Des insouciances de l'archivisme à une réflexivité constante : considérations éthiques en linguistique de terrain à l'ère du Traitement Automatique des Langues

Camille Noûs, Alexis Michaud

Des insouciances de l'archivisme à une réflexivité constante : considérations éthiques en linguistique de terrain à l'ère du Traitement Automatique des Langues

Camille Noûs

Laboratoire Cogitamus

Camille Noûs est un individu collectif qui symbolise notre attachement aux valeurs d'éthique et de probation que porte le débat contradictoire, et conscient de ce que nos résultats doivent à la construction collective ; le « Noûs », porteur d'un Nous collégial, est également une référence au concept d'« intellect » (νοῦς, *noûs*) de la philosophie grecque.

Alexis Michaud

Langues et Civilisations à Tradition Orale (LACITO), CNRS – Université Sorbonne Nouvelle, Villejuif, France

Linguiste de terrain, passionné par la documentation des langues rares et menacées et par les questions liées à l'archivage pérenne et la diffusion des corpus.

identifiant ORCID : 0000-0003-1165-2680

alexis.michaud@cnrs.fr

Ce texte se veut un essai de retour réflexif sur un parcours : deux décennies d'enquêtes linguistiques de terrain sur des langues rares, au Yunnan (Chine), et d'engagement en faveur de l'ouverture des données de terrain (corpus multimédia de langues rares). Le cadre temporel coïncide avec deux décennies de progrès spectaculaires de l'informatique, et notamment du Traitement Automatique des Langues. Au fil des réflexions, il ressort que l'adoption conséquente de principes de Science ouverte amène, non pas à se doter d'un ensemble de solutions qui permettent d'éviter les soucis, mais à faire une place centrale aux questions éthiques et socio-politiques soulevées par la constitution, la publication électronique et l'exploitation de ressources en langues rares. Ces thèmes appellent une réflexion collective constante, à laquelle soient associé·es des spécialistes de Traitement Automatique des Langues.

Mots-clés : données ouvertes, langues rares, archives, éthique et déontologie, Traitement Automatique des Langues

From Archival Activism to Constant Reflexivity: Ethical Considerations in Field Linguistics in the Era of Natural Language Processing

This text is an attempt to look back on two decades of linguistic fieldwork on endangered languages in Yunnan (China) and of activism in favour of opening up fieldwork data (multimedia corpora of undocumented or little-documented languages). The time frame coincides with two decades of spectacular progress in computer science, and in particular in Natural Language Processing. Looking back, it seems clear that the consistent adoption of Open Science principles implies that the ethical and socio-political issues raised by the creation, electronic publication and exploitation of fieldwork data must be given a central place (rather than hoping for ready-made solutions to deflect concerns and fend off trouble). These topics call for constant collective reflection involving specialists in Automatic Language Processing.

Keywords: open data, linguistic heritage and diversity, archives, research ethics, Natural Language Processing

从无所忌惮的数据存档激进主义到不断的反思：自动语言处理时代田野语言学的伦理考量

本文试图回顾二十年来在云南（中国）开展的濒危语言田野调查，以及为开放田野调查数据（未记录或记录较少的语言多媒体语料库）所做的努力。这二十年，是计算机科学，特别是自然语言处理技术取得巨大进步的二十年。回顾过去，我们似乎可以清楚地看到，始终如一地采用开放科学原则意味着：必须将实地调查数据的创建、电子出版和利用所引发的伦理和社会政治问题放在中心位置。而不是寄希望于依靠现成的解决方案来转移人们的关注和避免纠纷。这些问题的解决，需要语言学家与语言处理专家们协同合作，不断进行集体思考。

Introduction

Les études de cas peuvent présenter une certaine utilité pour aborder les délicates questions éthiques qui se posent autour de l'ouverture des données de la recherche. « Les politiques de science ouverte ont placé au cœur du travail scientifique l'enjeu de l'ouverture des données de la recherche (Galonnier *et al.* 2019). (…) On ne compte plus depuis quelques années les séminaires, journées d'étude, formations, tout comme les recommandations officielles sur le sujet. Cependant, très peu de publications rendent compte de manière documentée de l'expérience de chercheurs ou d'ingénieurs, pourtant les premiers concernés, ce qui signale un retard français comparativement au monde anglo-saxon et à l'Europe du nord. Ainsi, l'intérêt à restituer certaines expériences pour réfléchir collectivement à ces questions est réel et invite à décaler la perspective par rapport au seul cadre théorique des « bonnes pratiques » » (Bendjaballah et Garcia 2023).

Dans cette perspective, le présent texte[1] se veut un essai de retour réflexif sur le volet juridique et éthique d'un parcours de linguiste de terrain : deux décennies d'enquêtes sur des langues minoritaires (en voie de remplacement par la langue nationale) au Yunnan (Chine), de publication de ressources orales en archive ouverte, et d'exploitation de ces ressources, y compris dans le cadre de partenariats avec des chercheuses et chercheurs en Traitement Automatique des Langues (TAL).

Le cadre temporel de ce parcours coïncide en effet avec deux décennies de progrès rapides de l'informatique, d'une importance tout à fait centrale pour nos disciplines. Le fort potentiel applicatif du TAL pour la documentation des langues est bien identifié – et des réalisations viennent peu à peu le concrétiser –, et les collaborations entre TAListes et spécialistes de langues rares ouvrent d'importantes perspectives en recherche pour toutes les disciplines concernées, et donnent lieu à des partenariats pluridisciplinaires prometteurs. Mais avant même la mise en place d'éventuels projets communs avec des TAListes, le simple fait d'avoir un minimum d'information au sujet du domaine permet de prendre conscience du rythme très rapide de l'évolution du TAL. Cela sensibilise efficacement au fait qu'il n'est pas possible de prévoir à un instant *t* toutes les réutilisations qui seront faites des données à *t+1*, ce qui comporte des conséquences aussi déconcertantes qu'essentielles en matière de risques éthiques liés aux traitements réalisables sur les corpus (Badin *et al.* 2022, §52).

Au fil des réflexions, il ressort que le militantisme en faveur de l'ouverture des données de la recherche ne constitue pas par lui-même une réponse complète aux questions juridiques et éthiques soulevées par la constitution, la publication électronique et l'exploitation de ressources linguistiques en langues rares. Sur la base de ce constat (largement partagé) ressortent quelques pistes qui constituent autant de défis et de chantiers pour les Communs numériques.

1 Le choix de faire figurer Camille Noûs comme autrice symbolique de cet article vise à souligner la nature collégiale du travail scientifique. Camille Noûs est un individu collectif qui symbolise notre attachement aux valeurs d'éthique et de probation que porte le débat contradictoire, et conscient de ce que nos résultats doivent à la construction collective ; le « Noûs », porteur d'un Nous collégial, est également une référence au concept d'« intellect » (νοῦς, *noûs*) de la philosophie grecque. Je suis en cela l'exemple d'une collègue de mon laboratoire (Simon et Noûs 2021). Il va de soi que le principe fondamental de la responsabilité de l'auteur n'est nullement remis en cause : les propos tenus ici engagent la responsabilité pleine et entière de l'auteur non fictif. Mais cette précision en appelle une seconde : la présence de Camille Noûs n'est pas un canular. Ni la mention du laboratoire Cogitamus, institution rassemblant des scientifiques de tous horizons disciplinaires autour des valeurs communes d'une recherche intègre, désintéressée, aspirant à créer, perpétuer, réviser et transmettre les savoirs. Fiction qui permet à des scientifiques de se rejoindre et de faire face autrement, à ce monde-ci et à ce présent-là. Cet espace est, entre autres, un lieu de dialogue sur la politique des sciences, préoccupées des sociétés humaines.

Les insouciances de l'archivisme numérique à la naissance d'internet : l'ouverture des données comme mouvement à sens unique

Les années 1990 ont été marquées simultanément par la progression des technologies numériques et d'internet. Double révolution pour la conservation et la diffusion de données de langues rares. La numérisation rendait théoriquement possible la conservation indéfinie des données multimédia, puisque la copie numérique se fait à l'identique, alors que les bandes magnétiques « repiquées » perdaient en qualité à chaque copie, et que les supports magnétiques vieillissaient de façon lente mais irréversible (Bonnemason, Ginouvès et Pérennou 2001, 4). À peine entrées dans l'éternité numérique, ces ressources seraient partagées avec le monde entier, grâce à internet. La Corée du Nord et quelques zones de conflit figuraient des îlots anachroniques, fermés aux voyageurs et à internet, dans un monde ouvert : d'Europe de l'Ouest, on pouvait aller partout (la réciproque était moins vraie, la génération de SOS-Racisme était déroutée d'entendre répéter « la France ne peut pas accueillir toute la misère du monde »). Le rideau de fer était levé, la Chine n'était pas inaccessible, le Japon était devenu familier : un adversaire économique pas commode, comme les États-Unis. La mondialisation laminait les sociétés, les cultures et les langues, mais il y avait quelque chose à faire pour sauver du naufrage ce qui pouvait l'être.

Travailler à une archive numérique avant l'essor de la Science ouverte : le temps des pionniers

Le programme à réaliser paraissait clair : recueillir une documentation fiable et abondante au sujet des langues et des cultures en voie de disparition, en appliquant les méthodes éprouvées des linguistes et anthropologues (Bouquiaux et Thomas 1971), et en archivant les données au format numérique. Un problème paraissait avoir priorité sur tous les autres : mettre fin à une situation de déperdition quasi-généralisée des données de la recherche. Mais pour évidente qu'elle paraisse, cette priorité ne s'imposait pas à tout le monde, loin de là. Travailler à une archive numérique avant l'essor de la Science ouverte était un acte militant. On n'était pas seul, mais on n'était qu'une poignée de pionniers, dont les moyens d'action paraissaient assez dérisoires. Le « programme Archivage du LACITO » (Jacobson, Michailovsky et Lowe 2001), qui prendrait plus tard le nom de « collection Pangloss », était opérationnel, pour pérenniser, exploiter et diffuser les enregistrements des linguistes de terrain. Annotation (transcription, gloses interlinéaires, traductions) balisée selon la norme XML et synchronisée phrase par phrase avec l'enregistrement numérisé, outils logiciels génériques et librement disponibles, consultation au moyen d'un simple navigateur internet, via une interface qui permet à l'utilisateur de choisir les informations affichées : tout paraissait en place pour emporter rapidement l'adhésion des linguistes. Mais à notre grande surprise, l'effet d'entraînement tardait (bien que le site ait attiré l'attention de Nick Thieberger[2] et de quelques autres collègues, en France et à l'étranger).

Jusqu'aux années 2010, notre engagement pour l'archivage numérique (« archivisme numérique ») était marqué par une relative insouciance[3] vis-à-vis des questions éthiques. Il s'agissait d'abord de sauver les données.

> Les chercheurs et étudiants ont tendance à constituer leur propre corpus à mesure des besoins de leur recherche, plutôt que de raisonner en termes de patrimoine documentaire partagé. Les fonds d'archives sont peu connus, les grands corpus

2 Nick Thieberger posait à l'époque en Australie les bases de ce qui est devenu l'une des principales archives de langues rares : *Pacific and Regional Archive for Digital Sources in Endangered Cultures* (PARADISEC) (voir Harris, Thieberger et Barwick 2015 ; et, en français, Vernaudon *et al.* 2021, 327).

3 Le terme d'*insouciance* est une allusion à une communication dans laquelle Michel Ferlus, diachronicien, linguiste de terrain spécialiste des langues d'Asie, revient sur le parcours qui a mené de la collecte de données sur le terrain au fil d'enquêtes, depuis les années 1960, jusqu'à leur publication numérique, non prévue initialement : « Des enregistrements sans expérience à la numérisation : les insouciances du terrain » (Ferlus 2017).

> distribués sur internet dépassent souvent les budgets du chercheur individuel, tandis que l'on peut enregistrer soi-même un corpus d'une qualité technique satisfaisante.
>
> On voudrait souligner ici les limites de cette logique : il est illusoire de penser que l'on peut à tout moment créer le corpus dont on a besoin. Dans le cas des langues en danger, la mise en commun des données existantes est particulièrement nécessaire. Mais dans l'étude des grandes langues, le travail documentaire ne demande pas moins de sérieux. (Michaud 2002)

Comme on l'imagine, pareil cours de morale dispensé par un nouveau venu drapé dans des habits de bonne sœur à cornettes ne recueillait guère l'adhésion des collègues. Cette communication (ma première publication scientifique) n'a été citée que par un voisin de palier au sein du LACITO (Jacobson 2004). À l'époque, les pratiques de partage de données de la recherche étaient peu courantes. « Les chercheurs étaient recrutés pour faire des recherches, publier des résultats mais, pour la majorité d'entre eux, ils ne se considéraient pas investis d'une mission de service public pour la diffusion des données qu'ils collectaient » (Simonnot 2020). On se voyait renvoyé à de rares organisations spécialisées, telles que ELRA (*European Language Resources Association*, fondée en 1995), dont il paraissait pourtant clair qu'elles n'avaient pas pour mission de prendre en charge les données de chercheurs. Le travail de mise en forme et archivage de données était activement découragé par des collègues (bien intentionnés) qui jugeaient que ces tâches de technicien-ingénieur et non de chercheur ralentissaient d'autant le rythme de publication, qui seul comptait dans l'évaluation.

C'était le règne du « mérite publicationnel » jugé à l'aune de la bibliométrie à l'américaine (statistique des citations), à peine voilé derrière des discours au sujet d'une nécessaire prise en compte de facteurs d'évaluation *aussi bien quantitatifs que qualitatifs*. Face au constat évident des limites et des biais de la bibliométrie (Archambault et Larivière 2009), des propositions de rafistolages divers et variés se succédaient (Fassoulaki *et al.* 2002), sans remettre en cause le postulat selon lequel la solution, la bonne, serait une mesure quantitative enfin débiaisée. Pour l'heure, l'institution retirait son soutien financier aux revues de « catégorie C » (Gunthert 2004) et s'autorisait à modifier le classement proposé par les instances de recrutement lorsqu'il ne coïncidait pas avec les indicateurs quantitatifs. L'Institut des Sciences Humaines et Sociales du CNRS reléguait en rang non utile tel·le candidat·e dont la liste de publications était jugée trop peu fournie, supplanté·e par une personne que le Comité national avait, selon ses critères, placée moins haut dans son classement.

Dans un tel contexte, les occasions de se frotter aux questions juridiques et éthiques liées aux données étaient rares, et ces thèmes n'étaient guère abordés dans les cursus universitaires. L'attention du mouvement pour la constitution d'archives ouvertes (en particulier dans le cadre de *Open Language Archives Community*, OLAC, fondé en 2000) était d'abord tournée vers l'opérationnel : la mise en réseau, au plan mondial, d'archives numériques ouvertes. Si l'on s'arrête sur un document fondateur (Simons et Bird 2000), on peut relever que les questions juridiques et éthiques ne figuraient pas parmi les « sept piliers » identifiés, agencés de la manière suivante : utilisatrices et utilisateurs veulent des *données*, des *outils*, des *conseils* ; le réseau OLAC leur offre un *portail* d'accès unifié, le format partagé de *métadonnées* qui permet cette mise en réseau, une *évaluation* par les pairs, et une *standardisation* (des normes et protocoles qui permettent de s'assurer de la qualité des métadonnées). Bref, l'ouverture des données était un « mouvement à sens unique », pour reprendre l'expression que Frédérique Bordignon applique au mouvement en faveur du libre accès aux publications (Bordignon 2023).

Une certaine défiance vis-à-vis de l'approche nord-américaine des questions d'éthique de la recherche

Le cadre juridique indispensable à l'activité de publication de données était en place dans l'archive ouverte, mais n'apparaissait aux chercheurs que de façon discrète, voire marginale. À l'étape de la mise en forme des métadonnées pour le dépôt, le chercheur déposant apprenait être identifié comme détenteur du droit de copie *(copyright)* ; choix valorisant et rassurant, en cohérence avec le rôle central qu'accorde le CNRS au chercheur, « tête pensante » autour de laquelle gravite l'institution, et qui bénéficie de multiples services assurés par les personnels d'*appui à la recherche*. On se voyait proposer un choix quant à la licence sous laquelle étaient placés les documents déposés. Ce choix fournissait l'occasion d'un premier niveau d'information. (La collection Pangloss et la plate-forme Cocoon qui l'héberge ont très tôt recommandé les licences Creative Commons.) Le processus de publication des données ne comportait pas d'étape de formalisation par écrit de l'engagement pris par la déposante ou le déposant. Cela revenait à lui faire confiance pour les questions juridiques et éthiques liées à la collecte et la publication de données de terrain.

Les pratiques sur le terrain étaient conservatrices, et teintées d'une nette défiance vis-à-vis des procédures de validation éthique (par un *Institutional Review Board* ou *Research Ethics Review Committee*) dont on savait qu'elles avaient cours dans le contexte nord-américain. On devinait bien que l'Europe finirait par suivre le mouvement, mais on renâclait à l'idée de devoir emboîter le pas à une Amérique qui, en matière de morale, ne manifestait pas un grand sens de la nuance.

> La Russie n'apportait plus d'espoir ni de peur, rien d'autre qu'une désolation perpétuelle. Elle s'était retirée de notre imaginaire – que les Américains occupaient malgré nous, comme un arbre gigantesque étalant ses branches à la surface de la terre. Ils nous énervaient de plus en plus avec leur discours moral, leurs actionnaires et leurs fonds de pension, leur pollution de la planète et leur dégoût de nos fromages. (…) Des conquérants sans idéal sinon le pétrole et les dollars. Leurs valeurs et leurs principes – ne compter que sur soi – ne donnaient d'espérance à personne d'autre qu'eux et nous rêvions d'« un autre monde ». (Annie Ernaut, *Les Années*, 2008, p. 209)

Les « autres mondes » culturels, la linguistique de terrain a précisément pour objet d'aller à leur rencontre, et de prendre la mesure de leur diversité socio-politique, aisément méconnue (Woodbury 2011). Or une réglementation en matière d'éthique de la recherche ne peut que refléter des principes (culturellement relatifs) qui ont cours dans le pays où elle a été élaborée, et ne peut donc qu'être en décalage plus ou moins profond avec les situations concrètes du terrain. Les réactions de locutrices et locuteurs prié·es de signer des formulaires de consentement éclairé approuvés par un Comité d'éthique de la recherche (les exemples que j'ai en tête, employés au Vietnam, provenaient d'Allemagne et des États-Unis) me paraissaient éloquentes : le document (parfois une liasse comportant plus de mille mots) n'était pas lu, tout au plus rapidement survolé. La signature était précédée d'une brève explication au sujet de ce que représentait le document et de son caractère indispensable. La personne appelée à signer réagissait parfois par un simple regard (comme un clin d'œil de connivence : *allez on est sympa, on te le signe*), parfois par une brève phrase pour dire les choses, et relever le caractère incongru voire vaguement inquiétant de l'obligation de signer *(C'est bien long tout ça ! Bon, tiens, voilà)*. La bonne foi des collègues n'est pas en cause (elles et ils étaient tenus de suivre la procédure imposée par leur institution de rattachement), ni la qualité du protocole expérimental. Mais l'étape du consentement constituait clairement une formalité, perçue comme telle. Cela faisait penser à du *consent washing* (ou *ethics washing*), comme le *greenwashing* qui donne une image de responsabilité écologique pour pas trop cher.

Dans le contexte de cultures à tradition orale, il pouvait paraître plus respectueux d'éviter aux personnes qui nous enseignent leur langue et nous autorisent à recueillir leur parole l'intrusion de pratiques telles que le recueil d'un consentement écrit pour la collecte et la diffusion de données.

Faire signer une personne qui n'a pas la pratique de l'écrit peut susciter une anxiété à la mesure des craintes liées aux documents légaux, qui renvoient à la loi de l'État-nation, à laquelle est soumise la société étudiée (et qui ne fait pas toujours bon ménage avec le droit coutumier). Dans les localités du Yunnan où s'est déroulé mon terrain, l'écrit n'est pas inconnu : les enfants sont scolarisés en chinois, et chaque citoyen·ne est amené·e à signer des documents à certaines étapes de la vie, de sorte que l'on peut argumenter qu'il n'y a rien de véritablement exotique à faire signer un consentement éclairé. Mais les écrits légaux dont est familière la culture locale engagent des décisions importantes, de sorte que l'acte de signer un contrat comporte des connotations culturelles appartenant à un registre bien éloigné de celui qui est souhaitable pour une enquête linguistique. La sagesse populaire de là-bas comme d'ici suggère, on le sait, qu'il est bien préférable de n'avoir jamais affaire à la justice.

Les protocoles éthiques peuvent même manifester des tendances impérialistes, comme le relevait il y a près d'un demi-siècle un spécialiste de bioéthique. « I think the amateurs in ethics suffer in imperialism. They practice imperialism, which is constituted by establishing a rule and applying it ruthlessly. You know, wherever you go, you set your rule down. What we ought to be particularly good at by now is seeing that rules have very different meanings as situations approach or depart from a kind of a paradigm situation », relève Albert Jonsen, cité par Zachary Schrag (2010, 7). Semblables réflexions peuvent amener à rejeter en bloc l'idée d'une formalisation des pratiques en matière d'éthique de la recherche. Les protocoles éthiques seraient en décalage avec les réalités du terrain, au point d'être délétères et mensongers.

Ce point de vue est exposé dans un billet d'humeur écrit au vitriol : « La recherche sur les langues en danger et la dépravation morale des protocoles éthiques » (van Driem 2016). De même que, selon la formule pascalienne, *la vraie morale se moque de la morale*, l'éthique de la science se moquerait de l'éthique formaliste, dont la fonction première (dans un contexte nord-américain) serait tout bonnement – loin des préoccupations déontologiques affichées – de protéger l'institution contre le risque de coûteux procès qui lui soient intentés (Bowern 2010, 900).

La responsabilité éthique, la vraie, consiste à entretenir les meilleures relations avec les consultants linguistiques : adopter un comportement bien adapté au contexte culturel, respecter les consultants, valoriser les connaissances qu'ils partagent, leur expliquer l'ensemble du processus de documentation et de recherche, leur donner une juste compensation pour le temps qu'ils consacrent au travail, leur permettre d'accéder à la documentation linguistique produite ainsi qu'aux productions du chercheur. Et les associer au processus de documentation et de recherche, avec l'espoir que l'informateur (comme on disait avant la généralisation du terme plus valorisant de consultant linguistique) devienne un collaborateur et, idéalement, un collègue. Tout cela est exposé dans un ouvrage de référence au sujet de la méthode d'enquête sur les langues et civilisations à tradition orale (Bouquiaux et Thomas 1971), ouvrage dont on m'a remis un exemplaire lorsque j'ai rejoint le laboratoire de Langues et civilisations à tradition orale (LACITO, fondé en 1976), et que j'ai, depuis, recommandé à mon tour à des doctorantes et doctorants.

En outre, l'étude de langues rares dans des pays dont les politiques assimilationnistes (Gros 2014) visent, au plan linguistique, à la généralisation de la langue nationale, ne pouvait pas facilement prendre le tournant collaboratif et communautaire dont des projets nord-américains ou australiens donnent l'exemple (Snead et Cushman 2023).

> Depuis deux ou trois décennies, les préoccupations éthiques prennent une importance croissante dans le travail des linguistes. Il est en effet difficile de ne considérer que sous l'angle de l'intérêt intellectuel l'étude d'une langue en danger de disparition, ou au moins socialement dévalorisée, et dont les locuteurs sont parfois déplacés de leur lieu et de leur mode de vie traditionnels, fragilisés par la rupture avec leurs traditions culturelles, et relégués dans les marges défavorisées de la société. Des enquêtes internationales – malheureusement peu relayées en France – menées sur des populations autochtones d'Amérique, d'Australie ou des pays arctiques, montrent que

> la perte des langues ancestrales est l'un des éléments d'un mal-être général, qui se traduit aussi par des problèmes de santé (diabète, obésité, maladies cardiaques) et des comportements à risques comme l'alcoolisme, l'addiction aux drogues, le suicide, la délinquance, alors que le maintien des langues et des références culturelles réduit considérablement ces dangers. Il est difficile pour les linguistes de rester indifférents à ce constat, ou même de s'en tenir à une empathie compassionnelle, sans s'interroger sur la place qu'ils occupent dans un tel contexte, et sur les moyens d'y jouer un rôle positif. (Launey 2023, 730-731)

Dans des pays qui ont fermement exclu la mise en place de filières de scolarisation en langue maternelle (pourtant très positives au plan éducatif), il n'aurait pas été réaliste de se fixer des objectifs ambitieux en termes de revitalisation linguistique. La relative discrétion dans laquelle le travail se faisait limitait les échanges au sujet du projet documentaire et scientifique avec des membres de la communauté linguistique concernée. Or le fait que des locutrices et locuteurs soient partie prenante au projet de documentation linguistique figure parmi les principaux moteurs de réflexion sur les questions d'éthique, au sens large (Vapnarsky 2020 ; Marsault 2021, 27-38 ; Cox 2022). Le choix du nom donné à une langue (Bichurina 2017, 500-501 ; Michaud 2017c, 496-501), les décisions prises au stade crucial que constitue la mise à l'écrit d'une langue à tradition orale, comportent toutes sortes d'enjeux de politique linguistique, qui engagent la définition même de la communauté (indigène, autochtone, nation première – *Indigenous, Native, First Nations, Aboriginals* –, dans ses divers avatars juridiques au fil du temps), et soulèvent la question de la légitimité de telles ou telles personnes, en tant que membres, en tant que représentants, en tant que chercheurs « accrédités »… Ces thématiques, délicates à gérer (Launey 2023, 731-732), mettent les questions éthiques sur le devant de la scène. La critique du travail en mode non collaboratif – désigné par l'appellation péjorative de travail de « loup solitaire » : *lone wolf linguists* (Austin 2007), *lone ranger linguistics* (Dwyer 2006) – suscite en retour une mise en garde contre la dérive vers une interprétation trop étroite du principe de collaboration, qui risquerait de devenir un lit de Procuste.

> We take issue not with collaboration per se, but with the viewpoints that linguists practicing language documentation must collaborate with the community, that the linguist's goals should be subordinate to the goals of community members, or that solo research is necessarily unethical research. (Crippen et Robinson 2013, 124)

Ce contexte amène à des prises de position circonstanciées de la part des praticiennes et praticiens de la linguistique de terrain et de l'étude des langues rares. Ainsi, l'autrice et l'auteur d'une édition trilingue inuktitut-anglais-français de l'ouvrage *Chasseur au harpon*, de l'écrivain inuk Markoosie Patsauq, ont été amenés à clarifier par écrit les principes qui guident leur travail.

> *Taking Inuktitut seriously*
>
> What we mean by this is actually studying the language (common practice is to multiply incantatory statements about the value of indigenous languages, while working solely with English) and taking into consideration the power relations within which this language and its speakers function; simultaneously one must recognize the risks of doing this work as non-Inuit scholars, the criticism to which one is necessarily exposed (cultural appropriation, colonialism, etc.), and the need to respond to these. In all our work, we have underscored the value of respecting Inuktitut as worthy of critical appreciation; acknowledging its capacity to serve as a tool for expression, including literary, rather than simply an ethnographic curiosity; and collaborating with community members. (Henitiuk et Mahieu 2024, 3-4)

Mes premières enquêtes linguistiques sur le terrain (au Yunnan, dans le sud-ouest de la Chine), dans les années 2000, se déroulaient bien loin de ces préoccupations. Rétrospectivement, il ressort que ma pratique d'une ouverture des données « à sens unique » (plutôt que dans un processus collectif, réflexif et continu) n'était pas sans lien avec les parois de verre qui maintenaient une

certaine distance entre la communauté des locutrices et locuteurs – dont le représentant légitime, dans un État à parti unique, est un organe de ce parti – et le processus de constitution et exploitation des ressources produites. Sur le terrain, je ne faisais pas signer de document par les locutrices et locuteurs. Lorsque je me suis finalement essayé à recueillir un consentement sur le terrain, l'essai ne m'a pas paru concluant.

Autorisation de diffusion et consentement oral : un essai peu concluant

En dépit de réticences à céder à ce qui paraissait une mode d'outre-Atlantique, j'étais désireux de disposer de tous les arguments possibles en faveur du libre accès que je souhaitais pour les données de langue na (mosuo) sur laquelle je souhaitais mener une étude approfondie (travail qui se poursuit aujourd'hui, près de vingt ans après le début de cette collaboration). J'ai donc entrepris, au début des années 2010, de recueillir un consentement, par écrit et par oral, de la personne qui m'enseignait sa langue (le na de Yongning) depuis 2006.

Pour le document écrit (en chinois, le na étant une langue à tradition orale), un membre de la famille a signé à la place de la locutrice (qui n'écrit ni ne lit), pis-aller qui souligne la fragilité de l'exercice (la signature n'étant pas celle de l'intéressée), et qui cause de l'embarras en mettant en relief le décalage entre les compétences de la personne, qui n'a pas été « écolée »[4], et les normes de l'État-nation – alors qu'il importe au contraire de valoriser les savoirs dont la personne est détentrice, et qu'elle accepte de partager.

L'enregistrement d'un consentement oral peut en principe constituer une solution, en replaçant le propos dans l'espace de l'oralité et de la langue étudiée. Mais la mise en œuvre révèle d'autres fossés culturels. Je sentais bien que je forçais la main à la locutrice en lui demandant d'exprimer par oral son avis sur la question de la diffusion des données. Je n'ai pas renoncé au projet pour autant, car j'éprouvais la crainte qu'en l'absence d'un tel consentement oral, il ne me soit, à plus ou moins long terme, reproché d'avoir négligé un principe : de m'être tenu à l'écart des « bonnes pratiques » internationales en matière de documentation des langues. De fait, à la date de rédaction du présent texte (2024), la charge de la preuve incombe au chercheur, qui doit disposer d'éléments établissant (par exemple auprès d'une revue à laquelle on soumet un article reposant sur des données collectées sur le terrain) que la collecte de données répond aux exigences éthiques en vigueur. Un adage romain veut, paraît-il, qu'en matière de droit, *idem est esse aut non probari* : ne pas être, ou ne pas être prouvé, c'est tout un.

Mon idée était d'enregistrer un court message destiné aux auditeurs des enregistrements mis en ligne en accès public. La locutrice marquerait, à ce public absent, qu'elle est d'accord pour que n'importe qui puisse écouter les enregistrements et en savoir plus sur sa langue maternelle. Cet exercice imposé était plus difficile, pour mon enseignante de langue na (mosuo), que l'ordinaire des séances dont nous avions pris l'habitude. Lorsqu'elle enregistrait des récits, elle s'adressait à moi, sachant qu'un public beaucoup plus large aurait accès à l'enregistrement. Tandis que pour cet enregistrement de consentement oral, il lui était demandé de parler de notre collaboration, en ma présence, mais en s'adressant, dans sa propre langue, à un public abstrait, qui s'intéresserait, non pas à la langue et la culture, mais au processus d'enquête et à la relation de travail que nous avions liée.

Je suis particulièrement reconnaissant à ma prof d'avoir accédé à cette drôle de demande, qui lui avait été expliquée par moi-même et par son fils. Ce qu'elle a finalement dit[5] est, en substance : « Voilà maintenant trois, quatre ans que Diddeo [le nom qui m'a été donné en langue na] est venu chez moi. C'est mon fils qui le connaissait ; mon fils m'a présenté Diddeo pour que je lui enseigne la langue na ; je lui ai dit tout ce que je savais ; est-ce qu'il est calé ou pas, ça, c'est vous qui savez.

4 J'emprunte ce terme désuet à une grand-mère tenue à l'écart de l'école, et qui mesurait l'importance que les femmes fussent « écolées ».

5 Le document est disponible ici : https://doi.org/10.24397/pangloss-0004611. Chacune des ressources de la collection Pangloss bénéficie d'un identifiant DOI : voir Vasile et al. (2020).

Depuis trois ou quatre ans qu'il étudie, il comprend un peu la langue ; moi je parle, lui il écrit, jour après jour ; il est consciencieux, c'est quelqu'un de calé. Avec les enregistrements, vous êtes en mesure d'écouter, et de voir si vous arrivez à comprendre tout ça ». Ce message est exprimé dans une langue que bien peu de gens peuvent comprendre ; il faudrait le transcrire, mais je n'ai pas voulu faire le travail ordinaire de vérification avec la locutrice, ne souhaitant pas me replonger avec elle dans ce discours qui par certains aspects flotte dans le vide, et qui par d'autres est très personnel, puisqu'il met en scène nos deux personnes et la relation qui s'est établie. L'enregistrement, qui comporte par endroits une discrète note d'humour, pourrait être considéré comme trop éloigné du contenu attendu pour remplir la fonction souhaitée : établir la conformité de l'enquête linguistique avec la réglementation en matière d'éthique. Je pourrais reprendre l'échange sur ce thème avec la locutrice et sa famille, afin d'enregistrer une autre déclaration, plus conforme à tel ou tel modèle, mais cela reviendrait à restreindre encore la liberté d'appréciation qui lui est laissée dans la formulation du propos qu'on lui fait tenir.

En un mot, l'essai de recueil d'autorisation de diffusion paraissait non concluant : l'éthique réglementaire ne semblait décidément pas faire bon ménage, en pratique, avec l'éthique de la recherche.

Numérisation de sauvegarde de fonds anciens : priorité à l'ouverture

Dans le travail de sauvegarde de données anciennes (des fonds irremplaçables, qui constituent la base empirique de la recherche, mais qui dans bien des cas disparaissent avec leur auteur), la priorité allait pareillement à l'ouverture des données, plutôt qu'à l'approfondissement des questions juridiques et éthiques soulevées par l'entrée des données dans le monde numérique. Ainsi de la numérisation des collections d'enregistrements audio de Michel Ferlus, spécialiste des langues d'Asie orientale, réalisée à Hanoï avec un financement octroyé par le Comité pour la Science ouverte (anciennement Bibliothèque scientifique numérique) dans le cadre d'un Appel à projets « Numérisation du patrimoine » (Michaud 2017b). Un des critères d'admissibilité des projets était une mise en ligne, en libre accès, des documents numérisés, au plus tard à la date de fin du projet. Cette exigence clairement affichée a contribué à convaincre Michel Ferlus, initialement réticent à l'idée d'ouvrir au public des enregistrements qui à l'origine étaient réalisés comme simple documents de travail (Ferlus, 2017), de donner son accord à la diffusion de l'intégralité des données sous licence Creative Commons (CC BY-NC-SA 3.0 fr)[6]. La question du consentement des personnes concernées paraissait anachronique, s'agissant de données collectées depuis les années 1960, sur plus de trente langues de la péninsule indochinoise (Laos, Vietnam, Cambodge, Birmanie), généralement dans des localités enclavées. Recueillir un consentement *a posteriori* paraissait infaisable, et au fond assez futile en comparaison de l'enjeu de la conservation et l'ouverture des données.

Ce sont des échanges avec des spécialistes du droit qui ont amené un profond changement de perspective. Ils m'ont permis, non seulement de mesurer mon ignorance au sujet de thèmes juridiques aussi fondamentaux que le droit de copie (le *copyright* au sens légal du terme), mais de comprendre que le droit et l'éthique sont à voir comme des alliés (pour paraphraser l'intitulé d'une Journée d'étude qui s'est tenue en 2019[7]).

6 Cette observation fait écho à celle effectuée par Joséphine Simonnot au sujet des numérisations effectuées avec le soutien de la Mission Recherche et Technologie (MRT), qui avait fixé pour politique que les financements de la numérisation devaient s'inscrire dans le cadre de projets de diffusion au public (Simonnot 2020, §3).

7 La journée d'étude « Diffuser les données numériques en SHS : le droit et l'éthique comme alliés », organisée par le groupe de travail Éthique & Droit du Comité pour la Science ouverte, avec le soutien de l'URFIST Méditerranée, a été organisée le jeudi 3 octobre 2019 à la Maison méditerranéenne des sciences de l'homme (MMSH) d'Aix-en-Provence.

Un changement de perspective : le droit et l'éthique comme alliés pour l'ouverture des données de la recherche en sciences humaines et sociales

À partir du milieu des années 2010, j'ai eu la chance d'avoir des occasions d'échanges avec des juristes : tout d'abord avec Danièle Bourcier, introductrice en France des licences Creative Commons dont elle est responsable scientifique, puis, grâce à elle, avec Lionel Maurel, juriste de formation, conservateur des bibliothèques, qui s'intéresse aux biens communs, à la culture libre, au domaine public. Ces échanges m'ont conforté dans le choix d'ouvrir les données de la recherche ; ils m'ont en outre permis de mesurer que la posture qui consiste à se raidir contre le principe de recueil de consentement comporte des travers. Pour le chercheur, l'attitude paternaliste qui consiste à se poser en protecteur d'une communauté fragile (indigènes, peuple autochtone, minorité linguistique, groupe social marginalisé…) est propice à des dérives : s'arroger sur les gens une autorité morale, leur imaginer un devoir de gratitude proportionnel aux bontés qu'on estime avoir pour eux, se bâtir un fief scientifique (chasse gardée) dont on exclut les chercheurs qui viendraient menacer son monopole. Et invoquer le principe abstrait de protection de la vie privée pour légitimer le fait de garder par-devers soi des enregistrements, au risque qu'ils soient, au final, tout bonnement perdus. Certes, la parole est quelque chose de personnel, qui touche à l'intime, et les enregistrements sont recueillis dans le cadre d'une relation de confiance personnelle avec l'enquêteur. Pour autant, il n'est pas du tout évident que leurs auteurs aient souhaité limiter à cette unique personne la circulation d'enregistrements qui font partie de leur contribution à une entreprise visant aux progrès de la connaissance. Jean Malaurie relate que son compagnon Qaaqqutsiaq « a contribué à résoudre tous les problèmes et difficultés de l'expédition à laquelle il était fier de participer, avant tout parce qu'elle était scientifique » (Malaurie 1989, 449). Participer à un voyage de scientifique *(ilisimasassarsiorneq)* grandissait les participantes et participants inuit aux yeux des leurs et à leurs propres yeux (Malaurie 1989, 485). Les personnes que l'on enregistre ne méritent pas moins que nous d'être reconnues pour le travail réalisé, en lui-même souvent moins gratifiant pour elles que pour nous[8].

Si un contexte propice se présente, il n'y a rien d'inconvenant à avoir, pendant une enquête linguistique de terrain, des conversations au sujet de thèmes juridiques tels que la fonction et l'utilité des licences Creative Commons sous lesquelles on prévoit de placer des enregistrements. Au contraire, on contribue par là au partage d'une information juridique qui n'est pas sans valeur pour les citoyens de pays où ces licences, transposées en droit local, font partie du paysage contemporain ; et on facilite un débat, non vertical, au sujet de la question du partage des données. Aux antipodes d'un contrat de cession de droits patrimoniaux à titre exclusif, par lequel on prive la locutrice ou le locuteur de tous droits sur un document, les licences Creative Commons fournissent un cadre légal propice à une collaboration saine et équilibrée.

Les réflexions ainsi ouvertes se poursuivent depuis, dans un contexte institutionnel où les pratiques de Science ouverte sont désormais encouragées et soutenues.

De 2016 à nos jours : vers une généralisation des pratiques de Science ouverte

Dans la seconde moitié de la décennie 2010, la Loi pour une République numérique changeait la donne en matière de données de la recherche : l'ouverture des données ne serait plus un choix personnel militant, mais un devoir professionnel partagé (Arènes, Maurel et Rennes 2022). C'était là un bouleversement considérable : la carotte et le bâton allaient, bien mieux que les arguments de principe au sujet des vertus de la Science ouverte, convaincre enfin l'ensemble des chercheuses et chercheurs d'élaborer un plan de gestion des données, et de prendre le chemin de la publication

8 La personne qui m'a enseigné la langue na de Yongning est présentée dans le livre que j'ai écrit au sujet du système tonal de la langue (Michaud 2017c, 29-33). De tristes circonstances m'ont également amené à écrire un mot au sujet de ma première consultante de langue naxi (Michaud 2017a).

électronique des données de terrain. Documentation et recherche allaient enfin progresser de façon solidaire. Au lieu de dérouler un discours solitaire au sujet du rôle des chercheurs dans l'archivage et la diffusion des données de la recherche (Michaud 2002), il devenait possible de reformuler le même constat et les mêmes recommandations sur le mode du « puisque la réglementation impose désormais l'ouverture des données, autant que le processus soit bien accompagné, plutôt que de s'arc-bouter inutilement pour le freiner ». Une équipe de douze autrices et auteurs se constituait pour exposer leur vision d'une transition « Vers des politiques de données ouvertes dans les sciences phonétiques : ce que le domaine y gagne, et comment éviter les écueils ».

> …the phonetic sciences stand to gain greatly from data availability and citability (…) the present argument is made by insiders offering a critical introspective look, not as sniping outsiders. Our hope is to facilitate a base level of common understanding so that the field can deal with these core issues actively and manage ongoing transitions tactfully, rather than passively letting changes happen around us and belatedly realizing that data have evaporated and many research articles in phonetics play to our ears "ditties of no tone", like the silent Grecian urn in John Keats's *Ode* (Garellek *et al.* 2020, 3-4).

L'historique éditorial de cette prise de position collective, publiée dans une jeune revue brésilienne (en libre accès diamant) après avoir été rejetée par une revue-phare du domaine des sciences phonétiques, rappelait que presque rien n'avait bougé en deux décennies. La transition ne se ferait pas du jour au lendemain, mais c'était une raison de plus d'y aller de bon cœur. Il y avait du pain sur la planche, et il gardait une saveur de militantisme.

Inscription au registre des bases de données du CNRS : échanges avec la Déléguée à la Protection des Données

L'une des tâches à effectuer consistait à mettre la collection Pangloss en conformité avec la réglementation, par son inscription au registre des bases de données du CNRS. Les échanges avec Gaëlle Bujan, Déléguée à la Protection des Données du CNRS, ont fait ressortir des éléments rassurants. La licéité du traitement se fonde sur le fait qu'il est mis en œuvre dans le cadre de l'exécution d'une mission de service public. La sensibilité des données de la collection Pangloss est jugée Normale (les deux autres catégories possibles étant Sensible et À risque), et il n'y a pas d'obstacle de principe à la publication du nom des personnes ayant contribué à la création des ressources (à commencer par les locutrices et locuteurs – lorsqu'il n'y a pas de contre-indication,

cela va de soi : les métadonnées sont anonymisées si le contexte socio-politique le demande). Il n'y a donc pas d'obligation d'anonymisation des personnes, qui entrerait en conflit avec la politique de la collection de reconnaître le rôle de chacun (à commencer par les locutrices et locuteurs) dans la constitution des ressources orales[9].

Ces éclaircissements au plan juridique permettent non seulement de mettre l'archive en règle avec la législation (le Règlement général sur la protection des données du 24 mai 2016), mais aussi de fournir une information fiable aux collègues (dont les doctorantes et doctorants) qui montent un projet de recherche et doivent dorénavant le soumettre à l'examen du Comité d'éthique de la recherche de leur établissement. Le document d'inscription au registre du CNRS contient en effet des formulations en termes juridiques, validées par l'institution, qui constituent autant d'éléments de réponse à des questions qui se posent régulièrement dans le cadre de projets de recherche sur des langues rares.

9 Cette perspective rejoint celle de l'ethnographie contemporaine : « [l']évolution des cadres éthiques du métier d'ethnographe de ces dernières décennies (Caplan 2003 ; Fluehr-Lobban 2013) assigne en effet aux communautés le statut de copropriétaire des données résultant de la relation ethnographique, ce qui ouvre la voie à une coresponsabilité de leur préservation et de leur partage entre chercheur et communauté étudiée – et les incite à trouver les moyens de son exercice » (Heintz 2023).

Le séminaire « Science ouverte », lieu de formation et d'échange

Un séminaire « Science ouverte »[10] organisé par un collectif multi-métiers du campus CNRS de Villejuif en 2020-2021 et 2021-2022 visait à faire entrer de plain-pied les doctorantes et doctorants (et toutes autres personnes intéressées) dans les pratiques de Science ouverte, retour aux sources de l'ethos de la science[11]. L'Université Sorbonne Nouvelle, informée du projet, proposait qu'il soit mutualisé par l'ensemble des Écoles doctorales de l'établissement ; le séminaire était également accrédité par l'École doctorale de l'INALCO. Tous les doctorants de sciences humaines et sociales du campus de Villejuif pouvaient participer comme auditeurs libres, quelle que soit leur École doctorale de rattachement. Mastérants, doctorants, chercheurs, ingénieurs, collègues du service juridique de la Délégation CNRS se retrouvaient pour des séances dont la moitié était consacrée à une discussion libre, succédant à un exposé donné par des praticien·nes expérimenté·es. L'argumentaire du séminaire, calquant une formule de Rousseau, annonçait le choix d'allier toujours ce que la Science ouverte recommande avec ce que l'intérêt prescrit, afin que bonnes pratiques et perspectives de développement de carrière ne se trouvent pas divisées[12]. Est-ce que l'adoption de pratiques de Science ouverte constituerait pour les doctorantes et doctorants un atout professionnel, ou est-ce que l'obtention de leur Passeport pour la Science ouverte ouvrirait seulement la petite porte d'emplois précaires et ancillaires, comme gestionnaires des données d'autrui ? Les avis étaient partagés, et notablement différents d'une discipline à une autre. On convenait de laisser les intéressés faire leurs choix librement, au vu de l'information dispensée, et d'un contexte mondial qui se dégradait à vue d'œil. Des collègues attaché·es aux principes de l'ancien temps (l'esprit critique universitaire, sur fond de tolérance, d'ouverture et de liberté de pensée) étaient en butte à des attaques médiatiques (Eslen-Ziya, Giorgi et Ahi 2023) que légitimaient les déclarations de ministres de la République en exercice. Un président de la République disait tout haut ce qu'il pensait des « gens qui ne sont rien ». On pointait son erreur à notre manière, convaincue mais dérisoire, ou dérisoire mais convaincue (Michaud, Nguyễn et He 2020). *Médiapart* résumait l'état des services publics, malmenés de longue date, par la formule « après l'alerte, la dégringolade ». L'exécutif ne faisait plus semblant d'épargner à l'enseignement et la recherche l'application de politiques d'austérité sans fin.

À la différence de l'engagement écologico-social, lui aussi fondé sur un constat scientifique clair (Ripple *et al.* 2017), mais qui s'inscrivait sur fond de désespoir – « When hope dies, action begins » (Kaufmann *et al.* 2019) – et de répression politique méthodique (Middeldorp et Le Billon 2020), le militantisme pour la Science ouverte avait le soutien de l'institution, à tous les niveaux : universités, organismes de recherche, ministères, jusqu'à l'Union Européenne. Les sciences du langage figuraient parmi les disciplines bien avancées dans leur transition, fortes de revues et maisons d'édition en libre accès diamant (sans frais pour les autrices ni les lectrices), comme Glossa (Rooryck 2016) et Language Science Press[13], et de plate-formes spécialisées pour les données, telles que Cocoon[14]. On était invité à raconter sur un blog institutionnel ses aventures au

10 Voir : https://himalco.hypotheses.org/

11 L'expression est empruntée à Calimaq : https://scinfolex.com/2019/06/05/louverture-des-donnees-de-recherche-un-retour-aux-sources-de-lethos-de-la-science/

12 « Je tâcherai d'allier toujours, dans cette recherche, ce que le droit permet avec ce que l'intérêt prescrit, afin que la justice et l'utilité ne se trouvent point divisées. » Jean-Jacques Rousseau, *Du contrat social, ou Principes du droit politique*, livre premier.

13 Voir notamment : https://userblogs.fu-berlin.de/langsci-press/2018/07/11/what-it-means-to-be-open-and-community-based-the-unicode-cookbook-as-a-showcase/

14 Cocoon, pour *COllections de COrpus Oraux Numériques*, est une plate-forme hébergeant plus de trente collections d'enregistrements de parole (y compris les langues des signes), principalement dédiées à la recherche et la médiation scientifique, plutôt qu'à des usages commerciaux. La plate-forme Cocoon est donc complémentaire, et non concurrente, de plate-formes comme LDC (Linguistic Data Consortium) et ELRA (European Language Resources Association), tournées vers les grandes langues de communication et vers les corpus constitués dans le cadre de campagnes d'acquisition de données pour le Traitement Automatique des Langues. Voir : https://cocoon.huma-num.fr/

pays du libre accès[15]. Le dépôt dans HAL ne faisait plus débat : à partir de 2020, seules les publications présentes dans HAL pouvaient être signalées dans le rapport annuel d'activité CNRS. Un Datathon pour le dépôt, l'archivage et la diffusion de corpus oraux (linguistique, socio-linguistique, anthropologie, histoire orale)[16] constituait un moment privilégié pour un dialogue autour des données – dialogue qui s'avère généralement fertile en idées et en projets. Comme le relève Monica Heintz, relatant les effets induits par un état des lieux des données de la recherche au sein d'un labo d'anthropologie : « Certains chercheurs se sont aperçus qu'ils avaient encore des chantiers en cours, des données structurées mais non déposées, que certains corpus donnaient lieu à plus de publications que d'autres, qu'il y avait des formes de soutien insoupçonnées de la part des ingénieurs de recherche qu'on n'avait pas toujours su exploiter, etc. » (Heintz 2023, 244).

Entretenir un niveau minimum d'information juridique : une composante importante du soin apporté aux données de la recherche

Une leçon de l'étape de l'inscription au registre du CNRS, c'est qu'il importait, pour prendre soin[17] efficacement des données de la recherche (et encourager une culture des données dans les labos), d'acquérir, et de maintenir à jour, un niveau minimum d'information juridique autour des données de la recherche. Ce point de vue est solidaire de l'idée (suggérée par Lionel Maurel) selon laquelle, la loi européenne ayant fait le choix d'un niveau élevé de protection des personnes (données personnelles, vie privée…), satisfaire aux obligations légales (dans le respect de l'esprit des lois), est en soi un objectif exigeant, au point qu'il n'est pas évident qu'il soit nécessaire de lui ajouter une « surcouche » d'éthique formelle.

Cette perspective paraît tout à fait centrale pour le déploiement d'un modèle alternatif des comités d'éthique de la recherche, dans lesquels la réflexivité, plutôt que des procédures standardisées, soit au centre de l'éthique de la recherche (Gagnon 2010). Dans la recherche d'« un juste équilibre entre cette réflexivité et les nécessaires procédures à mettre en place » (Bazin et Goiseau 2023, 73), les exigences juridiques sont incontournables et doivent être intégrées (Bazin et Goiseau 2023, 93). En répondant à ces exigences de façon claire, nos disciplines se trouvent bien placées pour défendre leur point de vue, contre une bureaucratisation de l'éthique de la recherche.

La définition juridique de la notion d'« éditeur » *(« publisher »)* fournit un exemple de l'utilité d'un premier niveau d'information juridique pour une bonne compréhension des enjeux liés aux données de la recherche. En effet, les trois dimensions de la Science ouverte que constituent les publications, les données et les outils ne sont pas régies par la même réglementation, et en matière de droit, le bon sens à lui seul ne mène pas bien loin. Par exemple, quand on sait que le CNRS déconseille aux unités de recherche de jouer le rôle de maisons d'édition, on peut être perplexe de voir le labo de rattachement figurer dans le champ « éditeur » *(« publisher »)* pour les corpus de la collection Pangloss.

Pour tirer au clair les questions délicates de ce type, on peut bénéficier d'une information en ligne[18], et, pour les unités de recherche du CNRS, de l'aide d'interlocuteurs experts du service juridique de la Délégation CNRS et du pôle Science ouverte du CNRS – Sciences Humaines et Sociales (anciennement – jusqu'en 2023 – « Institut des Sciences Humaines et Sociales du CNRS »). Ainsi, les explications reproduites ci-dessous ont été fournies par Lionel Maurel

15 « Au pays des merveilles du libre accès : la préparation d'un premier ouvrage chez Language Science Press », Carreau de la BULAC (carnet de recherche de la Bibliothèque universitaire des langues et civilisations), 2017. https://doi.org/10.58079/m5ki

16 « Datathon de la parole, 8-10 novembre 2021 : dépôt, archivage et diffusion de documentation linguistique sur langues rares ». Les Carnets du LACITO. https://doi.org/10.58079/qpc8 Deuxième édition en 2023, troisième prévue en 2024, au DataLab de la Bibliothèque nationale de France (au sujet duquel on consultera la présentation de Carlin et Laborderie 2021).

17 L'expression « prendre soin » est un clin d'œil à l'ouvrage *Prendre soin : de l'informatique et des générations* (Alombert et al. 2021), en hommage à Bernard Stiegler.

18 Notamment la documentation produite par le Centre pour la Communication Scientifique Directe, disponible sur son site et sa chaîne Canal-U.

(Directeur Adjoint Scientifique en charge des questions de Science ouverte, d'édition scientifique et des données de recherche) dans un échange courriel à l'été 2023.

Les publications scientifiques sont régies par la loi sur la liberté de la presse, qui est articulée autour de la notion d'éditeur, lequel assume la responsabilité juridique liée à la publication et doit donc disposer pour cela d'une personnalité juridique. Une unité de recherche (Unité Mixte de Recherche ou Unité d'Appui et de Recherche), n'étant pas dotée de la personnalité juridique, ne peuvent structurellement jouer le rôle de maisons d'édition scientifique. Pour les données, en revanche, la notion pertinente n'est pas celle d'éditeur, mais plutôt celle de *producteur* (au sens d'*entité qui apporte les moyens nécessaires à la production des données*). En dernière analyse, les producteurs des données de recherche sont les établissements de recherche. Or les métadonnées de la plate-forme Cocoon (qui héberge notamment la collection Pangloss) recourent à la terminologie Dublin Core[19]. Dans ce cadre terminologique, c'est l'unité de recherche qui apparaît comme éditeur (« publisher »). Le Dublin Core, développé à la base pour cataloguer des publications scientifiques, et non des données, mobilise une notion de *« publisher »* qui est, au fond, inappropriée. Juridiquement, Huma-Num est un hébergeur, une autre qualification qui implique des règles de responsabilité limitée, sachant que c'est le CNRS, comme personnalité juridique, qui assume ici cette responsabilité. La mention d'une unité de recherche indiquée dans le champ *« publisher »* est donc à interpréter comme une mention de source, mais sans consistance juridique réelle, la qualité d'éditeur ne pouvant être revendiquée sur ce type d'objet. Si le CNRS est vigilant concernant la revendication de la qualité d'éditeur sur les publications scientifiques, il n'entend en revanche pas (du moins à ce jour) demander aux archives ouvertes de modifier leurs pratiques concernant l'indexation des données dans des plate-formes qui suivent les pratiques du réseau métier concerné.

Il ne fallait pas moins que cet échange pour comprendre la signification de la mention *« publisher »* utilisée dans les métadonnées de la collection Pangloss (et plus généralement de la plate-forme Cocoon), et pour voir confirmé son caractère non problématique en l'état actuel des pratiques et recommandations.

L'exigence d'un consentement éclairé, dont il semblait qu'elle ne pouvait être gérée d'une façon conforme à l'éthique (voir ci-dessus, section « Autorisation de diffusion et consentement oral : un essai peu concluant »), apparaît également sous un jour nettement moins problématique lorsqu'elle est abordée avec l'appui de spécialistes des questions juridiques. Céline Aires, Déléguée à la protection des données de l'Université Sorbonne Nouvelle, propose, en concertation avec le réseau des Délégués à la protection des données, le principe d'un consentement éclairé oral devant témoin lettré. Ce témoin est une personne de confiance (autre que la chercheuse), qui atteste par écrit que le consentement éclairé a bien été donné après une information adaptée au contexte culturel : que l'explication fournie a été comprise et acceptée. De la sorte, le chercheur n'est pas seul à décider que le participant est éclairé.

Grâce aux conseils avisés des collègues du CNRS et de ses institutions partenaires, la collection Pangloss semble parée au plan juridique (aussi bien qu'au plan technique et organisationnel) pour faire face à une montée en charge, et destinée à connaître un bel avenir. Dans ce contexte, les avancées rapides du Traitement Automatique des Langues introduisent un élément d'incertitude, suscitant de grands espoirs mais aussi des interrogations et des inquiétudes.

19 Dublin Core est un vocabulaire du web sémantique utilisé pour exprimer les données dans un modèle *« Resource Description Framework in Attributes »*. Il a été adopté comme une recommandation du *World Wide Web Consortium* (W3C) définissant une syntaxe permettant d'ajouter des données structurées dans un document XML.

Traitement automatique des langues rares : un potentiel considérable pour la documentation linguistique… et des craintes de vol à l'étalage

Le caractère stratégique des outils informatiques pour les sciences humaines et sociales (Delmas-Rigoutsos 2023) est particulièrement manifeste dans le cas des sciences du langage. La perspective d'un soutien du Traitement Automatique des Langues (TAL) à la documentation linguistique constitue un des principaux espoirs dans un domaine où les raisons d'espérer sont peu nombreuses, au vu du déclin rapide de la diversité linguistique mondiale, parallèle au déclin de la biodiversité. Comme évoqué en introduction, le fort potentiel applicatif du TAL pour la documentation des langues est bien identifié (Anastasopoulos *et al.* 2020), et des réalisations viennent peu à peu le concrétiser (Harrigan *et al.* 2023). L'alignement forcé texte-parole permet un accès facilité aux ressources (Littell *et al.* 2022), et les outils de Reconnaissance Automatique de la Parole facilitent la transcription de documents audio et vidéo dans des langues rares, en mode *transcription exhaustive* (Partanen, Hämäläinen et Klooster 2020 ; Liu, Spence et Prud'hommeaux 2022 ; Guillaume *et al.* 2022) ou en mode *fouille de documents* : recherche de mots-clefs (Hjortnæs, Partanen et Tyers 2021). La synthèse de la parole trouve sa place dans des projets de revitalisation (Pine *et al.* 2022), de même que toutes sortes d'autres outils, par exemple pour le traitement des paradigmes verbaux de langues polysynthétiques (Kuhn *et al.* 2020). Les collaborations entre TAListes et spécialistes de langues rares ouvrent d'importantes perspectives en recherche, du fait des défis spécifiques posés par le scénario à faibles ressources que constitue le travail sur des langues peu documentées (Jimerson, Liu et Prud'Hommeaux 2023 ; Lonergan *et al.* 2023 ; Fily *et al.* 2024 ; San *et al.* 2024).

L'historique des collaborations *linguistique de terrain + TAL* nouées depuis 2014 autour de la transcription automatique de la langue na de Yongning a fait l'objet d'une relation circonstanciée à destination d'un public de linguistes de terrain (Michaud *et al.* 2018), complémentaire des articles de TAL au sujet du volet informatique (Do, Michaud et Castelli 2014 ; Adams *et al.* 2018 ; Wisniewski, Guillaume et Michaud 2020). Les présentes réflexions, qui s'organisent autour du thème de l'éthique de la recherche, fournissent l'occasion de relever la place bien réelle que tenaient les questions d'éthique dans ces collaborations.

Éthique de la recherche et collaborations pluridisciplinaires

Un point central qui ressort des réflexions des équipes pluridisciplinaires « TAListes + linguistes » qui travaillent sur des langues rares est la nécessité d'un dialogue constant, afin de parvenir à un niveau suffisant de compréhension mutuelle, et de rester suffisamment en phase au sujet des objectifs aussi bien que du planning opérationnel. Une convergence au plan de l'éthique de la recherche constitue une excellente base, sans doute indispensable à des collaborations véritablement fructueuses. Les principes ne sont pas bien compliqués : écoute et respect mutuel, reconnaissance d'une complémentarité ainsi que de certaines différences dans les méthodes, perspectives et motivations[20].

Ce qui permet de s'engager en confiance dans des projets avec des spécialistes d'autres disciplines que la sienne, c'est le constat d'une convergence au sujet d'une échelle de valeurs, dont découlent des objectifs partagés. Si on considère la constitution, l'archivage et la diffusion d'une documentation sur une langue en danger comme une fin en soi, on pourra travailler en bonne intelligence avec des collègues d'autres disciplines et d'autres métiers. Informaticien·nes, enseignant·es, linguistes, archivistes, anthropologues, ethno-botanistes ou autres : là n'est pas l'important, dès lors qu'on partage un objectif commun. Si l'éthique est une réflexion sur les

20 Un podcast sur le thème « Unir informatique et linguistique » a été réalisé en 2023 dans la collection *Vox* du Labex EFL (Fondements Empiriques de la Linguistique – Empirical Foundations of Linguistics) : https://podcast.ausha.co/vox-podcast-labex-efl/unir-informatique-et-linguistique

valeurs qui orientent et motivent nos actions, la valeur accordée à la documentation linguistique constitue un *ethos* – un lieu où s'établir et vivre une collaboration.

> … it was hard for the linguist to fathom to what extent one or another of the strands of research explored by [the computer science researcher] would lead to achievements of practical use in language documentation. But mutual confidence was building up nonetheless, based on our shared commitment to language documentation. In the same way as some linguists feel more strongly than others about the value of language diversity, and come to identify language documentation and language description as priorities, some computer scientists consider language processing for under-resourced languages as their field of specialization. A computer scientist working for the first time on real data on a newly documented endangered language can be as excited as a linguist on a first trip to the field. The sense of a common goal fosters mutual interest between linguists and computer scientists, itself conducive to mutual understanding. (Michaud *et al.* 2018, 403)

Le caractère positif des expériences menées, ainsi que des réutilisations non anticipées (Guzmán *et al.* 2017 ; Flamein et Eshkol-Taravella 2021), nous amenait à souhaiter leur multiplication. Faciliter l'utilisation de corpus de langues rares par des TAListes nous apparaissait comme un objectif hautement désirable, stratégique pour nos disciplines.

> Le déploiement d'outils de traitement automatique de la parole comporte des enjeux évidents pour la documentation des langues, à une époque où le déclin de la diversité linguistique s'accélère (…). Inversement, les langues rares présentent à la recherche en informatique tout un éventail de défis dont l'intérêt est de plus en plus clairement perçu. Dans ce contexte, la mise à disposition de corpus de langues rares aisément accessibles, clairement versionnés et faciles d'utilisation paraît une nécessité tout à fait centrale. Dans le droit fil de la publication du corpus mbochi (bantou), nous avons déposé dans Zenodo deux corpus audio (avec transcriptions) de langues rares : le japhug et le na, langues minoritaires de Chine, de la famille sino-tibétaine. (Galliot *et al.* 2021)

Le risque théorique d'abus était connu, mais cela paraissait un risque à courir. La mise en libre accès de l'ensemble de la collection Pangloss avait également constitué une prise de risque ; vingt ans plus tard, on constatait que les retours d'utilisatrices et utilisateurs (certes peu nombreux) étaient uniformément positifs, ce qui allait dans le sens d'encourager le choix d'aller pareillement de l'avant en matière de partenariats côté TAL. Pourtant, il est peu à peu apparu que les contacts avec le TAL n'amenaient pas seulement des collaborations passionnantes.

Incertitudes liées à l'évolution technologique rapide en Traitement Automatique des Langues

Pour les linguistes, avant même la mise en place d'éventuels projets communs avec des TAListes, le simple fait d'avoir un minimum d'information au sujet du domaine permet de prendre conscience du rythme très rapide de l'évolution du TAL, ce qui sensibilise efficacement au fait qu'il n'est, de fait, pas possible de prévoir toutes les réutilisations qui pourront être faites des données (Badin *et al.* 2022, §52). Cette observation comporte des conséquences aussi déconcertantes qu'essentielles en matière de risques éthiques liés aux traitements réalisables sur les corpus.

Une solution consiste à limiter les usages autorisés, par le choix judicieux de la licence sous laquelle sont placées les ressources orales. Mais l'expérience récente (dans les années 2020) de l'ingestion massive de données par les entreprises du numérique (pour l'entraînement de modèles statistiques) montre que les grandes entreprises ne sont pas toujours des acteurs respectueux des licences. En outre, la rapidité de l'évolution technologique complique son encadrement juridique.

L'Union Européenne affiche de saines ambitions (« Parliament's priority is to make sure that AI systems used in the EU are safe, transparent, traceable, non-discriminatory and environmentally friendly »[21]), mais leur opérationnalisation pourrait se faire attendre. « For a technology that is advancing as quickly as generative AI, the time it takes to develop a treaty is just too long to make the adopted norms effective. We will end up with a cat-and-mouse game between treaties and technology, as we have already seen in the digital environment and in other areas » (Yu 2024, 11).

Manquements à l'éthique dans le déploiement d'outils de TAL : conséquences pour les archives ouvertes

Les dérives d'un modèle socio-économique prédateur de données façon « GAFAM » (vol à l'étalage généralisé de données, et déploiement d'outils de TAL hors du contrôle des sociétés concernées) viennent aujourd'hui remettre en question le choix d'un accès ouvert aux données de langues rares. En 2018, un *keynote speaker* en sécurité informatique mettait en garde contre le déploiement bâclé d'outils statistiques dits d'intelligence artificielle, et opposait, aux sirènes du solutionnisme technologique, la recommandation d'écouter la voix des poètes (Mickens 2018). En 2021, année où l'*Association for Computational Linguistics* mettait en place son Comité d'éthique, une étude relevait les dangers de la course au gigantisme dans les modèles de langues : outre les dégâts environnementaux et la reproduction de biais sociaux, le risque de graves malentendus est inhérent au fait d'exposer des humains à un message *qui n'a été produit par personne*. « [T]he human tendency to attribute meaning to text, in combination with large LMs' ability to learn patterns of forms that humans associate with various biases and other harmful attitudes, leads to risks of real-world harm, should LM-generated text be disseminated » (Bender *et al.* 2021, 618).

Ces observations ne donnent pas tort aux formulations (certes plus vagues) d'un de nos poètes, qui écrivait, cinq ans avant sa disparition :

> Nous sommes des milliards à tourbillonner, derviches, de plus en plus fortement, étourdis, vertigineux, solipsistes, reclus en notre *self*. (…) le tournis général … donné en spectacles innombrables de télé-vérité à tous les écrans de notre « vivre en direct » : équivalemment milliards d'opinions horaires à somme nulle, établissant le régime *« post-truth/non-truth »*, disons « trumpiste », de l'oralité humaine : ce sont nos *big data*, le mauvais infini, que les autres, les Big Data scientifiques, engloutiront. (Michel Deguy, *L'envergure des comparses* (2017), Paris : Hermann, pp. 39-40.)

Pour revenir à la question qui nous occupe, et aux responsabilités des linguistes qui collectent des données sur le terrain, l'époque de l'archivisme insouciant est bien révolue. L'emploi des données pour des usages que n'autorise pas la licence qui leur est apposée dans l'archive remet en cause le mode de fonctionnement des archives ouvertes (qu'il s'agisse d'images, ou de données de langues rares). On hésite à laisser à l'étalage des corpus qui pourraient accélérer la généralisation aux langues les plus rares des outils de surveillance de masse désormais bien implantés dans nos sociétés. La voix de la prudence peut conseiller d'éviter l'ouverture de corpus qui pourraient, dans un avenir proche, faire l'objet d'utilisations non souhaitables.

Conclusion : Pangloss et la Reine Rouge, ou le jardinage et son éthique

Au fil des réflexions, il ressort que l'adoption conséquente et résolue de principes qu'on appellerait aujourd'hui de Science ouverte ne constitue pas par elle-même une réponse complète (et encore moins une réponse pérenne, *future-proof*) aux questions juridiques et éthiques

21 « EU AI Act : first regulation on artificial intelligence », https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

soulevées par la constitution, la publication électronique et l'exploitation de ressources linguistiques en langues rares. Les enjeux se déplacent d'une recherche de solutions simples et claires qui permettent d'éviter les soucis (on fait ce qu'il faut pour satisfaire aux critères en vigueur en matière d'éthique, d'ouverture des données, de partage des outils, de libre accès aux publications, et on a l'esprit libre pour se concentrer sur la recherche) vers un processus réflexif continu. Une réflexion exigeante et vigilante, naturellement soucieuse de respect des personnes et des groupes sociaux, informe l'ensemble du processus de documentation et de recherche, du stade de la définition des protocoles d'enquête jusqu'aux choix concernant la gestion des données sur le moyen terme et le long terme.

La référence au maître Pangloss de Candide (non dénuée d'autodérision, quand on connaît la nouvelle de Voltaire) oriente la collection Pangloss vers une certaine conception de l'éthique : celle d'une petite équipe qui cultive son jardin, *les pieds sur terre*, se défiant des approches et préconisations « hors sol ». Le jardin de données qu'est la collection Pangloss, conçu comme un Commun de la connaissance en accès ouvert, doit-il devenir secret, pour se protéger contre des usages contraires à toute éthique ? Rien ne dit que la tentation du repli soit bonne conseillère. La métaphore du jardinage, qui souligne le caractère continu de l'activité de maintenance et d'amélioration de la collection, suggère aussi que l'archive est, comme le vivant, prise dans une évolution permanente qui impose d'être constamment mobile pour parvenir à rester en place (courir tant qu'on peut pour rester au même endroit, dit la Reine Rouge dans *De l'autre côté du miroir*). L'adaptation est toujours à recommencer, et l'extinction toujours possible.

Remerciements

Vifs remerciements à tous·tes les collègues du monde de la Science ouverte (plate-forme Cocoon, Huma-Num, Centre pour la Communication Scientifique Directe, Comité pour la Science Ouverte…) dont les réalisations, les réflexions et l'exemple constituent la base de tout. Mes remerciements aux professionnels des archives orales qui m'ont fait découvrir leur métier au fil des échanges : en particulier Michel Jacobson, Joséphine Simonnot, Florence Gétreau, Pascal Cordereix. Merci à l'équipe de la collection Pangloss (Séverine Guillaume, Léa Mouton, Balthazar Do Nascimento) pour son appui constant, et à Flora Badin, Natalia Cáceres, Raphaëlle Chossenot et Julie Giovacchini pour leur vision stratégique et leurs conseils judicieux.

Merci aux organisatrices de la Journée Éthique et TAL 2024 de l'Association pour le Traitement Automatique des Langues (ATALA), dont la sollicitation a fourni l'occasion du présent travail.

Merci à Aliyah Morgenstern, Julie Marsault et Séverine Guillaume pour leur relecture, et à Evangelia Adamou et Jacqueline Vaissière pour leurs réflexions au fil de nos échanges.

Merci aux collègues qui m'ont fait la confiance et l'honneur de me confier des données de terrain en vue de leur publication électronique, en particulier à Michel Ferlus.

Il va de soi que les points de vue exposés ici n'engagent que moi.

Références citées

Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird et Alexis Michaud. 2018. « Evaluating phonemic transcription of low-resource tonal languages for language documentation ». Dans *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, 3356-3365. Miyazaki. https://halshs.archives-ouvertes.fr/halshs-01709648.

Alombert, Anne, Victor Chaix, Maël Montévil et Vincent Puig, éd. 2021. *Prendre soin : de l'informatique et des générations*. Limoges : FYP.

Anastasopoulos, Antonios, Christopher Cox, Graham Neubig et Hilaria Cruz. 2020. « Endangered languages meet Modern NLP ». Dans *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, 39-45.

Archambault, Éric et Vincent Larivière. 2009. « History of the journal impact factor: Contingencies and consequences ». *Scientometrics* 79 (3) : 635-649.

Arènes, Cécile, Lionel Maurel et Stéphanie Rennes. 2022. « Guide d'application de la Loi pour une République numérique pour les données de la recherche ». Ministère de l'enseignement supérieur et de la recherche. https://doi.org/10.52949/31.

Austin, Peter K. 2007. « Training for language documentation: Experiences at the School of Oriental and African Studies ». Dans *Documenting and revitalizing Austronesian languages*, édité par D. Victoria Rau et Margaret Florey, 25-41. University of Hawai'i Press.

Badin, Flora, Caroline Cance, Céline Dugua, Layal Kanaan-Caillol, Anne-Lyse Minard et Katja Ploog. 2022. « Les données orales en linguistique: Questions éthiques et cadre juridique ». *Bulletin de l'Association Française d'Archives Sonores (AFAS)*, nº 48 (décembre) : 158-181. https://doi.org/10.4000/afas.7496.

Bazin, Yoann et Élise Goiseau. 2023. « Vers un modèle alternatif des comités d'éthique de la recherche : Quel équilibre entre procédures et réflexivité ? » *Revue française de gestion* 49 (1). Cairn/Isako : 73-100.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major et Shmargaret Shmitchell. 2021. « On the dangers of stochastic parrots: can Language Models be too big? ». Dans *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.

Bendjaballah, Selma et Guillaume Garcia. 2023. « Les sciences sociales à l'épreuve de l'ouverture des données de la recherche ». *Terrains & travaux* 43 (2). Cachan : ENS Paris-Saclay : 211-216. https://doi.org/10.3917/tt.043.0211.

Bichurina, Natalia. 2017. « Baptêmes d'une langue ou un peu de magie sociale dans le passé et dans le présent (francoprovençal–arpitan–savoyard) ». *Cahiers du Centre de Linguistique et des Sciences du Langage* 52 : 119-138.

Bonnemason, Bénédicte, Véronique Ginouvès et Véronique Pérennou. 2001. *Guide d'analyse documentaire du son inédit, pour la mise en place de banques de données*. Parthenay, France : Modal - AFAS.

Bordignon, Frédérique. 2023. « Maintenir l'intégrité de la littérature scientifique à l'heure de l'ouverture de la science : étude de cas, enjeux techniques et rôle des acteurs | 3e session de l'Assemblée des partenaires de HAL ». Video/mp4. Centre pour la Communication Scientifique Directe. https://doi.org/10.60527/7DMS-H251.

Bouquiaux, Luc et Jacqueline Thomas. 1971. *Enquête et description des langues à tradition orale. Volume I: l'enquête de terrain et l'analyse grammaticale*. 2nd edition 1976. Paris : Société d'études linguistiques et anthropologiques de France.

Bowern, Claire. 2010. « Fieldwork and the IRB: A snapshot ». *Language* 86 (4). Linguistic Society of America : 897-905.

Caplan, Patricia, éd. 2003. *The ethics of anthropology: debates and dilemmas*. London; New York : Routledge.

Carlin, Marie et Arnaud Laborderie. 2021. « Le BnF DataLab, un service aux chercheurs en humanités numériques ». *Humanités numériques*, nº 4. https://doi.org/10.4000/revuehn.2684.

Cox, Christopher. 2022. « Managing Data in a Language Documentation Corpus ». Dans *The Open Handbook of Linguistic Data Management*, édité par Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller et Lauren B. Collister, 277-286. The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0027.

Crippen, James A. et Laura C. Robinson. 2013. « In defense of the lone wolf: Collaboration in language documentation ». *Language Documentation and Conservation* 7. University of Hawaii Press : 123-135.

Delmas-Rigoutsos, Yannis. 2023. « Numérique et humanités : de l'ancillarité à la fécondité grâce à la modélisation computationnelle des connaissances ». *Humanités numériques* 7. https://doi.org/10.4000/revuehn.3359.

Do, Thi Ngoc Diep, Alexis Michaud et Eric Castelli. 2014. « Towards the Automatic Processing of Yongning Na (Sino-Tibetan): Developing a "light" Acoustic Model of the Target Language and Testing "Heavyweight" Models from Five National Languages ». Dans *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014)*, 153-160. St Petersburg. http://halshs.archives-ouvertes.fr/halshs-00980431/.

Driem, George van. 2016. « Endangered language research and the moral depravity of ethics protocols ». *Language Documentation and Conservation* 10 : 243-252.

Dwyer, Arienne M. 2006. « Ethics and practicalities of cooperative fieldwork and analysis ». Dans *Essentials of language documentation*, édité par Jost Gippert, Nikolaus Himmelmann et Ulrike Mosel, 178 : 31-66. Walter de Gruyter.

Eslen-Ziya, Hande, Alberta Giorgi et Ceren J. Ahi. 2023. « Digital Vulnerabilities and Online Harassment of Academics, Consequences, and Coping Strategies. An Exploratory Analysis ». *Feminist Media Studies*, novembre, 1-6. https://doi.org/10.1080/14680777.2023.2281268.

Fassoulaki, A., K. Papilas, A. Paraskeva et K. Patris. 2002. « Impact factor bias and proposed adjustments for its determination ». *Acta Anaesthesiologica Scandinavica* 46 (7). Wiley Online Library : 902-905.

Ferlus, Michel. 2017. « Des enregistrements sans expérience à la numérisation : les insouciances du terrain ». *La lettre de l'AFRASE* 93-94 : 12-13.

Fily, Maxime, Guillaume Wisniewski, Severine Guillaume, Gilles Adda et Alexis Michaud. 2024. « Establishing degrees of closeness between audio recordings along different dimensions using large-scale cross-lingual models ». Dans *Findings of the Association for Computational Linguistics (EACL 2024)*. Malta. http://arxiv.org/abs/2402.05581.

Flamein, Hélène et Iris Eshkol-Taravella. 2021. « Exploitation du corpus Enquêtes sociolinguistiques à Orléans (ESLO) par les outils du traitement automatique des langues et de la géomatique ». *Humanités numériques*, n° 3. https://doi.org/10.4000/revuehn.1911.

Fluehr-Lobban, Carolyn. 2013. *Ethics and Anthropology: Ideas and Practice*. Lanham, MD : AltaMira press.

Gagnon, Éric. 2010. « Le comité d'éthique de la recherche, et au-delà ». *Éthique publique* 12 (1) : 299-308. https://doi.org/10.4000/ethiquepublique.284.

Galliot, Benjamin, Guillaume Wisniewski, Séverine Guillaume, Alexis Michaud, Solange Rossato, Minh-Châu Nguyên et Maxime Fily. 2021. « Deux corpus audio transcrits de langues rares (japhug et na) normalisés en vue d'expériences en traitement du signal ». Dans *Journées scientifiques du Groupement de recherche « Linguistique informatique, formelle et de terrain » (GDR LIFT)*. Grenoble. https://halshs.archives-ouvertes.fr/halshs-03475436.

Galonnier, Juliette, Stefan Le Courant, Anthony Pecqueux et Camille Noûs. 2019. « Ouvrir les données de la recherche ? » *Tracés*, n° 19 : 17-33. https://doi.org/10.4000/traces.10588.

Garellek, Marc, Matthew Gordon, James Kirby, Wai-Sum Lee, Alexis Michaud, Christine Mooshammer, Oliver Niebuhr, *et al.* 2020. « Toward Open Data Policies in Phonetics: What We Can Gain and How We Can Avoid Pitfalls ». *Journal of Speech Science* 9 (1). https://halshs.archives-ouvertes.fr/halshs-02894375.

Gros, Stéphane. 2014. « The bittersweet taste of rice. Sloping land conversion and the shifting livelihoods of the Drung in Northwest Yunnan (China) ». *Himalaya* 34 (2) : 81-96.

Guillaume, Séverine, Guillaume Wisniewski, Benjamin Galliot, Minh-Châu Nguyễn, Maxime Fily, Guillaume Jacques et Alexis Michaud. 2022. « Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource settings ». Dans *Proceedings of Interspeech 2022*. Incheon, Korea. https://halshs.archives-ouvertes.fr/halshs-03625581.

Gunthert, André. 2004. « La punition des revues ». *Études photographiques*, n° 15. Société française de photographie : 2-3.

Guzmán, Gualberto A., Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock et Almeida Jacqueline Toribio. 2017. « Metrics for modeling code-switching across corpora ». Dans *Proceedings of Interspeech 2017*, 67-71.

Harrigan, Atticus, Aditi Chaudhary, Shruti Rijhwani, Sarah Moeller, Antti Arppe, Alexis Palmer, Ryan Henke et Daisy Rosenblum. 2023. *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-6)*. Association for Computational Linguistics (ACL) Anthology.

Harris, Amanda, Nick Thieberger et Linda Barwick. 2015. *Research, records and responsibility: Ten years of PARADISEC*. Sydney University Press.

Heintz, Monica. 2023. « Protéger ou invisibiliser ses interlocuteurs : peut-on ouvrir les données ethnographiques ? » Édité par Selma Bendjaballah et Guillaume Garcia. *Terrains & travaux* 43 (2). Cachan : ENS Paris-Saclay : 233-255. https://doi.org/10.3917/tt.043.0211.

Henitiuk, Valerie et Marc-Antoine Mahieu. 2024. « Tangled lines: what might it mean to take Indigenous languages seriously? » *Translation Studies* 17 (1). Taylor & Francis : 169-180.

Hjortnæs, Nils, Niko Partanen et Francis Tyers. 2021. « Keyword spotting for audiovisual archival search in Uralic languages ». Dans *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, 1-7. Association for Computational Linguistics.

Jacobson, Michel. 2004. « Corpus oraux en linguistique de terrain ». *Traitement automatique des langues* 45 : 63-88.

Jacobson, Michel, Boyd Michailovsky et John B. Lowe. 2001. « Linguistic documents synchronizing sound and text ». *Speech Communication* 33 [special issue: "Speech Annotation and Corpus Tools"] : 79-96.

Jimerson, Robert, Zoey Liu et Emily Prud'Hommeaux. 2023. « An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language ». Dans *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1008-1016.

Kaufmann, Sina Kamala, Michael Timmermann, Annemarie Botzki et Steffen Greiner, éd. 2019. *Wann wenn nicht wir*: ein extinction rebellion Handbuch*. Traduit par Ulrike Bischoff. Erweiterte deutsche Erstausgabe. Frankfurt am Main : S. Fischer.

Kuhn, Roland, Fineen Davis, Alain Désilets, Eric Joanis, Anna Kazantseva, Rebecca Knowles, Patrick Littell, Delaney Lothian, Aidan Pine et Caroline Running Wolf. 2020. « The Indigenous Languages Technology project at NRC Canada: An empowerment-oriented approach to developing language software ». Dans *Proceedings of the 28th international conference on computational linguistics*, 5866-5878.

Launey, Michel. 2023. *La République et les langues*. Cours et travaux. Paris : Raisons d'agir.

Littell, Patrick, Eric Joanis, Aidan Pine, Marc Tessier, David Huggins-Daines et Delasie Torkornoo. 2022. « Readalong studio: Practical zero-shot text-speech alignment for indigenous language audiobooks ». Dans *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, 23-32.

Liu, Zoey, Justin Spence et Emily Prud'hommeaux. 2022. « Enhancing documentation of Hupa with automatic speech recognition ». Dans *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Lonergan, Liam, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl et Ailbhe Ní Chasaide. 2023. « Towards Spoken Dialect Identification of Irish ». Dans *Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-Resourced Languages (SIGUL 2023)*, 63-67. ISCA. https://doi.org/10.21437/SIGUL.2023-14.

Malaurie, Jean. 1989. *Les derniers rois de Thulé: avec les Esquimaux polaires, face à leur destin ; Et après ? Retours à Thulé suivi d'un dossier « Débats et critiques »*. 5e éd. revue et Augmentée. Paris : Plon.

Marsault, Julie. 2021. « Valency-Changing Operations in Umóⁿhoⁿ : Affixation, Incorporation, and Syntactic Constructions, Les Modifications de La Valence Verbale En Umóⁿhoⁿ : Affixation,

Incorporation et Constructions Syntaxiques ». Thèse de doctorat, Université de la Sorbonne nouvelle - Paris III. https://theses.hal.science/tel-03573762.

Michaud, Alexis. 2002. « “Tu pourrais enregistrer un corpus pour moi?” Pour une charte de qualité des corpus ». Dans *XXIVe Journées d'Etude de la Parole*, 153-156. Nancy, France. https://shs.hal.science/halshs-01647020.

Michaud, Alexis. 2017a. « In memoriam He Qin (1973-2016) — 纪念和沁教授 ». Text/html. *Indo-Sinica · 震南古聲隨記*. https://doi.org/10.58079/Q69H.

Michaud, Alexis. 2017b. « Le projet de numérisation DO-RE-MI-FA : données des recherches de Michel Ferlus en Asie du Sud-Est ». *Lettre de l'AFRASE (Association Française pour la Recherche en Asie du Sud-Est)*, 2017.

Michaud, Alexis. 2017c. *Tone in Yongning Na: lexical tones and morphotonology*. Studies in Diversity Linguistics 13. Berlin : Language Science Press. http://langsci-press.org/catalog/book/109.

Michaud, Alexis, Oliver Adams, Trevor Cohn, Graham Neubig et Séverine Guillaume. 2018. « Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit ». *Language Documentation & Conservation* 12 : 393-429.

Michaud, Alexis, Minh-Châu Nguyễn et He Likun. 2020. « Voix de « ceux qui ne sont rien » en Asie du Sud-Est ». *Cahiers de littérature orale*, nº Hors-Série : 147-154. https://doi.org/10.4000/clo.7019.

Mickens, James. 2018. « Q: Why do keynote speakers keep suggesting that improving security is possible? A: Because keynote speakers make bad life decisions and are poor role models ». Dans *27th USENIX Security Symposium (USENIX Security 18)*.

Middeldorp, Nick et Philippe Le Billon. 2020. « Deadly environmental governance: authoritarianism, eco-populism, and the repression of environmental and land defenders ». Dans *Environmental governance in a populist/authoritarian era*, 24-37. Routledge.

Partanen, Niko, Mika Hämäläinen et Tiina Klooster. 2020. « Speech recognition for endangered and extinct Samoyedic languages ». Dans *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*. https://arxiv.org/abs/2012.05331.

Pine, Aidan, Dan Wells, Nathan Brinklow, Patrick Littell et Korin Richmond. 2022. « Requirements and motivations of low-resource speech synthesis for language revitalization ». Dans *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7346-7359.

Ripple, William J., Christopher Wolf, Thomas M. Newsome, Mauro Galetti, Mohammed Alamgir, Eileen Crist, Mahmoud I. Mahmoud, William F. Laurance et 364 scientist signatories from 184 countries 15. 2017. « World scientists' warning to humanity: a second notice ». *BioScience* 67 (12) : 1026-1028.

Rooryck, Johan. 2016. « Introducing Glossa ». *Glossa* 1 (1). Ubiquity Press : 1-3. https://doi.org/10.5334/gjgl.91.

San, Nay, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams et Dan Jurafsky. 2024. « Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens ». *arXiv preprint arXiv:2402.02302*.

Schrag, Zachary M. 2010. *Ethical Imperialism: Institutional Review Boards and the Social Sciences 1965-2009*. Baltimore, MD : John Hopkins University Press.

Simon, Camille et Camille Noûs. 2021. « The grammaticalization of plurality in the languages of Amdo ». *Himalayan Linguistics* 20 (3) : 49-81.

Simonnot, Joséphine. 2020. « Partager les archives sonores du musée de l'Homme sur le web avec la plateforme Telemeta ». *Bulletin de l'Association Française d'Archives Sonores (AFAS)* 46 : 88-101. https://doi.org/10.4000/afas.4056.

Simons, Gary et Steven Bird. 2000. « The seven pillars of open language archiving: A vision statement ». *Open Language Archives Community. Available at http://www.language-archives.org/docs/vision.html*. 2000.

Snead, Taylor et Ellen Cushman. 2023. « Building a community-centered archive for Cherokee language description, documentation, and reclamation ». *The Modern Language Journal* 107 (1). Wiley Online Library : 242-267.

Vapnarsky, Valentina. 2020. « Retour aux sources ? Circulation et virtualités des savoirs amérindiens à l'ère du numérique ». *Journal de la société des américanistes* 106 (2) : 79-103. https://doi.org/10.4000/jsa.19003.

Vasile, Aurelia, Séverine Guillaume, Mourad Aouini et Alexis Michaud. 2020. « Le Digital Object Identifier, une impérieuse nécessité ? L'exemple de l'attribution de DOI à la Collection Pangloss, archive ouverte de langues en danger ». *I2D - Information, données & documents* 2 : 156-175.

Vernaudon, Jacques, Nick Thieberger, Tamatoa Bambridge et Takurua Parent. 2021. « Un nouveau souffle numérique pour les corpus en langues océaniennes ». *Journal de la société des océanistes*, nº 153 (décembre) : 323-336. https://doi.org/10.4000/jso.13165.

Wisniewski, Guillaume, Séverine Guillaume et Alexis Michaud. 2020. « Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? » Dans *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*, édité par Dorothee Beermann, Laurent Besacier, Sakriani Sakti et Claudia Soria, 306-315. Marseille, France : European Language Resources Association (ELRA). https://halshs.archives-ouvertes.fr/hal-02513914/.

Woodbury, Tony. 2011. « Language documentation ». Dans *The handbook of endangered languages*, édité par Peter Austin et Julia Sallabank, 1 : 35-51. Cambridge : Cambridge University Press.

Yu, Peter K. 2024. « The future path of Artificial Intelligence and copyright law in the Asian Pacific ». *Computers and Law* 96 : 24-18.