakech Nafarroako Erribera
Commentaire	
Statut	Obligatoire

12.2 COUVERTURE TEMPORELLE	
Étiquette (anglais)	dc.Description
Définition	Les caractéristiques temporelles de la ressource (pas du contenu).
Exemple	1990/1993
Commentaire	
Codage	Voir les remarques de l'élément 5 DATE pour la syntaxe à préférer.
Statut	Obligatoire

13 DROITS	
Étiquette (anglais)	dc.Rights
Définition	Information à propos des droits détenus
Exemple	« Les droits sont détenus par les auteurs de ces enregistrements, à savoir les collecteurs et les informateurs, qui autorisent IKER à publier en ligne les données dans la bibliothèque numérique ANPERSANA »
Commentaire	Les droits comprennent typiquement une déclaration à propos des divers droits de propriété associés à la ressource, y compris les droits de propriété intellectuelle.
Codage	Texte libre
Statut	Obligatoire

13.1 DROIT D'ACCES	
Étiquette (anglais)	dc.AccesRights
Définition	Information à propos de qui peut accéder à la ressource.
Exemple	« accès public » « période d'embargo »
Commentaire	Les fichiers peuvent être déposés sous un accès ouvert ou sous embargo (protégés jusqu'à la première publication des résultats de recherche). C'est-à-dire que les producteurs des données peuvent déposer des fichiers restreints qui ne seront pas rendus publics, si ainsi souhaité. Une fois les données publiées, il n'y aura pas de restriction d'accès.
Codage	Texte libre. Utiliser exemples, si applicable.
Statut	Obligatoire

13.2 LICENCE	
Étiquette (anglais)	dc.License
Définition	Un document légal donnant permission officielle de faire quelque chose avec la ressource.
Exemple	- *Creative Commons* paternité, pas de modification [CC] [BY] [ND] - Paternité - Pas d'utilisation commerciale - Partage selon les Conditions Initiales [CC] [BY] [NC] [SA]
Commentaire	Si vous avez des questions au moment de faire le choix de la licence la plus adéquate, consulter le documentaliste. Vous pouvez également consulter le site officiel de *Creative Commons*, URL : http://creativecommons.fr/licences/
Codage	Utiliser nom officiel de la licence
Statut	Obligatoire

14 RELATION	
Étiquette (anglais)	dc.Relation
Définition	Une ressource liée ou apparentée.
Exemple	« Smith, John, *L'aquisition du L2 Basque par les…*, Université de X, thèse de doctorat, linguistique, 2016 » « TNA, HCA 31/270 » (autres archives) « Lamikiz, Xabier « Le Dauphin itsasontziaren testuinguru historikoa : Louisbourgeko euskaldunak, ekonomia atlantiarra eta gerra (1713–1758) », Lapurdum [En ligne], Numéro spécial 2 | 2015, mis en ligne le 01 septembre 2016, consulté le 03 avril 2017. URL : http://lapurdum.revues.org/2546 »
Commentaire	Cet élément se compose de titres, noms d'auteurs, lieux, dates, liens, etc. Souvent ça sert à signaler un document au sujet des données décrites par le moyen d'un lien : vers un travail de recherche, article de presse ou autre.
Codage	Texte libre.
Statut	Recommandé si applicable

15 IDENTIFIANT	
Étiquette (anglais)	dc.Identifier
Définition	Code numérique ou alphanumérique unique associé à la ressource
Exemple	
Commentaire	
Codage	Code automatiquement produit lors du dépôt. Susceptible d'évoluer vers l'assignation d'un identifiant pérenne.
Statut	Obligatoire

16 PROVENANCE	
Étiquette (anglais)	dc.Provenance
Définition	Une déclaration de tous les changements d'appartenance et de garde de la ressource depuis sa création qui sont importants pour son authenticité, son intégrité et son interprétation.
Exemple	- fourni par « The National Archives, High Court Admiralty (Londres), Royaume-Uni » le xx.xx.xxxx - fourni par PRENOM NOM le xx.xx.xxxx - dérivé de NOM COLLECTION D'ORIGINE - curation mené par PRENOM NOM (de la personne qui a élaboré la ou les collections telles qu'elles seront présentées sur la plateforme ANPERSANA ; normalement le documentaliste ou le chercheur même). - extrait de « ELAR, Endangered Languages Archive », dépôt id 0167, dépôt ELDP id MDP0268, *Koryak Ethnopoetics: Stories from Herders and Maritime Villagers* disponible sur URL : [https://elar.soas.ac.uk/Collection/MPI603817#items], accessible le 02.02.2002 - importé de XML
Commentaire	Pour la saisie de ce champ voir « Exemple » et « Codage ». Attention : « Provenance » à ne pas confondre avec « Source ». La provenance se réfère à l'appartenance et à la garde de la ressource, tandis que la source fait référence à l'objet même de la ressource d'origine, précédant la numérisation (voir « 9 SOURCE »).
Codage	- Se référer à la provenance en utilisant l'une ou plusieurs des versions en français du vocabulaire suivant : [provided by] = « fourni(e) par » [retrieved from] = « extrait(e) de » [imported from] = « importé(e) de » (de quel format, si format de destination différent de celui d'origine. Par exemple, on importe l'audio de provenance, qui est format WAV, pour le convertir ensuite à MP3, et gagner ainsi de l'espace. [created by] = « créé(e) par » - Le vocabulaire que nous recommandons ici, nous l'avons extrait et traduit à partir des propriétés PAV Table 2 : *Table 2 PAV provenance properties*, disponible sur URL : [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4177195/table/T2/], accessible le 14/04/2017 - Quand on modifie une collection pour lui donner une structure intellectuelle différente à celle de la provenance, cela veut dire qu'elle a suivi un processus de curation. Nous pouvons le signaler de la façon suivante : [*derived from*] = dérivé de [*curated by*] = curation menée par - Vocabulaire extrait et traduit à partir de *Table 3 PAV versioning properties*, disponible sur URL : [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4177195/table/T3/], accessible le 14/04/2017 - Ajouter la date associée, si connue, à chaque entrée de provenance : « le JJ. MM.AAAA », « entre le JJ. MM.AAAA et le JJ. MM.AAAA »
Statut	Obligatoire

ANNEXE 3 : Table de correspondances PGD d'IKER et éléments DC

Correspondances Dublin Core > Plan de gestion de données					**Notes**
Eléments DUBLIN CORE (DC) utilisés sur ANPERSANA (Anp.) : **- 15 éléments de base** **- 1 élément supplémentaire (S)** **- 8 éléments de raffinement ou *qualifiers* (Q)**			**Sections du modèle de PGD d'IKER**		
Num. DC	**Ordre Anp.**	**Intitulé**	**Num. section**	**Intitulé de l'entrée**	
1	1	Titre	1 -------	Titre du projet	
Q	1.1	Autre forme de titre	*	(correspondances inexistantes)	Dans l'élément DC, traduire titre : - au basque (si titre du projet en français) - au français (si titre du projet en basque) - au basque et au français (si titre du projet en autres langues)
2	2	Créateur	1 -------	Responsable scientifique du projet	
6	3	Contributeur	0 ------- 1 ------- 2 -------	Participants Responsable scientifique du projet Nom du responsable de la gestion de données au cours du projet de recherche	
5	4	Éditeur	1 -------	Coordinateur bénéficiaire	
7	5	Date	0 ------- 4.1. ---- 4.4. ----	Diffusion / Date début / Date fin Nature des données Dissémination	correspondances partielles
Q	5.1	Date de disponibilité	0 ------- 4.1. ---- 4.4. ----	Diffusion / Date début / Date fin Nature des données Dissémination	correspondances partielles
Q	5.2	Date de création	0 ------- 4.1. ---- 4.4. ----	Diffusion / Date début / Date fin Nature des données Dissémination	correspondances partielles

Q	5.3	Date d'acceptation	0 ------- 4.1. ---- 4.4. ----	Diffusion / Date début / Date fin Nature des données Dissémination	correspondances partielles
Q	5.4	Date du copyright/droit d'auteur	0 ------- 4.1. ---- 4.4. ----	Diffusion / Date début / Date fin Nature des données Dissémination	correspondances partielles
Q	5.5	Date de soumission	0 ------- 4.1. ---- 4.4. ----	Diffusion / Date début / Date fin Nature des données Dissémination	correspondances partielles
Q	5.6	Date de parution	0 ------- 4.1. ---- 4.4. ----	Diffusion / Date début / Date fin Nature des données Dissémination	correspondances partielles
Q	5.7	Date de modification	0 ------- 4.1. ---- 4.4. ----	Diffusion / Date début / Date fin Nature des données Dissémination	correspondances partielles
Q	5.8	Date de validité	0 ------- 4.1. ---- 4.4. ----	Diffusion / Date début / Date fin Nature des données Dissémination	correspondances partielles
8	6	Type	4.1. ----	Nature des données	
9	7	Format	4.1. ----	Formats des données	Attention, l'information à renseigner sur DC ne concerne que les formats de publication en ligne.
Q	7.1	Étendue de la ressource, taille, durée	4.2. ----	Stockage et enregistrement des données / Volumétrie prévisionnelle	La volumétrie sera définitive suite à la publication de données. A ce moment là, l'information peut être actualisée sur le PGD.
Q	7.2	Support	4.2. ----	Stockage et enregistrement des données / Support des données	Attention, l'information à renseigner sur DC ne concerne que les formats physiques préalables à la publication en ligne.
12	8	Langue	4.1. ----	Nature des données	
11	9	Source	4.1. ---- 4.1. ----	Réutilisation de données existantes Méthode de production des données	
3	10	Sujet, mots clés	*	(correspondances inexistantes)	
4	11	Description	4.1. ----	Nature des données	

Q	11.1	Résumé	1 ------- 4.1. ----	Objectifs du projet Nature des données	
Q	11.2	Table de matières	*	(correspondances inexistantes)	
14	12.1	Couverture spatiale	4.1. ----	Nature des données	
14	12.2	Couverture temporelle	4.1. ----	Nature des données	
15	13	Droits	2. ------- 4.1. ---- 4.4. ----	Propriété des données Réutilisation des données Embargo	
Q	13.1	Droit d'accès	2. ------- 4.1. ---- 4.4. ---- 4.4. ---- 4.4. ---- 4.4. ----	Propriété des données Réutilisation des données Principe général de diffusion Potentiel de réutilisation Dépôt et dissémination des données Embargo	
Q	13.2	Licence	2. ------- 4.1. ---- 4.4. ----	Propriété des données Réutilisation des données Type de licence	
13	14	Relation	4.4. ----	Existence de publications associées aux données	
10	15	Identifiant	*	(correspondances inexistantes)	
S	16	Provenance	4.1. ----	Réutilisation des données	

ANNEXE 4 : Description des entités ajoutées à la cartographie de Delay-Artous

Catégories, quoi, qui, pourquoi	Accronyme Site web Sources	Présentation
France Entrepôt de données Qui le fait : IKER CNRS UMR 5478 Ce que ça fait : Diffusion des données scientifiques des études basques Pourquoi : Visibiliser le travail du laboratoire et contribuer aux études basques	Bibliothèque numérique ANPERSANA : https://anpersana.iker.univ-pau.fr/ Source complémentaire : www.iker.cnrs.fr/-entrepot-de-donnees-.html?lang=fr	Anpersana est une bibliothèque numérique recevant et diffusant des sources primaires de la recherche dans le domaine des études basques (manuscrits, carnets de recherche, images, sons, et autres documents multimédia). C'est une application informatique du centre de recherche sur la langue et les textes basques CNRS IKER (UMR 5478) résolument tournée vers l'accès libre à l'instar d'autres sites d'information du laboratoire, telles que l'archive ouverte Artxiker, la revue annuelle de l'unité Lapurdum et le catalogue de bibliothèque Kutxa. La sélection de ressources numériques est constituée de documents de première main utilisés dans le cadre d'un travail de recherche. Tous les documents sont orientés vers les études basques et les disciplines des sciences humaines et sociales (langue et culture d'expression basques). Les objectifs de la bibliothèque numérique sont de : - produire un catalogue de données de la recherche à l'échelle du laboratoire en anticipant l'obligation de dépôt des données de recherche en archive ouverte pour les projets financés sur fonds publics (programme Horizon 2020), - conserver et préserver des documents de première main créés par les chercheurs, enseignants-chercheurs, doctorants; leurs donner une seconde chance d'être utilisés, - documenter autant que possible les ressources numériques avec des métadonnées descriptives, méthodologiques, techniques et juridiques, classer les ressources numériques dans des collections, leurs attribuer des mots clés (indexation matière) et les géolocaliser, - proposer un accès restreint aux responsables de collections pendant la phase de gestion des contenus numériques puis diffuser librement les données sur le web une fois la préparation achevée, - contribuer à renforcer la visibilité des recherches du laboratoire à l'international. Fonctionnalités de recherche : Anpersana est couplée à un moteur de recherche avancée. Il est possible de définir des recherches en texte intégral, d'utiliser des opérateurs booléens, et des requêtes de type expression exacte. Les documents sont indexés par sujet et géolocalisés sur une carte interactive. Ils peuvent être regroupés dans une collection numériques marquant une typologie particulière. Spécificités techniques La bibliothèque numérique fonctionne avec le programme Omeka, un logiciel libre, sous licence GPL, développé par le Center for History and new Media. Ecrit en PHP, Omeka utilise le système de gestion de base de données relationnelle (SGBDR) MySQL et le serveur web Apache. La

		technologique de la bibliothèque numérique et les normes retenues pour la description des contenus devrait assurer l'interopérabilité des données. Utilisation des (meta)données Les données d'Anpersana publiées sur le site sont conditionnées par son régime de protection légale. La bibliothèque numérique ne met en ligne que des documents natifs ou transformés dans le respect du droit d'auteur, et donc sous réserve de l'autorisation préalable des éventuels ayants droits. Toute utilisation commerciale des documents publiés sur Anpersana est exclue. Il est demandé à l'utilisateur qui souhaiterait réutiliser les (meta)données de mentionner le titulaire du droit d'auteur et de ne pas transformer les enegsitrements sonores sans accord préalable de l'éditeur scientifique. Equipe de production Initiée dans le cadre de l'appel à projets numérisation 2013 du Ministère de la Culture et de la Communication, la bibliothèque numérique Anpersana est le résultat de plusieurs mois de travail collectif entre l'unité CNRS IKER (UMR 5478), l'association Bigun Kakol, le Département TIC de l'Université de Bordeaux et la Direction du numérique (DN), Université de Pau et des Pays de l'Adour. En 2017, Anpersana intègre un corpus de lettres basques du XVIIIe siècle dans la collection Le Dauphin collectés aux Archives de l'Amirauté britannique (Hight Court Admiralty) à Londres dans le cadre du programme thématique 2016-2020 de la Maison des Sciences de l'Homme d'Aquitaine (MSHA). Responsable du projet Anpersana : Jean-Philippe Talec, CNRS IKER (UMR 5478) Hébergeur : Direction du numérique (DN), Université de Pau et des Pays de l'Adour

Catégories, quoi, qui, pourquoi	**Accronyme Site web Sources**	**Présentation**
France Entrepôt de données sonores Qui le fait : La TGIR Huma-Num Ce que ça fait : Archivage pérenne et diffusion des données sonores issues de la recherche en SHS Pourquoi :	COllection de COrpus Oraux Numériques, COCOON : https://cocoon.huma-num.fr/exist/crdo/	Une plateforme technique qui accompagne les producteurs de ressources orales, à créer, structurer et archiver leurs corpus ; un corpus pouvant se composer d'enregistrements (en général audio) accompagnés éventuellement d'annotations de ces enregistrements. Les ressources déposées sont dans un premier temps cataloguées et stockées, puis, dans un deuxième temps archivées dans l'archive de la TGIR Huma-Num. L'auteur et son institution restent responsables des documents déposés et peuvent bénéficier d'un accès restreint et sécurisé à leurs données, pendant une période définie, si le contenu de l'information est considéré sensible.

Accompagner les producteurs des données sonores dans leur gestion		

Catégories, quoi, qui, pourquoi	**Accronyme Site web Sources**	**Présentation**
International Plateforme de gestion de projets de recherche et outil de collaboration Qui le fait : L'entreprise Center for Open Science, COS Ce que ça fait : Permet le développement des projets de recherche, la gestion des contenus de recherche, dont les données et leur diffusion Pourquoi : Créer une infrastructure gratuite de soutien à la recherche	Open Science Framework, OSF : https://osf.io/ Source complémentaire : https://cos.io/our-products/open-science-framework/	The Open Science Framework (OSF) provides free and open source project management support for researchers across the entire research lifecycle. As a collaboration tool, the OSF helps researchers work on projects privately with a limited number of collaborators and make parts of their projects public, or make all the project publicly accessible for broader dissemination. As a workflow system, the OSF enables connections to the many services researchers already use to streamline their process and increase efficiency. As a flexible repository, it can store and archive research data, protocols, and materials. Structured projects: Access files, data, code, and protocols in one centralized location and easily build custom organization for your project - No more trawling emails to find files or scrambling to recover from lost data Controlled access: Control which parts of a project are public or private, making it easy to collaborate and share with the community or just your team Enhanced workflow: Automate version control, get persistent identifiers for projects and materials, preregister your research, and connect your favorite third party services directly to the OSF Extend Your Research. Automatically create a preprint, meeting abstract, or add to your institutional repository. Manage multi-institutional projects.

ANNEXE 5 : Questionnaire

Les données de la recherche et la bibliothèque numérique ANPERSANA

Ce questionnaire est destiné aux chercheurs du laboratoire IKER. Il permettra d'identifier les usages et pratiques des chercheurs dans la gestion des données produites et collectées au cours du travail de recherche, ainsi que de connaître quelles sont leurs perceptions au sujet de l'ouverture des données. Cette enquête est conçue en vue de continuer l'avancement du projet de bibliothèque numérique ANPERSANA et de développer, éventuellement, des formations adaptées aux besoins du public de notre centre de documentation.

Bonjour,

Dans le cadre du projet de la bibliothèque numérique ANPERSANA, certaines des données primaires collectées au cours des projets de recherche du laboratoire sont désormais disponibles en accès libre sur la plateforme web : https://anpersana.iker.univpau.fr/

La dernière contribution, la correspondance de la collection Le Dauphin, composée de 49 lettres en basque-labourdin de l'époque, XVIIIe siècle, a récemment été numérisée et mise en ligne. L'objectif est en effet de continuer l'enrichissement de la bibliothèque numérique en diffusant des manuscrits, images, sons et tout autre type de document contenant des données brutes issues de la recherche dans les études basques.

Nous vous invitons à répondre toutes les questions de ce questionnaire. Nous sommes intéressés par vos usages et pratiques au sujet des données de la recherche. Cela nous permettra, éventuellement, de développer des formations sur la gestion de données, ainsi que de vous accompagner dans l'élaboration d'un Plan de gestion des données. Vos perceptions seront également utiles pour la suite du projet de bibliothèque numérique, son enrichissement et amélioration.

Pour remplir ce questionnaire environ 10 minutes vous seront nécessaires.

« Ces informations sont collectées en vue de permettre la réalisation de la présente enquête ; elles sont susceptibles de permettre une identification indirecte, en fonction des champs renseignés. Seules M. Talec et M. Fuentes Zamalloa ont accès à l'ensemble de ces données, et les résultats ne permettront pas l'identification des personnes. En application de la loi n°78-17 modifiée, vous disposez d'un droit d'accès, de rectification, et de suppression aux données vous concernant en contactant Jean-Philippe Talec ou Aitor Fuentes Zamalloa »

Il y a 51 questions dans ce questionnaire

VOTRE STATUT ET DOMAINE DE SPECIALISATION

1 Vous êtes : (Veuillez sélectionner une seule des propositions suivantes)

Chercheur CNRS

Enseignant chercheur universitaire

Chercheur postdoctoral (Post-Doc)

Doctorant

ITA

ITRF

Autre …….

2 A quelle ou quelles organisation-s êtes-vous rattaché-e ? (Veuillez choisir toutes les réponses qui conviennent)

CNRS

UBM

UPPA

UPV/EHU

Autre: …….

3 Quel est votre âge ?

(Veuillez sélectionner une seule des propositions suivantes)

22-35 ans

36-45 ans

46-55 ans

56-68 ans

69 ans ou plus

4 Dans le domaine des études basques, quelle est votre spécialité principale ?

(Veuillez choisir toutes les réponses qui conviennent)

Linguistique

Littérature

Didactique / enseignement

Autre : …….

5 Veuillez préciser brièvement votre sous-domaine, ou vos sous-domaines de spécialisation.

(p. ex. : dialectologie, sémantique, littérature du XIXe, l'enseignement en contexte multilingue, etc.).

Veuillez écrire votre réponse ici : …….

6 Vous êtes impliqué-e dans la recherche depuis :

(Veuillez sélectionner une seule des propositions suivantes)

< 5 ans

entre 5 et 10 ans

entre 10 et 20 ans

> 20 ans

LES DONNEES DE LA RECHERCHE DANS LE CADRE DE VOTRE PROJET : QUESTIONS ADMINISTRATIVES

7 Avez-vous été porteur d'au moins un projet au cours des 5 dernières années ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

8 De quels types de projets avez-vous été porteur au cours des 5 dernières années ?

(Répondre à cette question seulement si les conditions suivantes sont réunies : La réponse n'était 'Non' à la question '7 [G02Q07]' (Avez-vous été porteur d'au moins un projet au cours des 5 dernières années ?))

(Veuillez choisir toutes les réponses qui conviennent)

Européen

National

Régional

Départemental

Autre: …….

9 Précisez l'intitulé de votre projet.

(Répondre à cette question seulement si les conditions suivantes sont réunies : La réponse n'était 'Non' à la question '7 [G02Q07]' (Avez-vous été porteur d'au moins un projet au cours des 5 dernières années ?))

Veuillez écrire votre réponse ici : …….

10 Est-ce que des recommandations concernant la gestion des données de la recherche sont explicitées dans le descriptif de votre projet ?

(Répondre à cette question seulement si les conditions suivantes sont réunies :La réponse n'était 'Non' à la question '7 [G02Q07]' (Avez-vous été porteur d'au moins un projet au cours des 5 dernières années ?)

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

LA GESTION DE VOS DONNEES

11 Signalez les types de données numériques que vous produisez lors de vos recherches.

(Veuillez choisir toutes les réponses qui conviennent)

Enregistrements sonores

Bases de données

Lexiques

Données statistiques

Corpus de données linguistiques

Transcriptions phonétiques

Transcriptions de texte en orthographe moderne

Numérisations de texte

Numérisations d'image

Vidéos

Questionnaires d'enquêtes complétés

Texte encodé (XML, SGML, etc.)

Cartes (historique, sociolinguistique, dialectologique, etc.)

Autre : …….

12 Êtes-vous familier avec le terme métadonnées ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui, je connais bien le concept.

Oui, à peu près

Non

13 Où stockez-vous les fichiers numériques de vos données de la recherche ?

(Veuillez choisir toutes les réponses qui conviennent)

Dans mon ordinateur personnel ou de travail

Dans un disque dure externe

Online (par le moyen d'un opérateur de services de stockage en ligne ou cloud service)

Dans le(s) serveur(s) de ma structure de rattachement

Dans une clé USB

Dans un DVD

Dans un CD

Autre : …….

14 Avez-vous été formé à la gestion des données produites au cours vos activités de recherche ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

15 Par quelle instance ?

(Répondre à cette question seulement si les conditions suivantes sont réunies : La réponse n'était 'Non' à la question '14 [G03Q14]' (Avez-vous été formé à la gestion des données produites au cours vos activités de recherche ?))

Veuillez écrire votre réponse ici : …….

16 Sur quelle durée ?

(Répondre à cette question seulement si les conditions suivantes sont réunies : La réponse n'était 'Non' à la question '14 [G03Q14]' (Avez-vous été formé à la gestion des données produites au cours vos activités de recherche ?))

Veuillez écrire votre réponse ici : …….

18 Savez-vous qu'est-ce qu'un Plan de gestion des données (PGD) ou *Data Management Plan* (DMP) ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui, je connais bien le concept

Oui, à peu près

Non

19 Saviez-vous que certains programmes, comme le Programme européen pour la recherche et l'innovation Horizon 2020, exigent un Plan de gestion des données (PGD) pour le financement d'un projet de recherche ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui

J'en ai entendu parler.

Non

20 Avez-vous élaboré un plan de gestion des données pour votre projet de recherche actuel ?

(Répondre à cette question seulement si les conditions suivantes sont réunies : La réponse n'était 'Non' à la question '7 [G02Q07]' (Avez-vous été porteur d'au moins un projet au cours des 5 dernières années ?))

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non, mais je pense le faire.

Non

21 Pour vous, avoir un plan de gestion des données, c'est ou ce serait :

(Veuillez sélectionner une seule des propositions suivantes)

Très utile pour mes activités de recherche

Utile pour mes activités de recherche

Ni utile ni inutile pour mes activités de recherche

Inutile pour mes activités de recherche

Une contrainte pour mes activités de recherche

Je ne sais pas

DROITS ET ASPECTS ETHIQUES

22 Est-ce que vos données vous appartiennent ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

Juste certaines de mes données

Je ne sais pas

23 Est-ce que vous avez les droits de diffusion de vos données ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

Juste de certaines de mes données

Je ne sais pas.

24 Savez-vous qu'en tant qu'auteur de vos données, ou avec l'autorisation des titulaires des droits, vous pouvez les protéger en choisissant parmi différentes licences en fonction de vos besoins ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

25 Lors d’une éventuelle diffusion de vos données, choisiriez-vous l’une des 6 licences Creative Commons ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

Juste pour certaines de mes données

Je ne sais pas.

26 Est-ce qu’il y a une période d’embargo associée à la publication ou diffusion de vos données ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

Je ne sais pas.

27 Dans le cadre de votre projet actuel, est-ce que vous avez déjà publié un article scientifique au sujet de vos données ou avec des résultats obtenus à partir de vos données ?

(Répondre à cette question seulement si les conditions suivantes sont réunies : La réponse n’était 'Non' à la question '7 [G02Q07]' (Avez-vous été porteur d'au moins un projet au cours des 5 dernières années ?))

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

28 Est-ce que la diffusion publique de vos données nécessiterait d’une prise en compte des questions éthiques, tel que l’anonymat de vos sources ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

Je ne sais pas

29 Est-ce que vous qualifieriez le contenu de vos données comme « sensible » ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

Je ne sais pas.

COLLABORATION SCIENTIFIQUE

30 Partagez-vous vos données avec d'autres chercheurs ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

31 Avec qui ?

Répondre à cette question seulement si les conditions suivantes sont réunies : La réponse n'était 'Non' à la question '30 [G05Q30]' (Partagez-vous vos données avec d'autres chercheurs ?))

(Veuillez choisir toutes les réponses qui conviennent)

Avec des chercheurs de la même équipe

Avec des chercheurs de la même université ou laboratoire

Avec des chercheurs d'autres institutions

Autre: …….

32 Avez-vous des inquiétudes concernant le partage de données ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

33 Lesquelles ?

(Répondre à cette question seulement si les conditions suivantes sont réunies : La réponse n'était 'Non' à la question '32 [G05Q32]' (Avez-vous des inquiétudes concernant le partage de données ?))

(Veuillez choisir toutes les réponses qui conviennent)

Peur de perdre l'avantage scientifique

Les questions juridiques et éthiques

L'utilisation abusive des données

La mauvaise interprétation des données

Le manque de ressources (techniques, financières, personnel, etc.) Le manque de protection des politiques et des droits appropriés

Autre: …….

VOTRE AVIS SUR L'OUVERTURE DES DONNEES DE LA RECHERCHE

Signalez votre degré d'accord ou désaccord avec les propositions suivantes

(5 étant « complétement d'accord », 3 « ni d'accord ni pas d'accord » et 1 « pas du tout d'accord ».)

(Veuillez sélectionner une seule option)

34 L'ouverture des données de la recherche contribue à la transparence de la recherche.

1 2 3 4 5

35 L'ouverture des données de la recherche améliore les conditions de reproductibilité des expériences scientifiques.

1 2 3 4 5

36 L'ouverture des données de la recherche est économiquement viable.

1 2 3 4 5

37 L'ouverture des données de la recherche n'est qu'une mode passagère.

1 2 3 4 5

38 La culture de l'open access en général n'est qu'une mode passagère.

1 2 3 4 5

39 L'ouverture des données de la recherche est positive pour les Sciences humaines et sociales.

1 2 3 4 5

40 L'ouverture des données de la recherche dans les études basques est spécialement nécessaire par rapport à la majorité des autres domaines en Sciences humaines et sociales.

1 2 3 4 5

41 L'ouverture des données de la recherche dans les études basques est inutile en comparaison avec la majorité des autres domaines en Sciences humaines et sociales.

1 2 3 4 5

42 L'impact que mes publications peuvent avoir auprès de mes pairs est important pour moi.

1 2 3 4 5

46 Je voudrais en savoir plus sur le projet de bibliothèque numérique ANPERSANA et sur les avantages ou inconvénients de diffuser mes données sur la plateforme

1 2 3 4 5

43 L'impact que mes publications peuvent avoir auprès du grand public est important pour moi.

1 2 3 4 5

44 J'ai déjà consulté la bibliothèque numérique ANPERSANA au moins une fois.

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

45 Le projet de bibliothèque numérique ANPERSANA est positif pour rapprocher le public non-chercheur aux études basques.

1 2 3 4 5

49 Sur quel plateforme voudriez-vous les diffuser ?

(Répondre à cette question seulement si les conditions suivantes sont réunies : La réponse n’était 'Non' à la question '48 [G06Q48]' (Souhaiteriez-vous diffuser certaines de vos données ?))

(Veuillez choisir toutes les réponses qui conviennent)

ZENODO

ANPERSANA

NAKALA

OSF

Autre: ……..

47 A un moment, je voudrais être accompagé-e ou formé-e sur :

(Veuillez choisir toutes les réponses qui conviennent)

Les bonnes pratiques en gestion de données

L'élaboration d'un Plan de gestion de données

La gestion de références et l'automatisation dans la création d'une bibliographie

L'utilisation d'outils et plateformes de collaboration scientifique

Les techniques de recherche de ressources sur le web

Le cycle de vie de données de la recherche

Les métadonnées

Les bases des données

Les enjeux de l'ouverture des données de la recherche

Les questions juridiques des données de la recherche (droits de propriété, de diffusion, de réutilisation, etc.)

La numérisation et l'édition d'images

Je ne souhaite pas être formé-e ou accompagné-e.

Autre: ……

48 Souhaiteriez-vous diffuser certaines de vos données ?

(Veuillez sélectionner une seule des propositions suivantes)

Oui

Non

Je ne sais pas.

Nous vous remercions de votre participation à cette enquête.

ANNEXE 6 : Croissement de résultats du questionnaire

	Du total des répondants	L'âge				Le statut					Expérience			
		22-35	36-45	46-55	56-68	Doctorant	Post doc	Enseignant chercheur	Chercheur CNRS	Autre ou sans réponse	-5	entre 5 et 10	entre 10 et 20	+20
Partage = OUI	65,38 (17/26)	29,41 (5/17)	29,41 (5/17)	11,76 (5/17)	29,41 (5/17)	30,77 (8/26)	15,38 (4/26)	19,23 (5/26)	19,23 (5/26)	15,38 (4/26)	26,92 (7/26)	38,46 (10/26)	15,38 (4/26)	19,23 (5/26)
Souhaitent diffuser leurs données = OUI	46,15 (12/26)	22,22 (2/9)	50 (5/10)	100 (2/2)	60 (3/5)	37,50 (3/8)	50 (2/4)	20 (1/5)	80 (4/5)	--	28,57 (2/7)	40 (4/10)	75 (3/4)	60 (3/5)
Souhaitent diffuser leurs données = ne sait pas	50 (13/26)	77,78 (7/9)	40 (4/10)	0	40 (2/5)	62,50 (5/8)	50 (2/4)	80 (4/5)	0 (0/5)	--	71,43 (5/7)	50 (5/10)	25 (1/4)	40 (2/5)
Souhaitent diffuser leurs données = NON	3,85 (1/26)	0	1 (1/10)	0	0	0	0	0	20 (1/5)	--	0	10 (1/10)	0	0
Inquiétudes = OUI	26,92 (7/26)	42,86 (3/7)	57,14 (4/7)	0	0	42,86 (3/7)	14,29 (1/7)	14,29 (1/7)	28,57 (2/7)	--	28,57 (2/7)	42,86 (3/7)	28,57 (2/7)	0

	Du total des répondants	Rôle des acteurs institutionnels												Formation		PGD			
		Structure de rattachement					Porteurs d'un projet = OUI	Préconisations explicitées dans le projet = OUI	Type de projet					Oui	Non	Oui	Oui dans l'avenir	Non	Ne répondent pas
		CNRS	UBM	UPPA	EHU	Autres org.	50,00 (13/26)	46.15 (6/13)	Eur.	Nat.	Rég.	Dép.	Autre	3,85 (1/26)	96,15 (25/26)	11,54 (3/26)	7,69 (2/26)	30.77 (8/26)	50 (13/26)
Partage = OUI	65,38 (17/26) *	47,06 (8/17)	17,65 (3/17)	11,76 (2/17)	11,76 (2/17)	29,41 (5/17)	84.62 (11/13)	100,00 (6/6)	17,65 (3/17)	23,53 (4/17)	41,18 (7/17)	11,76 (2/17)	23,53 (4/17)	100 (1/1)	64 (16/25)	100 (3/3)	50 (1/2)	87.50 (7/8)	--
Inquiétudes = OUI	26,92 (7/26)*	--	--	--	--	--	--	16.67 (1/6)	--	--	--	--	--	--	--	--	--	--	--
Embargo = NON	46,15 (12/26)	62,50 (5/8)	40 (2/5)	28,57 (2/7)	71,43 (5/7)	--	53,85 (7/13)	50 (3/3)	66,66 (2/3)	40 (2/5)	62,50 (5/8)	66,66 (2/3)	--	--	--	100 (3/3)	100 (2/2)	25 (2/8)	--
Embargo = ne sait pas	53,85 (14/26)	37,50 (3/8)	60 (3/5)	71,43 (5/7)	28,57 (2/7)	--	46,15 (6/13)	50 (3/3)	33,33 (1/3)	60 (3/5)	37,50 (3/8)	33,33 (1/3)	--	--	--	0	0	75 (6/8)	--

	PGD			L'âge								Préconisations explicitées dans le projet	
	Oui		Non	22-35		36-45			46-55		56-68	OUI	NON
Stockage - bonnes pratiques (11/26)*	2/7		5/7	5/10		4/10			1/10		0	3/5	2/5
Stockage pratiques risquées (15/26)*	1/10		9/10	4/15		6/15			1/15		4/15	2/5	3/5
Connaissent métadonnées	3/3		7/10	5/9		6/10			1/1		4/5	5/6	5/7
Connaissent métadonnées à peu près	2/3	1/3	1/7	6/7	0/5	5/5	3/6	3/6	1/1	0/1	3/5	1/5	
Connaissent PGD	--			--		--			--		--	4/6	2/2
Ont PGD	--			--		--			--		--	1/5	2/7

Inquiétudes = OUI		26,92 (7/26)
Préconisations non-explicitées dans le projet		85,71 (6/7)
Types d'inquiétude	Perdre l'avantage scientifique	57,14 (4/7)
	Questions juridiques et éthiques	28,57 (2/7)
	Utilisation abusive des données	28,57 (2/7)
	Mauvaise interprétation	14,29 (1/7)
	Manque de ressources	28,57 (2/7)
	Manque de protection	0
	Questions politiques	14,29 (1/7)

		Linguistique	Littérature	Didactique / enseignement	Autre	Du total des participants
		53,33 (16/30)	26,66 (8/30)	10 (3/30)	10 (3/30)	
Partage =	OUI	100 (16/16)	37,5 (3/8)	33,33 (1/3)	33,33 (1/3)	80,77 (21/26)
Type des données	Questionnaires complétés	50 (8/16)	25 (2/8)	33,33 (1/3)	66,66 (2/3)	50 (13/26)
	Enregistrements sonores	43,75 (7/16)	37,5 (3/8)	66,66 (2/3)	66,66 (2/3)	53,85 (14/26)
	Transcriptions de texte	43,75 (7/16)	25 (2/8)	66,66 (2/3)	66,66 (2/3)	50 (13/26)
	Corpus de données linguistiques	25 (4/16)	12,5 (1/8)	33,33 (1/3)	33,33 (1/3)	26,92 (7/26)
	Bases de données	25 (4/16)	0	0	100	26,92 (7/26)
	Transcriptions phonétiques	18,75 (3/16)	0	0	33,33 (1/3)	15,35 (4/26)
	Données statistiques	18,75 (3/16)	12,5 (1/8)	33,33 (1/3)	66,66 (2/3)	26,92 (7/26)
	Num. texte	6,25 (1/16)	25 (2/8)	33,33 (1/3)	33,33 (1/3)	19,23 (5/26)
	Num. image	6,25 (1/16)	0	0	66,66 (2/3)	11,53 (3/26)
	Texte encodé	6,25 (1/16)	0	0	33,33 (1/3)	7,69 (2/26)
	Cartes	6,25 (1/16)	0	0	33,33 (1/3)	7,69 (2/26)
	Wiki	6,25 (1/16)	0	0	0	3,84 (1/26)
	Vidéos	0	12,5 (1/8)	33,33 (1/3)	0	3,84 (1/26)
	Lexiques	0	12,5 (1/8)	0	0	3,84 (1/26)
	Guides d’analyse	0	12,5 (1/8)	0	0	3,84 (1/26)
	Textes littéraires	0	12,5 (1/8)	0	0	3,84 (1/26)
Souhaitent diffuser =	OUI	(7/16)	37,5 (4/8)	33,33 (1/3)	33,33 (1/3)	46,15 (12/26)*
Souhaitent diffuser =	ne sait pas	(8/16)	50 (4/8)	33,33 (2/3)	66,66 (2/3)	50 (13/26)*
Souhaitent diffuser =	NON	1	0	0	0	3,85 (1/26)*

	OUI	NON	De/sur/pour certaines des données	Ne sait pas	Préconisations dans le projet = connaissent les notions de la 1re colonne
Sont propriétaires des données	53,85 (14/26)	3,85 (1/26)	23,08 (6/26)	19,23 (5/26)	6/6
Droits de diffusion	30,77 (8/26)	7,69 (2/26)	23,08 (6/26)	38,46 (10/26)	6/6
Savent que la licence peut être choisie par l'auteur-propriétaire	30,77 (8/26)	69,23 (18/26)	--	--	3/6
Si diffusion, choix de licence CC ?	19,23 (5/26)	11,54 (3/26)	7,69 (2/26)	61,54 (16/26)	--
Embargo	0	46,15 (12/26)*	--	53,85 (14/26)	3/6
Données sensibles	23,08 (2/26)	69,23 (18/26)	--	7,69 (2/26)	--
Questions éthiques se posent lors d'une éventuelle diffusion	53,85 (14/26)	15,38 (4/26)	--	30,77 (8/26)	--

ANNEXE 7 : Grilles d'entretien remplies

Entretien nº1, id.: E1PEDS			
THEMES / OPTIONS			REPONSES
Données de la recherche	Types produits		Données orales / archives orales organisées en corpus, notamment bases de données. Quelques données écrites aussi.
	Enregistrements sonores dans votre domaine		La majorité des collègues travaillent avec des données orales. Parfois des données écrits issues des revues où il y a des sections en « texte libre », équivalent à l'oral.
	Stockage, Format des fichiers et métadonnées		Les corpus sont gérés par une entreprise externe d'aide à la gestion de bases de données. Le service informatique de l'entreprise est intégré dans le projet de recherche. Serveur de l'entreprise et de l'université. WAV Métadonnées produites dans la mesure où elles facilitent la consultation et les recherches. Métadonnées descriptives sur l'informant, sur la donnée de recherche, ayant toujours à l'esprit que celui qui accèdera aux données ne sera pas toujours un linguiste.
Partage	Inclut diffusion		C'est la mise en ligne des données secondaires. Les brutes ou primaires, parfois issues d'entretiens élicités mais surtout dans le cas des entretiens libres, sont éthiquement les plus compromettants. Selon le sujet, l'intimité, les problèmes, les aspect personnels sont évoqués. Dans ce sens on applique un filtre et donc on produit une donnée secondaire, avec Praat, que l'on peut mettre en ligne, car les données sont anonymisées, sans contenu compromettants… Pas seulement au niveau de l'identité, au niveau du contenu toute donnée pouvant désigner l'identité est éditée aussi. La voix n'est pas modifiée car ils ont l'autorisation signée. Les données brutes ne sont jamais partagées, même pas avec les collègues.
	Exclut diffusion		
	Inqui études	Oui	
		Non	
Diffusion	Réalisation	Oui	
		Non	
	Perception	Avantages	
		Inconvénients	
Quand l'accès aux données est restreint …			On demande au producteur de pouvoir accéder aux données, en prenant de précautions car l'on sait que les questions d'éthique se jouent, surtout pour les entretiens libres, par rapport aux entretiens élicités. On essaye de traiter les données des autres tel qu'on aimerait que l'on traite les nôtres.
Contribue à l'acquisition de compétences	PGD	Oui	Tout à fait. Sans doute.
		Non	
	Précisions	Oui	Pendant la rédaction du projet on se rend compte du besoin d'aborder le sujet des données. Mais surtout à la fin, quand on a une masse de données si importante, c'est une grande responsabilité. Donc à la fin du projet on aime écrire un article sur la méthodologie et les données. Le sujet des données s'impose à la fin du projet, qu'on l'ait pris en compte au début ou non il se présente, c'est inéludable.
		Non	
Formation			Les chercheurs du laboratoire sont systématiquement informés sur les formations de l'UrfIST de Bordeaux.

Entretien nº2, id.: E2DDDG			
THEMES / OPTIONS			REPONSES
Données de la recherche	Types produits		Enregistrement. Un choix, car les données vidéo trop lourdes.
	Enregistrements sonores dans votre domaine		Enregistrements et un peu de vidéo aussi.
	Stockage, Format des fichiers et métadonnées		Praat et Audacity pour les enregistrements longs WAVE pour la qualité en phonologie Transcriptions sur Word. Classification par dossiers. Hiérarchie : village, locuteur. Dans le dossier locuteurs 3 documents : profil du locuteur et transcriptions en Word, enregistrement des entretiens en WAVE. Ne sait pas qu'est-ce que les métadonnées. Sur ordinateur et deux disques durs. Ne sait pas qu'il y a un service de l'université pour le stockage.
Partage	Inclut diffusion		Interprète partage comme étant = diffusion. Préoccupé par la confidentialité et l'intimité des participants.
	Exclut diffusion		
	Inquiétudes	Oui	Questions juridiques.
		Non	
Diffusion	Réalisation	Oui	
		Non	
	Perception	Avantages	Ne sait pas s'il y a des avantages personnels. En tout cas, il y a la question du temps qui se pose aussi par rapport à la gestion de données nécessaire et préalable à la diffusion.
		Inconvénients	
Quand l'accès aux données est restreint …			Assume que les données ne soient pas accessibles et fait confiance à l'auteur. Pas de démarche de mise en contact avec le producteur des données.
Contribue à l'acquisition de compétences	PGD	Oui	Croit qu'il a des avantages, parce qu'on le dit, mais n'est pas trop sur ce que c'est qu'un PGD.
		Non	
	Précisions	Oui	Dans le contrat du projet il n'y a pas de recommandation. Mais il y a la charte éthique de l'université. A son avis dans la charte les données ne sont pas abordées.
		Non	
Formation			Ne sait pas qu'est-ce que l'UrfIST et ne connait pas des dispositifs de formation sur la documentation pour les chercheurs.

Entretien nº3, id.: E3PNAL			
THEMES / OPTIONS			REPONSES
Données de la recherche	Types produits		Enregistrements de productions orales lues et les transcriptions
	Enregistrements sonores dans votre domaine		Enregistrements aussi et transcriptions.
	Stockage, Format des fichiers et métadonnées		WAVE et MP3 OpenOffice Word 3 Disque dures Connait les serveurs de l'université Connait les métadonnées et a créé des tableaux décrivant les données.
Partage	Inclut diffusion		Décrit le partage à tous les niveaux : entre collègues et la diffusion. Considère que les données ne sont pas à lui. Elles appartiennent au projet.
	Exclut diffusion		
	Inquiétudes	Oui	Au niveau personnel aucune inquiétude.
		Non	
Diffusion	Réalisation	Oui	Veut le faire. Mais attend la permission des locuteurs.
		Non	
	Perception	Avantages	Pas d'avantages personnelles. C'est s'exposer à la critique aussi, mais c'est comme ça la recherche. Ça vaut la peine.
		Inconvénients	
Quand l'accès aux données est restreint …			N'a jamais senti le besoin d'utiliser des données autres que les résultats. Mais ça aurait pu lui arriver
Contribue à l'acquisition de compétences	PGD	Oui	N'a pas entendu parler.
		Non	
	Précisions	Oui	Mais seulement sur comment faire les enregistrements. Aucune recommandation pour après l'enregistrement.
		Non	
Formation			N'a pas entendu parler de l'UrfIST.