Breathing digital life into Oceanic language corpora

Jacques Vernaudon, Nick Thieberger, Tamatoa Bambridge, Takurua Parent

Un nouveau souffle numérique pour les corpus en langues océaniennes

par

Jacques Vernaudon[*], Nick Thieberger[**], Tamatoa Bambridge[***] et Takurua Parent[****]

RÉSUMÉ

Les savoirs autochtones sont intimement liés aux langues océaniennes et aux textes, oraux ou écrits, produits dans ces langues. Les corpus primaires, que ce soit des écrits ou des enregistrements audio ou vidéo, recueillis sur le terrain par les chercheur·euses et conservés comme base de leurs publications académiques sont longtemps restés inaccessibles aux communautés source. Avec l'essor des archives numériques, ces ressources peuvent désormais être mises en ligne, permettant ainsi aux communautés locales de se réapproprier les informations collectées auprès de leurs aïeux. Les chants, les styles de performance, les mots, les récits sont autant de liens puissants avec le patrimoine et l'environnement. Nous présentons ici les enjeux scientifiques, éthiques et méthodologiques liés à ces archives numériques, plus particulièrement à travers une présentation de la *Pacific and Regional Archive for Digital Sources in Endangered Cultures* (PARADISEC) et d'une base linguistique locale *Anareo*, en cours de développement en Polynésie française en collaboration avec PARADISEC.

MOTS-CLES : archives numériques, enregistrements, langues océaniennes, savoirs autochtones, documentation linguistique

ABSTRACT

Indigenous knowledge is intimately linked to Oceanic languages and to oral or written texts produced in these languages. Primary recordings collected in the field by researchers and held as the basis of their academic publications, have long remained inaccessible to source communities. With the rise of digital archives, they can now be put online, allowing local communities to reclaim information collected from their ancestors. The songs, the performance styles, the words, the stories are all powerful links with heritage and the environment. We present here the scientific, ethical and methodological issues linked to these archives, more particularly through a presentation of the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), and that of the local *Anareo* linguistic database under development with PARADISEC in French Polynesia.

KEYWORDS : digital archives, recordings, Oceanian languages, indigenous knowledge, linguistic documentation

1. Introduction

[*] Université de la Polynésie française, équipe d'accueil Sociétés traditionnelles et contemporaines en Océanie (EASTCO EA 4241), jacques.vernaudon@upf.pf

[**] Université de Melbourne et Pacific and regional archive for digital sources in endangered cultures (PARADISEC), thien@unimelb.edu.au

[***] Université Paris Sciences & Lettres : EPHE-UPVD-CNRS, USR 3278 Centre de recherches insulaires et observatoire de l'environnement (criobe), tamatoa.bambridge@criobe.pf

[****] Université de la Polynésie française, takurua.parent@etudiant.upf.pf

D’une génération à une autre, les connaissances et les savoir-faire sur l’environnement en Océanie sont transmises par imitation de pratiques qui ne sont pas nécessairement verbalisées, mais aussi par des mots et des textes (récits, chants, explications, etc.) produits dans des langues particulières. On peut citer en exemple les taxinomies vernaculaires dont l’attention se concentre sur les espèces remarquables ou utiles aux yeux des sociétés. En tahitien, par exemple, l’espèce de bananier *Musa troglodytarum* reçoit plus de vingt-cinq noms vernaculaires selon ses variétés soigneusement distinguées (ex. *'āpiri*, *'āpura*, *farafara*, *māmara*, *'ōmene,* etc.). Les textes issus de la tradition orale comportent d’innombrables détails sur les espèces et leurs utilisations. Exemple fameux, le récit tahitien de la capture du soleil par Māui, recueilli par Orsmond en 1825 (Henry, 1993), énumère les plantes sélectionnées par le héros dans la fabrication des fibres de son lasso[1] : le *rō'ā*, *Pipturus argenteus*, arbuste dont les racines étaient autrefois utilisées pour les lignes de pêche ; le *mati*, *Ficus tinctoria*, dont l’écorce servait à la fabrication des filets ; le *'ie'ie*, *Freycinetia impavida*, dont les racines aériennes sont utilisées pour les paniers ; le *more*, écorce interne du *pūrau*, *Hibiscus tiliaceus*, avec laquelle on confectionne des cordes, des nattes, des costumes et, autrefois, des sandales ; le *nape*, cordelette en fibres de bourre de coco, *Cocos nucifera*. Les langues offrent également une terminologie appropriée aux environnements géomorphologiques occupés par les populations (ex. les termes *hoa* et *karena* en pa'umotu, qui désignent respectivement une petite ouverture dans le récif et un pinacle madréporique dans le lagon d’un atoll). Aussi, même si la transmission des savoirs et savoir-faire sur l’environnement ne relève pas exclusivement d’une modalité verbale et même si ces savoirs autochtones peuvent toujours s’exprimer, par la traduction, dans la langue d’observateurs étrangers, et se transmettre, y compris au sein des populations locales, dans une langue empruntée, leur vitalité est au moins partiellement liée à celle des langues océaniennes et à la diffusion intergénérationnelle des textes, oraux ou écrits, produits dans ces langues.

Cependant, d’ici à la fin du siècle, en raison des processus de mondialisation, de la pression des anciennes langues coloniales et des créoles véhiculaires, et de la transformation des modalités de transmission du savoir aux jeunes générations dans un cadre scolaire emprunté à l’Occident, il ne restera sans doute plus que quelques centaines des 1 500 langues autochtones parlées en Océanie. De manière plus générale, Campbell *et al.* (2013), sur la base d’analyses des données du catalogue des langues en danger, estiment que la moitié des langues du monde aura disparue d’ici à la fin du XXI^{e} siècle.

Beaucoup d’éléments des corpus recueillis par les linguistes, les musicologues et les ethnographes concernent des connaissances sur les pratiques traditionnelles. Ce matériau, lorsqu’il est accessible, peut nourrir la revitalisation des pratiques locales et la revalorisation des cultures autochtones, comme l’illustrent un certain nombre de programmes linguistiques, par exemple à Hawa'ii ou en Nouvelle-Zélande, qui s’appuient sur des documents d’archives pour réactiver, dans des journaux et d’autres média, des mots et des expressions sortis de l’usage. Les enregistrements sonores et écrits sont aussi une ressource précieuse pour réapprendre des chansons et des récits.

Tant que ces documents sont détenus uniquement par le·la chercheur·euse qui les a collectés, ou qu’ils sont archivés dans son université ou dans un musée, ils restent inaccessibles aux personnes enregistrées et à leurs familles qui sont intéressées au premier chef par ces données. Les publications universitaires, dont la production est le but principal de la plupart des

[1] *Hāmani a'era Māui i te here e here i te rā. Here pa'ari tāna i ha'a, e here rō'ā, e here mati, e here 'ie'ie, e here more 'e te here nape, 'o tē tūtau-ana'e-hia i te tai.* (Henry, 1993 : 446)
‘Māui confectionna un lasso pour attraper le soleil. C’était un lasso solide qu’il fabriqua, en *rō'ā*, en *'ie'ie*, en *more* et en *nape*, toutes [ces fibres] avaient été mises à tremper dans l’eau de mer.’ (notre traduction)

chercheur·euses, ne sont généralement pas rédigées dans un style ou une langue accessible au grand public et sont souvent coûteuses et de diffusion restreinte, parfois non disponibles dans les pays où le travail de terrain a été réalisé. L'essor du numérique et d'Internet ouvre cependant de nouvelles perspectives en matière d'archivage et d'accessibilité des corpus recueillis sur le terrain (Thieberger et Harris, 2020).

Cette dynamique récente s'accompagne d'enjeux complexes, à la fois scientifiques, technologiques et éthiques, et conduit parfois à des injonctions contradictoires. Par exemple, le mouvement actuel en faveur de la science ouverte et des données librement accessibles ne tient pas toujours pleinement compte des droits des peuples autochtones. Les principes FAIR[2] (*Findable, Accessible, Interoperable, Reusable*), par exemple, se concentrent principalement sur les caractéristiques techniques des données afin d'en fluidifier la circulation, mais ignorent les différents contextes sociaux et historiques dont elles sont issues et les asymétrie de pouvoir, par exemple entre les institutions publiques ou privées et les communautés locales. Le principe de libre circulation des données peut aller à l'encontre de l'intérêt des peuples autochtones qui revendiquent au contraire un plus grand contrôle, pour le bénéfice de leur communauté, sur la diffusion et l'utilisation des savoirs et savoir-faire dont ils sont détenteurs, conformément à d'autres principes, comme les « CARE principles for Indigenous data governance » : *Collective benefit*, *Authority to control*, *Responsibility*, *Ethics* [3] (Thieberger, 2020).

En s'appuyant sur l'exemple de deux archives numériques, l'une développée en Australie, l'autre en Polynésie française, cet article explore les transformations induites dans le rapport aux corpus primaires sur les savoirs autochtones, auxquels on reconnaît désormais une valeur aussi importante que celle de l'analyse scientifique qui est produite à partir de ces corpus. Les implications méthodologiques de l'archivage numérique sur le recueil des données sont également abordées. Cette présentation illustre les nouvelles modalités de transmission des savoirs offertes par ces plateformes et comment l'implication des académies de langues locales et des chercheur·euses autochtones permet, à la source, d'adresser les enjeux éthiques liés à l'élaboration et au pilotage des archives numériques (Harris *et al.*, 2019b).

2. L'archive PARADISEC

Pour valoriser les corpus empiriques primaires recueillis sur le terrain et répondre aux enjeux de conservation et de transmission du patrimoine symbolique de l'humanité, une équipe de chercheurs soutenue par l'Université de Sydney, l'*Australian National University* et l'Université de Melbourne a développé depuis 2003 un ambitieux programme d'archivage intitulé *Pacific and regional archive for digital sources in endangered cultures* (PARADISEC)[4]. Ce programme a pour objectif de collecter et de conserver des documents menacés de disparition provenant de la région Pacifique et de faciliter l'accès à ces ressources grâce à une plateforme de consultation en ligne. PARADISEC a développé un protocole de restauration des supports dégradés, de numérisation et d'indexation dans le respect des normes partagées. Ces standards permettent un moissonnage de l'information et de consultation selon des critères de communicabilité définis par les auteurs des enregistrements et les locuteurs-informateurs ou leurs ayant droits (Thieberger, 2020 ; Thieberger et Harris, 2021). Cette structure héberge à ce

[2] Cf. https://www.go-fair.org/fair-principles/
[3] Cf. https://www.gida-global.org/care.
[4] En français, Archives de sources numériques sur les cultures menacées de la région Pacifique. http://www.paradisec.org.au/

jour 600 collections dans 1 278 langues, avec plus de 14 000 heures d'enregistrements audio et 1 600 heures de vidéos.

PARADISEC s'est employé à sensibiliser des institutions du Pacifique pour numériser les enregistrements audio et vidéo des différentes formes d'expressions culturelles traditionnelles, parfois avec l'aide de petites subventions de bailleurs de fonds comme le Programme des archives en péril (*Endangered Archives Programme*) ou le Programme de documentation sur les langues en danger (*Endangered Languages Documentation Programme*). Ces actions ont déjà permis de sécuriser les collections suivantes : 180 bandes audio du Centre culturel de Vanuatu, 230 bandes audio du Musée national des îles Salomon, 190 bandes audio de l'Université Divine Word à Madang en Papouasie Nouvelle-Guinée et une collection de 180 enregistrements audio de Madang conservés au musée de Bâle. Sans le travail de PARADISEC, ces collections risquaient de devenir inexploitables et d'être perdues. De même, PARADISEC mène depuis 2012 une enquête, intitulée « *Lost and Found* », dont l'objectif est d'identifier des collections de bandes audio qui ont besoin d'être numérisées et archivées. Grâce à ce programme, plusieurs collections ont été localisées. En s'appuyant sur le réseau international d'archives, chaque collection peut être prise en charge, soit dans ou à proximité du pays d'origine de son possesseur, par exemple à l'*Archive of the Indigenous Languages of Latin America* (AILLA) au Texas, l'Institut Max Planck en psycholinguistique à Nimègue ou à la *School of Oriental and African Studies* (SOAS) à Londres, soit être expédiée en Australie afin d'y être numérisée[5].

Il reste encore beaucoup à faire d'autant que de nouvelles collections à numériser ne cessent jamais d'être découvertes et il convient d'encourager la création d'un plus grand nombre d'archives implantées localement et mieux placées pour obtenir la confiance des partenaires institutionnels locaux (académies, musées, archives locales, services gouvernementaux, etc.) et des collectionneurs qui possèdent encore des enregistrements à numériser. Il importe de souligner ici qu'il ne s'agit pas d'une simple tâche mécanique de conversion de formats. Ce travail est nécessairement fondé sur la construction d'une relation de confiance avec les détenteur·trices des corpus et l'opérateur d'archivage, comme PARADISEC, doit garantir aux communautés à la source des enregistrements que les fichiers numérisés seront mis à leur disposition. La numérisation implique également le recueil du consentement préalable et éclairé des personnes enregistrées, qu'elles souhaitent être nommément citées ou demeurer anonymes. Cela implique de publier des métadonnées sur internet en utilisant des termes standards qui permettent au catalogue d'être moissonné par des moteurs de recherche puissants. Cette configuration, à son tour, favorise la visibilité du catalogue sur internet et le rend plus facile à trouver, en particulier par des diasporas dont les membres sont situés loin de leur lieu d'origine.

Si PARADISEC est aujourd'hui l'archive linguistique qui compte dans son catalogue le plus d'enregistrements issus du Pacifique, elle n'est ni la seule, ni la première. Un premier effort d'archivage linguistique de matériel textuel strictement numérique a débuté en 1991 en Australie avec l'*Aboriginal Studies Electronic Data Archive* (ASEDA), qui stockait des fichiers de données pour des dictionnaires de langues aborigènes (Thieberger, 1995). L'objectif était que ces données lexicographiques puissent être réutilisées pour de nouvelles éditions de

[5] Voici quelques exemples de collections identifiées dans le cadre de cette enquête et le lien vers leur archivage numérique : Don Kulick, 82 enregistrements : Tayap (Papouasie Nouvelle-Guinée) http://catalog.paradisec.org.au/collections/DK1 ; Zygmunt Frajzyngier, 65 enregistrements : Afrique du Nord http://catalog.paradisec.org.au/collections/ZF1 ; Wolfgang Sperlich, 29 enregistrements : Namakira (Vanuatu) http://catalog.paradisec.org.au/collections/WS1 ; Lamont Lindstrom, 23 enregistrements : Kwamera (Vanuatu) http://catalog.paradisec.org.au/collections/LL1

dictionnaires et, puisqu'il s'agit d'une compilation de fichiers numériques, qu'elles puissent être explorées comme un corpus lexical de termes de nombreuses langues australiennes.

Par ailleurs, à partir de 1992, avant que quiconque ne se soit engagé dans ce type de problématique, une équipe de chercheur·euses et d'ingénieurs du laboratoire du CNRS « Langues et civilisations à tradition orale » (LACITO) ont développé le projet précurseur *Archivage*, créant un éditeur XML et une application en ligne pour associer des médias audio et leurs transcriptions (Jacobson *et al.*, 2001). Désormais nommé *Pangloss*, ce programme continue d'explorer les possibilités des médias numériques dans une perspective visionnaire qui trouve encore peu d'équivalents. On peut citer également les archives DOBES (*Dokumentation bedrohter Sprachen*), débutées en 2000 par l'institut Max Planck en psycholinguistique à Nimègue, et la *Kaipuleohone language archive* fondée en 2008 à l'Université de Hawai'i. Ces quatre archives rassemblent aujourd'hui la plupart des enregistrements audio numérisés sur le Pacifique (François, 2018).

L'essor de ces archives numériques a été encouragé par le développement remarquable des technologies informatiques et d'internet, mais il s'inscrit aussi dans le contexte d'une importante transformation épistémologique dans le champ de la linguistique, débutée dans les années 1990, avec l'émergence de la « documentation linguistique » (*language documentation* en anglais) en tant que discipline autonome. Ce changement de paradigme a été théorisé, entre autres, par Himmelmann (1998, 2006), lequel en donne la définition suivante :

> « […] a field of linguistic inquiry and practice in its own right which is primarily concerned with the compilation and preservation of linguistic primary data and interfaces between primary data and various types of analyses based on these data. […] a language documentation is a lasting, multipurpose record of a language. » (Himmelmann, 2006 : 1)

Auparavant, les corpus linguistiques, s'ils existaient, étaient destinés principalement à servir de sources empiriques pour produire des travaux académiques comme des grammaires ou des dictionnaires, lesquels documents étaient les seuls à avoir une valeur dans le champ universitaire. Les corpus tombaient souvent dans l'oubli. Désormais, on reconnait aux corpus une valeur intrinsèque qui ne peut être subsumée par les productions académiques qui en découlent. Ils doivent donc être constitués avec soin et conservés durablement pour servir des objectifs multiples, dont la production de descriptions linguistiques, mais pas exclusivement. Comme le résume Himmelmann :

> « The goal is not a short-term record for a specific purpose or interest group, but a record for generations and user groups whose identity is still unknown and who may want to explore questions not yet raised at the time when the language documentation was compiled. » (Himmelmann, 2006 : 2)

Par exemple, lorsque Thieberger effectuait dans les années 1990 un travail de terrain sur le nafsan, une des langues d'Efate, Vanuatu, le développement d'outils numériques d'aide au traitement des enquêtes linguistiques débutait à peine. Le LACITO a développé une méthodologie pour créer des transcriptions synchronisées de médias (Jacobson, Michailovsky et Lowe, 2001) permettant de construire un corpus de textes liés aux enregistrements audio. À titre comparatif, la première version du logiciel de transcription ELAN[6] ne fut publiée qu'en

[6] ELAN (acronyme de « EUDICO linguistic annotator ») est un logiciel gratuit de transcription développé par l'Institut Max Planck en psycholinguistique à Nimègue, Pays-Bas. Il est désormais largement utilisé par les linguistes et autres chercheurs travaillant sur des corpus audio ou vidéo.

2002. L'équipe du LACITO a généreusement partagé ses outils avec Thieberger, qui les a utilisés pour construire à son tour un corpus d'enregistrements qui peuvent être cités (Thieberger, 2004), permettant ainsi à sa grammaire de nafsan de comporter des liens entre les phrases sous forme écrite et leurs sources média audio. Cet ensemble de documents a continué de croître, mais la citation de fichiers qui ne sont accessibles que sur l'ordinateur d'un chercheur n'est pas une configuration idéale. Thieberger a donc travaillé avec des collègues pour créer une archive dans laquelle les enregistrements linguistiques pourraient être déposés et conservés pour un accès à long terme. Ainsi est née PARADISEC en 2003 (Thieberger et Barwick, 2012).

Le simple fait de numériser, de décrire et de rendre les documents accessibles en ligne transforme profondément la méthodologie et le produit de la recherche, à plusieurs égards. Premièrement, l'opération préserve les documents eux-mêmes et révèle leur existence au monde, conformément aux recommandations convergentes des principes FAIR et CARE cités plus haut. Les principaux bénéficiaires sont les personnes enregistrées et leurs descendants, qui n'étaient pas toujours considérés auparavant comme des partenaires intéressés. En conditionnant la consultation des données à l'acceptation d'une licence d'utilisation, les archives peuvent garantir l'accès aux enregistrements de manière uniforme et stable, en évitant de se fier à la seule disponibilité d'un·e chercheur·euse pour traiter les demandes de copies d'enregistrements et à ses bonnes grâces pour l'accès aux enregistrements, surtout s'il·elle n'est pas nécessairement en mesure de localiser et de mettre à disposition l'ensemble de ses enregistrements réalisés dans le passé. Le·la chercheur·euse a également intérêt à confier la tâche de l'archivage à une structure spécialisée car il·elle peut ainsi continuer à accéder à ses enregistrements sur le long terme, malgré la dégradation des supports initiaux ou la perte de ses propres fichiers informatiques.

Du point de vue méthodologique, des enregistrements identifiables de manière pérenne peuvent être créés dès le travail d'enquête et déposés en ligne depuis le terrain, ou peu de temps après. Ils peuvent ainsi être cités à toutes les étapes ultérieures de la recherche grâce à un permalien[7] fourni par l'agence d'archivage. C'est important car cela permet un traçage des enregistrements source, que l'on peut réécouter à tout moment, en particulier lorsqu'ils sont utilisés dans une analyse descriptive. Les affirmations d'une publication de recherche peuvent ainsi être étayées avec des enregistrements pouvant être cités et réécoutés directement dans leur contexte d'origine. Cette démarche garantit ainsi une méthodologie de recherche vérifiable, là où auparavant cela n'était tout simplement pas possible. De plus, l'archivage numérique peut commencer dès le début d'un projet de recherche et non à la fin d'une carrière universitaire, comme c'était le cas auparavant. En réalisant l'archivage dès les étapes initiales de la recherche, les métadonnées sont encore fraîches dans l'esprit de l'enquêteur·trice et la création d'une collection n'est donc pas aussi onéreuse et hasardeuse que lorsqu'on doit l'entreprendre trente ans après la réalisation de l'enquête (Barwick, 2004).

Un autre enjeu épistémologique majeur est qu'un corpus de recherche bien construit et accessible en ligne peut être réutilisé par d'autres chercheurs·euses pour explorer des sujets qui n'avaient pas été pris en compte par le·la chercheur·euse initial·e. Ana Krajinovic, étudiante en doctorat sous la direction de Thieberger, a commencé son analyse des marqueurs temporels, aspectuels et modaux en nafsan en travaillant sur la collection archivée de ce dernier, avant qu'elle ne l'accompagne sur le terrain pour effectuer une enquête plus approfondie (Krajinovic, 2019). Rosey Billington a étudié en postdoctorat la phonétique du nafsan et, avec la professeure Janet Fletcher, sa prosodie, elle aussi en s'appuyant sur le

[7] Un permalien est un lien hypertexte qui pointe vers une URL dont le contenu web est stable et pérenne.

corpus déjà accessible en ligne (Billington *et al.*, 2018). La même collection a été utilisée dans le projet DoReCo[8] et le projet Multi-CAST[9].

Au fur et à mesure que de nouvelles technologies se développent, les collections numériques comportant à la fois des enregistrements et leurs transcriptions fournissent une base d'apprentissage pour les logiciels de reconnaissance vocale (Foley *et al.*, 2019), favorisant ainsi la réalisation de davantage d'enregistrements qui peuvent être, au moins partiellement, retranscrits automatiquement. Mais rien de tout cela n'est possible si les enregistrements ne sont pas correctement archivés.

Sur la base de ses dix-huit années d'expérience dans la conservation de données linguistiques et ethnographiques, l'équipe de PARADISEC opère désormais la migration de l'ensemble de sa collection dans l'*Oxford Common File Layout*, un standard émergeant qui vise à rendre les collections disponibles dans un format ouvert[10]. Il permet à chaque item (image, texte, vidéo, audio, etc.) d'être autonome avec ses propres métadonnées et d'être ainsi indépendant de tout catalogue monolithique[11] de base de données. Les services permettant d'exploiter la collection, y compris les outils de recherche et de visualisation pour divers types de fichiers, peuvent tous être implémentés en tant que microservices ou widgets individuels, et peuvent être partagés avec d'autres archives utilisant les mêmes normes.

Les microservices offrent l'avantage de pouvoir être paramétrés pour fonctionner sur de petits appareils de rapatriement, comme le Raspberry pi[12], afin que les collections soient accessibles sur des téléphones ou des tablettes. Un catalogue en ligne est accessible pour des utilisateurs d'internet, mais il ne convient pas à une utilisation hors ligne si l'on veut partager les fichiers à partir d'un disque dur. L'un des microservices conçu par PARADISEC, appelé « chargeur de données » (*data loader*), offre cette fonctionnalité[13]. Il permet d'extraire des fichiers du catalogue avec des éléments de la collection pour les déposer sur un disque dur, puis de consulter en local des fichiers grâce à une interface HTML créée à cet effet.

Une des conditions de la pérennité des bases numériques est la stabilité de leur financement, ce qui ne va pas de soi eu égard aux coûts importants en matière d'équipement et de ressources humaines. Jusqu'à présent, PARADISEC a été financé par des subventions ponctuelles provenant de sources diverses. PARADISEC offre également un service payant de numérisation de bandes analogiques qui génère quelques ressources propres. L'Université de Sydney lui a fourni des postes à temps partiel et deux salles dans son « Conservatorium of Music ». L'Australian National University et l'Université de Melbourne ont également mis à disposition des bureaux. Il est désormais entré dans l'usage d'intégrer les coûts d'archivage dans les demandes de financement pour les projets de recherche afin de soutenir le traitement archivistique et la conservation des données. PARADISEC a ainsi été l'opérateur désigné de plusieurs projets de ce type avec une part de financements dédiée à l'archivage, ce qui lui a permis d'obtenir des fonds complémentaires. Nous considérons que le processus de création

[8] http://doreco.info/

[9] https://multicast.aspra.uni-bamberg.de/

[10] Les spécifications techniques d'un format « ouvert » sont publiques et sans restriction d'accès ni de mise en œuvre, contrairement à celles d'un format « fermé » ou « propriétaire » qui sont gardées secrètes par les entreprises l'ayant développé.

[11] Par catalogue « monolithique », nous entendons ici un catalogue tout en un, qui gère les fichiers et leur contenu, présente les métadonnées, permet de lire, écouter ou visualiser les fichiers, etc. Cette configuration s'oppose à celle qui distingue, d'une part, le stockage des fichiers autonomes et de leurs métadonnées et, de l'autre, les microservices qui permettent de les gérer et de les exploiter.

[12] Raspberry pi est un petit microprocesseur qui peut fonctionner comme serveur wifi, sans connexion internet : https://www.raspberrypi.org/.

[13] https://language-archives.services/about/data-loader

d'une collection et de sa description en vue de son inclusion dans une archive constitue déjà une avancée majeure pour la plupart des chercheur·euses. Ainsi, même si nous ne pouvons pas garantir la longévité de la collection au-delà des cinq prochaines années, nous stockons toutes les métadonnées avec les éléments de la collection, de sorte qu'il y a moins de risque pour l'ensemble de la collection si le site web et le catalogue venaient à péricliter. Nous sommes convaincus qu'un service national de données qui conservera les données de recherche verra le jour à l'avenir en Australie[14], et la collection de PARADISEC sera un excellent candidat pour être hébergée par un tel service.

3. Un exemple d'archive locale : la base Anareo

Comme il a été dit plus haut, il importe d'encourager l'implantation d'archives plus proches des communautés autochtones et dans lesquelles elles sont plus directement impliquées, conformément aux principes CARE. C'est l'objet de cette section qui présente un programme déployé en Polynésie française en partenariat avec PARADISEC.

Les archipels de Polynésie française présentent une diversité linguistique moins importante que celle que l'on trouve en Mélanésie, mais qui reste inaccoutumée dans l'ensemble ethnolinguistique polynésien, lequel compte une quarantaine de langues. Charpentier et François (2015 : 22) estiment que « le nombre exact de "langues différentes" parlées traditionnellement en Polynésie française oscille entre cinq et huit ». Toutes ces langues présentent des variétés dialectales, l'ensemble le plus diversifié étant le pa'umotu aux Tuamotu, avec neufs sous-ensembles. On compte deux à trois variétés dialectales pour les Marquises ainsi qu'aux îles de la Société, probablement quatre pour les Australes.

Les descriptions disponibles pour ces langues et dialectes sont d'inégales quantité et qualité. Le tahitien, devenu langue d'évangélisation dès le début du XIX^{e} siècle et disposant le premier d'un système d'écriture (Nicole, 1988), a fait l'objet d'une attention soutenue de la part des missionnaires protestants, puis des philologues amateurs et des linguistes. Depuis la première esquisse grammaticale et le premier dictionnaire du pasteur anglican John Davies en 1851, pas moins de six lexiques et dictionnaires et une dizaine de grammaires ont été publiés à ce jour sur cette langue (Vernaudon, 2018). Cependant, en dépit de l'essor de l'outil informatique, jusqu'à récemment les ressources en tahitien et sur le tahitien restaient peu accessibles via internet. Les langues des autres îles ou archipels (Marquises, Tuamotu, Gambier, Australes, Rapa), ou les variantes du tahitien (îles-sous-le-vent, Maupiti) disposent quant à elles de descriptions souvent anciennes ou partielles. Certaines de ces langues sont doublement menacées dans leur transmission, parce qu'elles sont concurrencées par le français et le tahitien. Il existe encore trop peu d'enregistrements sonores référencés de récits et de chants qui permettraient d'en préserver un dense témoignage, utile à des fins de transmission patrimoniale et de recherche. Quant aux langues pour lesquelles des enregistrements ont été réalisés, ces derniers sont dispersés dans des collections publiques ou privées, sans indexation commune et standardisée, sur des supports dont la pérennité est limitée : cassettes audio, minidiscs, Digital Audio Tapes (DAT). Certains experts, comme ceux de l'*International Association of Sound Archives*[16], estiment par exemple qu'au-delà de 2025, les enregistrements sur cassettes audio seront illisibles, soit parce que les bandes elles-mêmes seront dégradées, soit parce que les machines en état de lire les bandes seront difficiles à

[14] On pense, à titre comparatif, à la très grande infrastructure de recherche Huma-Num qui offre en France un support à l'archivage pour la recherche en sciences humaines et sociales.
[16] Cf. https://www.iasa-web.org/blog/magnetic-tape-alert-project-final-report-published

trouver (*cf.* également NFSA, 2014). Ces paramètres compromettent à la fois l'accessibilité des ressources enregistrées et leur conservation sur le long terme.

Pour tenter de combler progressivement ce hiatus technologique et descriptif, mais aussi pour favoriser une interaction plus directe avec les communautés locales, un partenariat est engagé depuis 2013 entre l'Université de la Polynésie française (UPF)[17] et PARADISEC afin de développer une base numérique locale dédiée aux langues autochtones de Polynésie française. Cette base, nommée *Anareo* – littéralement, la « grotte des langues » en tahitien –, a pour vocation d'agréger des données en langues polynésiennes, lexicales et textuelles, écrites ou orales, et de les rendre interopérables.

3.1. Le dictionnaire tahitien-français de l'Académie tahitienne

La première opération réalisée a consisté à se rapprocher de l'Académie tahitienne afin de convertir son dictionnaire tahitien-français (1999), réalisé initialement avec le logiciel Microsoft Word, en une base de données lexicales. Depuis le début de son travail lexicographique en 1989, l'Académie tahitienne a compilé le contenu des dictionnaires et des lexiques antérieurs, en particulier le dictionnaire de John Davies (1851), en normalisant l'orthographe des entrées du dictionnaire, en traduisant les sources anglaises en français, en enrichissant les articles de commentaires encyclopédiques et d'exemples, puis en complétant le vocabulaire du tahitien par des néologismes (ex. *roro uira* pour ordinateur, *hōvai* pour hydrogène, *arutaimāreva* pour environnement). La première édition de 1999 du dictionnaire de l'Académie tahitienne, qui a servi de support à l'opération de conversion en base de données, compte 13 880 entrées lexicales, du tahitien vers le français. En revanche, elle ne comporte pas de liste inverse du français vers le tahitien. La version la plus complète de ce dictionnaire était stockée sur un Mac IIfx âgé de plus de vingt ans dont le système de gestion de la base de données était obsolète. L'Académie tahitienne a exporté une copie de ces données dans plusieurs fichiers Microsoft Word. Thieberger a utilisé le service en ligne OxGarage[18] en identifiant les expressions régulières[19] pour convertir ces fichiers sous la forme d'un document unique au format structuré, compatible avec le logiciel lexicographique Toolbox[20] (fig. 1).

FIGURE 1. – Étapes de la conversion en partant du fichier MS Word : (en haut à gauche) extrait d'un article à partir d'un fichier MS Word source ; (en haut à droite) identification des expressions régulières sous OxGarage ; (en bas) conversion dans un format compatible avec Toolbox avec l'indication des balises de champs.

Ce document a ensuite été converti en une base numérique et un logiciel de gestion en mode SaaS[21] permet d'en modifier les contenus. La base est accessible sur le cloud à partir de tout appareil doté d'un navigateur web, ce qui permet un travail collaboratif avec plusieurs contributeurs, dont des membres référents de l'Académie tahitienne, auxquels est attribué un

[17] Plusieurs composantes internes de l'UPF sont impliquées dans ce programme : l'équipe d'accueil « Sociétés traditionnelles et contemporaines en Océanie » (EASTCO), la Maison des Sciences de l'Homme du Pacifique, la bibliothèque universitaire et la direction des systèmes d'information.
[18] https://oxgarage.tei-c.org/
[19] Les expressions régulières sont des formules qui subsument sous une forme synthétique un ensemble de chaînes de caractères qui partagent des propriétés communes. Pour le traitement automatisé d'une série d'articles de dictionnaire, elles permettent de dégager les éléments constitutifs récurrents de chaque article et de leur assigner une balise (entrée du dictionnaire, partie du discours, définition, exemple, traduction de l'exemple, etc.).
[20] https://software.sil.org/toolbox
[21] SaaS : *Software as a Service*. Dans cette configuration, les données et les applications ne sont pas embarquées sur le matériel informatique de l'utilisateur, mais elles sont stockées sur les serveurs du prestataire.

code d'accès. La base nourrit un site de consultation publique du dictionnaire de l'Académie tahitienne[22]. Ce dictionnaire en ligne, mis en service en juin 2017, a reçu plus de 61 000 visiteurs pour 170 000 consultations en 2020[23].

La conversion du dictionnaire en base de données a été l'occasion d'ajouter des informations supplémentaires pour les mots tahitiens, en particulier :

- leur entrée inverse en français, c'est-à-dire le mot français à partir duquel un utilisateur peut trouver la vedette tahitienne de l'article, ce qui permet une recherche du français vers le tahitien ;
- l'indication du champ sémantique des mots tahitiens (ex. navigation, parenté, plante, etc.) ;
- l'étymon des mots tahitiens, lorsqu'ils sont disponibles sur le site du « *Polynesian Lexicon Project Online* » (POLLEX) (Greenhill et Clark, 2011) : dans le dictionnaire en ligne, les liens intitulés « voir POLLEX » redirigent l'utilisateur vers une page précisant, en l'état actuel des connaissances en linguistique historique polynésienne, l'étymon et les cognats associés à cette entrée.

Nous donnerons ici deux illustrations de l'intérêt de ce travail pour la connaissance des savoirs polynésiens. La production des entrées inverses des 13 880 mots du dictionnaire tahitien-français et l'accès à la base via le site de consultation public de l'Académie tahitienne permet désormais de révéler en un clic la richesse remarquable du vocabulaire tahitien, ancien et contemporain, sur l'environnement. En tapant le mot « mer » par exemple, on obtient 40 résultats (fig. 2).

FIGURE 2. – Écran de visualisation du dictionnaire en ligne de l'Académie tahitienne : premier résultat de la liste, sur 40, pour le mot « mer » (capture réalisée le 06/08/20)

Nous avons extrait ces résultats pour les restituer dans le tableau 1. Les mots ou locutions suivis d'un astérisque dans le tableau ont été collectés par les missionnaires protestants au début du XIX^{e} siècle et sont aujourd'hui hors d'usage et inconnus des académiciens.

TABLEAU 1. – Extraction des résultats obtenus dans le dictionnaire en ligne de l'Académie tahitienne en cherchant les équivalents de traduction du mot « mer » (recherche réalisée le 06/08/20)

Entrées en tahitien	**Définitions**
tai	indique la direction de la mer ; mer ; eau salée ; sel
miti	agitée (mer) ; eau salée, eau de mer ; sel ; mer
moana	profond ; bleu foncé ; océan
aeha'a	la haute mer, le large, l'océan ; difficulté, danger.
*'are fatu moana**	état de la mer caractérisé par un niveau très élevé avec de grosses vagues

[22] http://www.farevanaa.pf/dictionnaire.php
[23] Statistiques de Google Analytics, consultées le 25/04/21.

*aroine**	mer entre le récif et le rivage
'arufe'efe'e	mer très houleuse
*'aruhao**	mer qui se brise d'une manière inhabituelle
*'arumatara**	mer claire et dégagée
*'aru 'ona'ona**	mer continuellement agitée
*‘arupapa’itohe**	mer qui frappe la poupe ; quelqu’un qui dit du mal des gens dans leur dos
*'arupupure**	mer écumante
'aruriri	mer qui se brise en envoyant son écume vers les nuages
*'aru tāhopu**	mer qui se brise et vient tomber aux pieds d'une personne
*'aru tānunanuna**	mer où les vagues se succèdent rapidement
*'aru tāpo'ipo'i**	mer où les vagues se succèdent rapidement
*'aru tūatea**	mauvaise mer qui peut être prévue et en vue de quoi on peut se préparer
*maniataeaha'a**	mer d’huile
moana 'afā	mer qui a des fosses profondes en son milieu
moana fa'a'oa'oa	mer troublée ; émotion interne
moana hāuriuri	mer insondable
*moana pūnao**	mer avec des gouffres au milieu
moana tārere	mer insondable
moana tauana	mer avec des grottes sous-marines
*moana tere-'ore-hia**	mer inexplorée ; femme qui n'a pas eu de relations sexuelles
*moana timatima**	mer sombre
*moana tumatuma**	mer sombre
moana uriuri	mer sombre parce que profonde

taiātea	le large, la pleine mer
*taifa'a'aro**	pleine mer quand aucune terre n'est en vue
taihāuriuri	mer abyssale, reconnaissable à sa couleur bleu sombre
taihorahora	mer quand les vagues commencent à déferler
taihōtū	mer très grosse
*taimara**	mer, quand elle est devenue sacrée à cause d'un chef
taiōtua	mer en dehors des récifs
*taivaha**	mer qui n'existe que dans les récits ou l'imagination de quelqu'un
taivahatete	mer qui ne cesse de briser et de mugir
taivavao	mer qui s'enfle et brise sur les récifs alors que le lagon entre le récif et le rivage est calme et les pâtés de coraux découverts
tua	le large, la haute mer
*tuateaehā**	la pleine mer quand la terre n'est plus en vue

En renouvelant la recherche pour « vent », le dictionnaire en ligne livre 62 noms de vents. Pour « vague », il donne 33 équivalents de traduction en tahitien. Ce dernier corpus d'exemples est d'ailleurs proposé par Manon Sanguinet et Frédéric Torrente dans leur *Guide des aires marines éducatives de Polynésie française* (2020 : 29) pour illustrer « la richesse lexicale relative aux vagues et à la houle et donc […] la connaissance fine des habitants des îles ».

3.2. La régulation des taxonomies vernaculaires

Une autre opération importante, relative à la connaissance autochtone sur l'environnement, est réalisée à partir de la base lexicale du dictionnaire tahitien-français dans *Anareo* grâce au regroupement des mots en champs sémantiques. Ce regroupement permet d'extraire en particulier les noms vernaculaires de plantes (486), de poissons (249) et d'oiseaux (75). Une fonction d'enrichissement taxinomique offre la possibilité à des contributeur·trices spécialisé·es de mettre à jour ou de compléter les noms scientifiques associés aux noms tahitiens d'espèces. Un premier projet a ainsi été engagé en 2019 autour des noms de plantes avec la collaboration du botaniste Ravahere Taputuarai. Les noms de plantes répertoriés dans le dictionnaire tahitien-français de l'Académie tahitienne (1999) correspondent soit à des plantes indigènes (ex. *toi, 'Alphitoniza zizyphoides'*), soit à des plantes introduites (ex. *rā'au marumaru, 'Albizia lebbeck'*), soit à des plantes nommées dans la bible et pour lesquelles les missionnaires protestants du XIX^{e} siècle ont créé un nom tahitien, mais ces espèces sont non représentées en Polynésie française (ex. *hēteta*, 'myrte'). De nombreuses plantes ont été correctement identifiées par l'Académie tahitienne et le dictionnaire fournit comme traduction du nom tahitien à la fois la dénomination vernaculaire française et le nom scientifique. Le

nom binomial établi en 1999 au moment de la première édition du dictionnaire doit cependant parfois faire l'objet d'une mise à jour selon l'évolution de la taxinomie botanique. D'autres noms, recueillis au XIX^{e} siècle, correspondent à des plantes non identifiées ou insuffisamment décrites (ex. l'entrée *nuna* est suivie de la seule mention « arbre qui pousse sur les roches »). Dans ce cas, la contribution des botanistes est d'autant plus indispensable pour tenter d'associer le nom tahitien à un nom scientifique, entre autres grâce à la comparaison avec les autres langues polynésiennes. La fonction de corrélation entre les deux index, l'un vernaculaire, l'autre scientifique, permet de gérer des liens qui ne sont pas tous bijectifs : il arrive qu'un seul nom tahitien corresponde à plusieurs noms scientifiques ou qu'un seul nom scientifique corresponde à plusieurs noms tahitiens (fig. 3).

Figure 3. – Capture d'écran de la fonction taxinomique pour les plantes sous Anareo

L'identification scientifique des plantes associées aux noms vernaculaires est utile à la fois au travail lexicographique sur le tahitien, mais aussi à la recherche en linguistique comparative qui vise à reconstruire l'histoire des langues polynésiennes. Elle permet de suivre l'évolution des taxinomies vernaculaires dans l'espace et dans le temps. Ainsi, au-delà des cas de conservation (on trouve par exemple des reflets du proto-polynésien **talo* dans pratiquement toutes les langues polynésiennes pour *Colocasia esculenta*), un même nom vernaculaire peut désigner deux espèces indigènes différentes dans deux langues polynésiennes (par ex., le nom polynésien *maire* correspond à *Microsorum commutatum* en tahitien, mais à *Nestegis sp.* en māori de Nouvelle-Zélande), ou réciproquement une même espèce peut être nommée différemment dans deux langues (par exemple, *Artocarpus altilis* est nommé *'uru* en tahitien et *mei* en marquisien).

L'index taxinomique est complété par des photos de plantes réalisées par Ravahere Taputuarai lors de ses enquêtes de terrain. Une fois achevées, les fiches et leurs illustrations sont soumises à l'Académie tahitienne en prévision d'une régulation progressive des entrées du dictionnaire et d'une implémentation des illustrations. L'objectif final de ce programme est de fournir aux utilisateurs du dictionnaire tahitien-français en ligne de l'Académie tahitienne, lorsqu'ils recherchent un nom vernaculaire de plante, à la fois une identification scientifique la plus précise possible et des images de la plante pour en faciliter la reconnaissance. Un travail équivalent est engagé pour les noms de poissons et d'oiseaux.

Les outils développés pour le tahitien sont désormais déployés dans d'autres langues polynésiennes. *Anareo* héberge déjà des bases lexicographiques en gestation en mangarevien, en rurutu, en marquisien et en pa'umotu avec sept contributeurs issus des communautés autochtones qui travaillent en ligne.

3.3. L'archive orale Anavevo

Un autre projet s'inscrit dans la collaboration scientifique et technologique engagée entre PARADISEC et l'Université de la Polynésie française, avec le soutien de l'ARC *Centre of Excellence for the Dynamics of Language* (COEDL) et du Fonds Pacifique. Il vise à développer localement des compétences de haut niveau nécessaires à la réalisation d'une archive numérique de corpus oraux dans les langues de Polynésie française. Cette archive, intitulée *Anavevo* (en tahitien, la « grotte des échos »), est intégrée à la base *Anareo.* Le projet s'articule autour du développement informatique d'une plateforme de dépôt, d'archivage et de consultation des enregistrements et la réalisation d'un inventaire des corpus d'enregistrements oraux déjà réalisés en langues autochtones de Polynésie française (service du Patrimoine archivistique et audiovisuel, service de la Culture, Maison de la culture, académies, Société des études océaniennes, collections privées) et de l'évaluation de leur état de conservation et

de leurs conditions juridiques d'accès. Le programme engagé a déjà permis la numérisation et l'archivage d'un corpus de 52 cassettes audio enregistrées dans les années 1980 par Éric Conte, principalement aux Tuamotu, dans le cadre de ses recherches ethno-archéologiques, et de 32 enregistrements en langue mangareva. PARADISEC est en mesure de proposer une numérisation professionnelle des bandes, en utilisant un convertisseur analogique-numérique de haute qualité et des machines de lecture calibrées qui permettent le réglage azimutal des têtes. Si nécessaire, les bandes peuvent également être nettoyées et portées à basse température et, dans les cas extrêmes, lubrifiées pendant la lecture, pour garantir la capture optimale du signal.

Toujours dans le cadre de ce projet, PARADISEC a assuré une mission de formation à Tahiti sur la création et la gestion d'archives linguistiques numériques en novembre 2019, auprès d'une quarantaine de participants, dont des membres des académies de langues locales, des représentants des services du gouvernement de la Polynésie française engagés dans la préservation du patrimoine culturel (service de la culture, service du patrimoine archivistique et audiovisuel) et des chercheurs spécialistes des langues polynésiennes. Le développement d'un prototype pour l'archivage et la consultation de corpus audio ou vidéo a débuté à l'occasion de cette mission. La crise sanitaire de la Covid-19 a retardé les étapes de mise en place de la base, mais la livraison de l'interface de consultation publique est prévue pour novembre 2021.

Le lancement du projet *Anavevo* a déjà eu des incidences méthodologiques sur les enquêtes de terrain réalisées en Polynésie française. En guise d'illustration, nous livrons ici un retour d'expérience à la suite d'une recherche réalisée par une équipe du *Rāhui Forum and Resource Center*. Le terme *rāhui*, 'prohibition, interdire', désigne en tahitien un ancien mode de gouvernance et de gestion des ressources utilisé par les chefs et les *tahu'a*, spécialistes traditionnels. Pendant des siècles, le *rāhui* a été utilisé par les pêcheurs-horticulteurs au sein des communautés locales pour assurer la santé et la durabilité à long terme des ressources terrestres et marines. Différentes formes de *rāhui* sont encore pratiquées en Polynésie française aujourd'hui et sont généralement des systèmes socio-écologiques axés sur un bassin versant et le milieu marin côtier adjacent. Les *rāhui* émergent traditionnellement d'une communauté locale en fonction de ses besoins et de ses préoccupations environnementales et politiques et sont gérés par la communauté sur la base d'une vision partagée et de pratiques bien établies.

Des recherches-actions récentes (cf. *inter alia*, Bambridge, 2016 ; Bambridge *et al.*, 2019) ont encouragé un dialogue entre les *leaders* communautaires – aujourd'hui principalement les maires –, les *tahu'a* et des chercheur·euses en sciences naturelles et humaines, afin de favoriser le développement, l'évaluation et le pilotage des *rāhui* comme moyen de maintenir la résilience culturelle polynésienne et la gestion des ressources naturelles, et d'améliorer la gestion de l'environnement, la prévention de la pollution et la gestion des pêches.

Pour renforcer les sites de *rāhui* existants et en étendre le réseau sur l'ensemble de la Polynésie française, un Forum et un centre de ressources du Rāhui (*Rāhui Forum and Resource Center*, RFRC) a été créé afin d'accompagner les communautés de pêcheurs et autres parties prenantes qui utilisent les lagons et les environnements insulaires terrestres associés en fournissant à ces communautés les informations, les compétences et les outils pour une gestion durable de ces espaces.

Parmi les actions du RFRC, des campagnes de sensibilisation sont réalisées auprès des communautés locales, accompagnées d'un recueil de témoignages et de récits liés au *rāhui*. Dans le cadre d'un partenariat avec l'UPF en appui sur la base *Anavevo*, il s'agit d'archiver durablement les données recueillies sur le terrain, dès l'étape de l'enquête, de les rendre

accessibles aux communautés locales grâce aux technologies internet et d'assurer l'interopérabilité de ces données avec les principales archives mondiales.

Une première enquête dans le cadre de ce partenariat a été réalisée en juillet 2020 à Bora-Bora, pendant dix-huit jours, auprès de 26 pêcheurs et acteurs culturels ou environnementaux. 21 hommes et 5 femmes ont participé à l'enquête. Parmi eux, 11 hommes et 1 femme sont âgés de 64 à 78 ans, 13 d'entre eux dont les 4 autres femmes ont entre 39 et 54 ans et le dernier participant est âgé de 27 ans. L'enquête portait sur le thème de la pêche hier, aujourd'hui et demain, avec pour objectif de documenter, au travers d'enregistrements de qualité, les techniques, les lieux et l'organisation de la pêche en laissant une place importante à ce qui fait patrimoine du point de vue des pêcheurs. L'enquête était conduite par Takurua Parent, étudiante polynésienne en master 2 (« Langues, cultures, et sociétés océaniennes » à l'UPF), chargée de présenter le projet, d'obtenir le consentement des personnes enquêtées et d'enregistrer les interventions. Elle était accompagnée dans l'enquête par Patrick Rochette, pêcheur de Teahupoo, expert local dans les histoires anciennes de Taiarapu (partie est de l'île de Tahiti) et membre du comité de gestion du *rāhui* de Teahupoo au titre de la culture, et de Rakamaly Madi Moussa, biologiste au Centre de recherches insulaires et observatoire de l'environnement (CRIOBE), chargé de géo-référencer à l'aide d'un GPS les lieux d'entretien et les sites remarquables indiqués par les personnes enquêtées. L'ensemble de ce travail était supervisé sur le terrain par Tamatoa Bambridge.

Les entretiens ont été menés en langue tahitienne, dans la variété dialectale pratiquée à Bora-Bora. Chaque entretien dure entre une et trois heures, en fonction des personnes. L'entretien était semi-directif, les enquêteurs se chargeant de relancer la discussion ou d'aborder un autre thème en fonction du déroulement. Mais d'entretiens semi-directifs, les enquêtes sont rapidement devenues des dialogues, des échanges d'expérience, où Patrick Rochette s'attachait à faire émerger les noms anciens, à demander des explications sur des mots inconnus des enquêteurs.

Pour certaines personnes, l'entretien était complété par une visite de sites, terrestres ou lagonaires, au cours de laquelle des points géo-référencés étaient enregistrés pour donner un cadre géographique aux histoires (techniques, lieux de pêche, formes d'organisation de pêche collective). Ces déplacements ont donné lieu à de nouveaux enregistrements. Au final, l'ensemble des entretiens représente un volume d'environ 28 heures d'enregistrements audio, lesquels seront déposés progressivement sur *Anavevo* avec plusieurs niveaux d'accessibilité en fonction du vœu des personnes enquêtées.

La perspective d'un archivage pérenne a conduit les enquêteurs à apporter un soin particulier dans l'information aux personnes enquêtées sur le cadre du recueil de leurs témoignages et pour l'obtention de leur consentement éclairé, à l'oral puis par écrit, avec l'indication du degré d'accessibilité souhaité. L'interlocuteur peux faire le choix d'être mentionné ou non avec son enregistrement.

L'optique d'un archivage sur le long terme a été source de motivation pour les informateurs pour qui archivage et conservation riment avec partage et héritage à léguer aux enfants de l'île. D'autre part, les enquêteurs y ont vu l'occasion d'enrichir la connaissance scientifique grâce à la centralisation des données permettant une accessibilité plus facile pour les chercheur·euses. C'est ainsi que chacun des membres de l'équipe, puisque issus de domaines de recherche variés (biologie, anthropologie et linguistique), a contribué de manière considérable au recueil de propos inédits par sa participation aux entretiens.

Beaucoup des personnes sollicitées regrettent un manque d'intérêt de la part de leur entourage pour les savoirs qu'elles souhaitent partager. Leur désir de transmission, déjà manifeste en amont de l'entretien, s'accroît une fois la démarche d'archivage et de mise en ligne des

enregistrements expliquée en détails. Confier leurs connaissances – excepté les confidences au sujet de discordes avec d'autres individus – à des fins d'archivage et de consultation semble être un acte voulu et assumé. Leur principale motivation demeure le partage de leur héritage patrimonial avec les générations futures de leur île qui pourront consulter facilement et librement ces témoignages.

Il est important de noter que beaucoup des pêcheurs rencontrés vivent principalement de la pêche. En ce sens, l'enregistrement, l'archivage et la consultation des propos des enquêtés peut, à quelques occasions et de façon tout à fait légitime, les conduire à éviter de dévoiler leurs techniques et lieux de pêche actuels de peur que d'autres pêcheurs en prennent connaissance et s'en emparent. De ce point de vue, les enregistrements peuvent manquer d'informations. Ils n'en demeurent pas moins d'une grande richesse. Ayant appris le déroulement des enquêtes menées par l'équipe, des habitants de l'île sensibles aux questions de préservation du patrimoine oral local, pour des questions de réappropriation identitaire et culturelle semble-t-il, ont à leur tour volontairement débuté l'enregistrement de récits en vue d'un dépôt dans la base *Anavevo*. Toujours dans l'optique d'un archivage numérique pérenne et de qualité, des améliorations techniques pourraient être soumises à réflexion telles que l'utilisation d'outils plus adaptés (ces « enquêteurs volontaires » enregistrent à l'aide de leur smartphone) et l'initiation à des modalités de recueil différentes visant, par exemple, à amener les enquêteurs volontaires à réfléchir à la relation enquêteur/enquêté ou encore apprendre à ne pas imposer leur point de vue aux informateurs.

4. Conclusion

Les archives linguistiques remplissent plusieurs fonctions importantes et complémentaires et il est nécessaire qu'elles soient plus largement utilisées par les chercheur·euses qui recueillent des témoignages linguistiques et culturels singuliers. Les archives doivent permettre la restitution des documents à la communauté source selon les principes « FAIR » tels que discutés pour PARADISEC par Barwick et Thieberger (2018) : les documents sont trouvables, accessibles, interopérables, réutilisables. Les archives offrent une sauvegarde sûre aux chercheur·euses qui peuvent continuer à accéder à leurs propres fichiers au fil du temps. Elles rendent les enregistrements disponibles pour que d'autres puissent les utiliser, avec des licences qui expliquent ce qui peut et ne peut pas être fait avec ces sources. Elles publient des métadonnées pour faciliter au maximum la recherche d'enregistrements. Elles comprennent la nature des documents linguistiques et peuvent donc, par exemple, présenter les transcriptions et les médias ensemble et permettre la recherche d'occurrences de mots ou d'expressions grâce aux transcriptions. Elles participent à l'agrégation des enregistrements à l'échelle internationale, ce qui permet de générer une page web comportant les ressources répertoriées pour chaque langue[24]. Elles fournissent des recommandations sur la façon de réaliser des enregistrements sur le terrain, de les annoter et de les enrichir au fil du temps, ce qui favorise ainsi la mise en place d'une méthodologie explicite de l'enquête. Elles établissent un identifiant permanent pour les données, ce qui permet de les citer dans la recherche, afin que d'autres puissent vérifier l'existence de ces éléments empiriques et qu'ils puissent entendre l'énoncé ou voir l'événement qui est cité dans le cadre d'une démonstration théorique.

En raison à la fois de leur rôle essentiel pour le soutien de la recherche et des enjeux d'accessibilité pour les communautés source, il faut encourager la création de davantage

[24] C'est ce que fait par exemple le site OLAC (Open Language Archives Community) http://www.language-archives.org/

d'archives numériques dans le monde. Il existe actuellement dix-huit archives qui sont membres du réseau d'archives numériques des langues et des musiques en danger (*Digital Endangered Languages and Musics Archives Network*, DELAMAN), et soixante-et-un projets répertoriés par la communauté des archives linguistiques ouvertes (*Open Language Archives Community*, OLAC), dont environ la moitié sont de véritables archives linguistiques (les autres étant des répertoires de ressources linguistiques).

Les enregistrements d'expressions culturelles, comme les savoirs traditionnels transmis dans des histoires et des chansons par exemple, peuvent offrir une puissante ressource identitaire et jouer un rôle important dans le renouveau culturel, en particulier lorsque les forces coloniales ont empêché leur transmission aux jeunes générations. Mais si ces enregistrements restent sous un format analogique dans les centres métropolitains, ils sont pratiquement inaccessibles pour les membres des communautés où ils ont été réalisés et ils ne servent pas les intérêts de ces communautés.

Bibliographie

BAMBRIDGE Tamatoa (dir.), 2016. *The Rāhui: Legal pluralism in Polynesian traditional management of resources and territories*, Canberra, Australian National University press.

BAMBRIDGE Tamatoa, François GAULME, Christian MONTET et Thierry PAULAIS, 2019. *Communs et océan*, Papeete, Au Vent des îles et Agence française de développement.

BARWICK Linda, 2004. Turning it all upside down… Imagining a distributed digital audiovisual archive, *Literary and Linguistic Computing* 19 (3), pp. 253-263.

BARWICK Linda et Nick THIEBERGER, 2018. Unlocking the archives, in V. Ferreira et N. Ostler (dir.), *Communities in Control: Learning tools and strategies for multilingual endangered language communities. Proceedings of the 2017 XXI FEL conference*, Hungerford, FEL, pp. 135-139.

BILLINGTON, Rosey, Janet FLETCHER, Nick THIEBERGER et Ben VOLCHOK, 2018. Acoustic correlates of prominence in Nafsan, in J. Epps, J. Wolfe, J. Smith, & C. Jones (dir.), *Proceedings of the 17th Australasian International Speech Science and Technology Conference*. Sydney, Australasian Speech Science and Technology Association, pp. 137-140.

CAMPBELL Lyle, Nala Huiying LEE, Eve OKURA, Sean SIMPSON et Kaori UEKI, 2013. New Knowledge: Findings from the Catalogue of Endangered Languages (ELCat). *3rd International Conference on Language Documentation and Conservation*, 28 février 2013.

CHARPENTIER Jean-Michel Alexandre FRANÇOIS, 2015. *Atlas linguistique de la Polynésie française*, Berlin et Papeete, Mouton de Gruyter et Université de la Polynésie française.

DAVIES John, 1851. *A Tahitian and English Dictionary, with Introductory Remarks on the Polynesian Language and a Short Grammar of the Tahitian Dialect*, Tahiti, London Missionary Society Press.

FOLEY Ben, Alina RAKHI, Nicholas LAMBOURNE, Nicholas BUCKERIDGE et Janet WILES, 2019. Elpis, an accessible speech-to-text tool. *Interspeech 2019: Show & Tell Contribution* Graz, Austria.

FRANÇOIS Alexandre, 2018. In search of island treasures: Language documentation in the Pacific, in B. McDonnell, A. L. Berez-Kroeker et G. Holton, Reflections on Language Documentation, 20 Years after Himmelmann 1998, *Language Documentation & Conservation Special Publication* 15, pp. 276-294.

HARRIS Amanda, GAGAU Steven, KELL Jodie, THIEBERGER Nick et Nick WARD, 2019. Making Meaning of Historical Papua New Guinea Recordings: Collaborations of Speaker Communities and the Archive. *International Journal of Digital Curation*, 14, pp. 136–149

HENRY Teuira, 1993. *Tahiti aux temps anciens*, Paris, Société des Océanistes.

HIMMELMANN Nikolaus, 1998. Documentary and descriptive linguistics, *Linguistics* 36 (1), pp. 161–195.

—, 2006, Language documentation: What is it and what is it good for?, in G. Jost, N. Himmelmann et U. Mosel (dir.), *Essentials of Language Documentation, Trends in Linguistics,* Studies and Monographs 178, Berlin & New York, Mouton de Gruyter, pp. 1-30.

JACOBSON Michel, Boyd MICHAILOVSKY et John B. LOWE, 2001. Linguistic documents synchronizing sound and text, *Speech Communication* 33, pp. 79-96.

KRAJINOVIC Ana, 2019. *Tense, mood, and aspect expressions in Nafsan (South Efate) from a typological perspective: The perfect aspect and the realis/irrealis mood.* PhD Dissertation. University of Melbourne/Humboldt University.

KRAUSS Michael, 1992. The World's Languages in Crisis, *Language* 68 (1), pp. 4-10.

NFSA (National Film and Sound Archive of Australia), 2014. *Deadline 2025 : collections at risk.* https://www.nfsa.gov.au/collection/curated/deadline-2025-0

NICOLE Jacques, 1988. *Au pied de l'écriture*, Papeete, Haere po no Tahiti.

SANGUINET Manon et Frédéric TORRENTE, 2020. *Moana. Le Chemin de l'Océan. Guide des Aires marines éducatives de Polynésie française*, Papeete, OFB, MSHP et DGEE.

THIEBERGER Nicholas, 1995. The Aboriginal Studies Electronic Data Archive (ASEDA), *International Journal of the Sociology of Language* 113, pp. 147-149.

—, 2004. Documentation in practice: Developing a linked media corpus of South Efate, in P. Austin (dir.), *Language documentation and description* 2, London, Hans Rausing Endangered Languages Project, SOAS, pp. 169-178.

—, 2020. Technology in Support of Languages of The Pacific: Neo-Colonial or Post-Colonial?, *Asian-European Music Research Journal*, 5, pp.17-24.

THIEBERGER Nicholas et Amanda HARRIS, 2020. Be Not Like the Wind: Access to Language and Music Records, Next Steps. *Proceedings of the Language Technologies for All* (LT4All), pp.101-103

VERNAUDON Jacques, 2018. Les métalangues du tahitien à l'école, *Contextes et Didactiques* [En ligne], 12 , mis en ligne le 15 décembre 2018, URL : https://www.contextesetdidactiques.com/1129