Le Digital Object Identifier, une impérieuse nécessité ? L'exemple de l'attribution de DOI à la Collection Pangloss, archive ouverte de langues en danger

Aurelia Vasile, Séverine Guillaume, Mourad Aouini, Alexis Michaud

Accepté pour publication en 2020 dans *I2D - Information, données & documents*,
revue de l'Association des professionnels de l'information et de la documentation (ADBS)

Le *Digital Object Identifier*, une impérieuse nécessité ? L'exemple de l'attribution de DOI à la Collection Pangloss, archive ouverte de langues en danger

Aurelia VASILE, Séverine GUILLAUME, Mourad AOUINI, Alexis MICHAUD

La Collection Pangloss est une archive ouverte de langues à tradition orale, née il y a plus de vingt ans d'une vision qu'on dirait aujourd'hui de Science ouverte : elle consiste à associer étroitement documentation et recherche, pour leur bénéfice mutuel. L'article revient sur les problématiques d'identification de la ressource, et sur le contexte « socio-scientifique » actuel, pour mettre en perspective le choix – qui peut paraître paradoxal – d'attribuer un Digital Object Identifier (DOI) à chaque document de la Collection Pangloss. Les étapes de la mise en œuvre sont également abordées, dans leurs dimensions méthodologiques et techniques.

MOTS CLÉS : Collection Pangloss, DataCite, DOI, Identifiants numériques, Métadonnées, Science ouverte

Aurelia VASILE
Diplômée du master Technologies Numériques Appliquées à l'Histoire de l'École nationale des chartes, ingénieure d'études au CNRS, Aurelia VASILE travaille dans le traitement et l'analyse des données scientifiques à la Maison de Sciences de l'Homme de Clermont-Ferrand.

Séverine GUILLAUME
Informaticienne spécialisée en Traitement Automatique des Langues, ingénieure d'études au CNRS, Séverine GUILLAUME gère la Collection Pangloss et cogère la plate-forme Collection de Corpus Oraux Numériques (Cocoon).

Mourad AOUINI
Informaticien spécialisé en Traitement Automatique des Langues, ingénieur d'études au CNRS, Mourad AOUINI met en place des solutions TAL en exploitant des recherches, des méthodes et des techniques issues de la science des données (Data-science), de l'intelligence artificielle (l'apprentissage profonds) ainsi que des formalismes symboliques.

Alexis MICHAUD
Chercheur au CNRS, Alexis MICHAUD mène des enquêtes de terrain au sujet de langues tibéto-birmanes et austroasiatiques. Il a à cœur de contribuer au développement de la collection Pangloss, archive ouverte de langues rares et menacées

Digital Object Identifiers as an absolute must? Why DOIs were assigned in the Pangloss Collection, an open archive of endangered languages

The Pangloss Collection is an open archive of endangered languages, developed over more than twenty years in what would currently be called an Open Science perspective: to arrive at a close association between documentation and research, for their mutual benefit. The article looks back at the general topic of resource identification, and at the current “socio-scientific” context, in order to put into perspective the somewhat paradoxical choice of assigning a Digital Object Identifier (DOI) to each document in the Pangloss Collection. The stages of implementation are also outlined, broaching methodological as well as technical topics.

KEYWORDS : DataCite, DOI, Digital Identifiers, Metadata, Pangloss Collection, Open Science

1. Contexte général : Science ouverte et données FAIR

L’ouverture des données de la recherche est un volet fondamental de ce qu’on désigne aujourd’hui comme la Science ouverte *(Open Science)*. Le mouvement vers une science ouverte n’est rien de moins qu’« un retour aux sources de ce que MERTON appelait "l’Ethos de la Science", à savoir les valeurs et les normes morales encadrant l’activité des membres de la communauté scientifique » (MAUREL 2019). Ce mouvement paraît actuellement appelé à s’amplifier, fort du soutien croissant des communautés de chercheurs (qui prennent notamment conscience de la ponction opérée sur les budgets de recherche et d’enseignement par des maisons d’édition en situation d’oligopole), des pouvoirs publics[1], et d’une législation qui établit les droits et les devoirs des institutions de recherche et des chercheurs[2].

1. Ministère de l’Enseignement supérieur, de la Recherche et de l’Innovation, *Plan national pour la science ouverte,* 4 juillet 2018. Ce document représente plus qu’une recommandation et amorce une véritable politique publique autour de la science ouverte. http://cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf (consulté le 29 novembre 2019).

2. Code des relations entre le public et l’administration (CRPA), la Loi pour une République numérique du 7 octobre 2016, Code de la propriété intellectuelle (CPI).

Néanmoins, si « le besoin d'une Science Ouverte a été invoqué de tout temps, et par tous les bords politiques » (HOCQUET 2019, 531), le cheminement vers un libre accès aux publications scientifiques est tortueux et jalonné de paradoxes (SUBER 2016), et la mise à disposition des données collectées dans le cadre des projets de recherche est dans une situation plus complexe encore : dans une phase de tâtonnement méthodologique, technologique et juridique. Le partage des données tarde à prendre de l'ampleur dans un milieu de la recherche qui accorde trop souvent aux publications une primauté quasi-exclusive dans l'évaluation.

> Faute d'un signal extrêmement clair et crédible indiquant que, s'il joue le jeu de l'ouverture des données qu'il a accumulées, le chercheur sera reconnu, apprécié et récompensé d'une manière quelconque mais significative, il lui sera très difficile de s'engager spontanément dans cette voie. Là encore, vaincre la tendance individualiste au profit de la collectivité demandera un effort considérable, bien concerté et simultané. On perçoit immédiatement l'ampleur de la difficulté d'une telle mise en place.
> (RENTIER 2018, 29)

C'est dans ce contexte que s'inscrit l'initiative décrite ici : l'attribution de DOI à la Collection Pangloss, archive ouverte de langues en danger. L'approche adoptée consiste à écouter le souhait formulé par des chercheurs aux yeux de qui le *Digital Object Identifier* constitue une impérieuse nécessité (on reviendra sur les facteurs qui façonnent cette conviction). Le déploiement de ces identifiants vise à contribuer à une meilleure intégration des dimensions complémentaires que constituent les publications, les données et les outils de la science. Autrement dit, le DOI, qui ne s'imposait pas comme une nécessité archivistique, est employé comme cheval de bataille pour la diffusion de pratiques de Science Ouverte.

Après une présentation du projet scientifique qui sous-tend la Collection Pangloss, on reviendra sur les problématiques d'identification de la ressource et sur les diverses solutions existantes (types d'identifiants pérennes), avant de présenter en détail les identifiants DOI – qui, avons-nous découvert au fil du processus de déploiement, gagnent à être connus, car ils présentent plusieurs avantages. Enfin, nous en viendrons aux premiers éléments de bilan qui peuvent en être tirés, et aux nouvelles perspectives qui s'ouvrent.

2. La Collection Pangloss : une archive ouverte de langues rares

La Collection Pangloss (MICHAILOVSKY et al. 2014) est une archive ouverte d'enregistrements de langues rares. Depuis ses débuts dans les années 1990, elle s'inscrit dans une logique de conservation, de structuration et de mise à disposition[3] de données de la recherche qui présentent une haute valeur patrimoniale. Les documents qu'elle héberge sont des fichiers audio ou vidéo, accompagnés (pour partie) de fichiers d'annotation (transcription, traduction et notes). Ces ressources sont des récits de littérature orale, des discours spontanés (récits de vie, récits d'artisans...) et autres matériaux recueillis sur le terrain dans des aires géographiques variées. La Collection Pangloss comprend actuellement des ressources dans plus de 170 langues peu documentées, majoritairement sans tradition écrite. Les ressources de la Collection Pangloss sont consultables sur un site web dédié (https://pangloss.cnrs.fr/).

La Collection Pangloss fait partie des collections (plus de trente à ce jour) hébergées par la plate-forme Cocoon (COllections de COrpus Oraux Numériques), lesquelles partagent la caractéristique d'être 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 plateformes 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.

La Collection Pangloss a d'emblée été conçue comme un maillon dans un dispositif mondial structuré autour de normes pour les métadonnées, en particulier OLAC (*Open Language Archive Community*), et de protocoles ouverts de partage (OAI-PMH: *Open Archives Initiative Protocol for Metadata Harvesting*). Cocoon est associée au consortium européen CLARIN (*Common Language Resources and Technology Infrastructure*) par le biais de la Très Grande Infrastructure de Recherche Huma-Num, qui en est l'un des partenaires. La Collection

3. Si les questions juridiques et éthiques n'entrent pas dans le cadre du présent exposé, on précisera néanmoins que l'accès aux données est « aussi ouvert que possible, et aussi restreint que nécessaire ». Des considérations dictées par le respect des locuteurs et des cultures étudiées peuvent amener à restreindre l'accès à certains documents (pour une durée déterminée en fonction de la situation). Le déposant peut également différer l'ouverture des données afin de se réserver une exclusivité d'exploitation. Cela concerne en particulier des chercheurs non titulaires, afin d'éviter que des collègues plus expérimentés ne tirent parti avant eux de leurs données, ce qui risquerait de les cantonner dans un rôle ancillaire de « pourvoyeurs de données ».

Pangloss, grâce à l'adoption de standards descriptifs et technologiques, répondait aux exigences des données FAIR avant même que ce concept (lancé en 2007 comme un des principes de l'OCDE en faveur de l'accès aux données de la recherche financée sur fonds publics (OCDE 2017) n'occupe l'espace discursif de la science ouverte : des données Faciles à trouver, Accessibles, Interopérables, Réutilisables.

Pourtant, ce qui est facile à trouver pour certains utilisateurs ne l'est pas nécessairement pour d'autres. La problématique de l'identification de ressources numériques, capitale dans le monde de l'Information scientifique et technique d'aujourd'hui, se pose différemment selon les catégories d'utilisateurs et force est de constater qu'il n'existe pas aujourd'hui un unique système d'identifiants qui satisfasse à tous les besoins. Pour aborder cette question et expliquer pourquoi le DOI est apparu comme une nécessité pour la Collection Pangloss, les paragraphes qui suivent détaillent plusieurs systèmes d'identifiants et leurs objectifs en termes de pérennisation et de valorisation. Il s'agit de préciser la place des DOI dans le monde des identifiants numériques et les enjeux de leur utilisation dans le cadre du projet Pangloss.

3. Les systèmes d'identifiants et leurs objectifs

3.1 Localisation et identification dans le contexte du web

Inutile de rappeler que la bonne identification des documents est une préoccupation de longue date des bibliothécaires et archivistes, qui ont utilisé des systèmes de cotation et de numérotation pour garantir l'unicité d'un document. Avec l'apparition du web, la notion d'identifiant a atteint une autre dimension, car elle fait partie intégrante de l'architecture du web : elle est l'une de ses trois notions fondatrices, aux côtés de la notion de *représentation* et de celle de *ressource* (BERMÈS et POUPEAU 2012).

Dans une première phase, l'URL (*Uniform Resource Locator*) identifiait un document sur le web par le biais de sa localisation sur un serveur. Mais la fragilité des liens URL, sensibles au renommage des fichiers et au changement de la structure des sites ou des noms de domaine, hypothéquait d'emblée leur utilisation pour la citation des ressources du web (HILSE et KOTHE 2006). Un déplacement de paradigme s'est rapidement produit, du localisateur (l'URL) vers

l'identifiant (l'URI : Uniform Resource Identifier). Ce dernier est mis en avant depuis la fin des années 1990 par le consortium W3C comme un concept sensiblement différent de l'URL, car il est défini comme un identifiant unique de ressource et non pas comme un indicateur de localisation (MASINTER, BERNERS-LEE, et FIELDING 2005). Cette nuance est importante car un URI est conçu pour assurer l'identification d'une ressource indépendamment de sa localisation sur le web. Cela signifie qu'un URI peut ne pas être actionnable.

C'est sur ce même principe que s'appuient des mécanismes d'identification qui ont ambitionné de résoudre le problème de périssabilité des URL. Déployés sous l'égide d'autorités mondialement reconnues, ils ont été présentés comme des systèmes d'identifiants pérennes. Les premières propositions se sont cristallisées autour des identifiants PURL (*Persistent uniform resource locator*) et *Handle*. Elles ont ouvert la voie des identifiants généralistes permettant d'identifier, sur le web, une typologie variée d'objets. PURL a été mis en place par l'OCLC (*Online Computer Library Center*) en 1995. L'usage du terme « identifiant » pour désigner le PURL est néanmoins impropre, parce qu'il ne garantit pas l'unicité globale du nom de la ressource. C'est une technologie de redirection du lien URL permettant de retrouver une donnée même si la localisation a changé, à condition de mettre à jour la nouvelle localisation sur le serveur qui assure la redirection. *Handle*, développé à partir de 1994 et géré aujourd'hui par le CNRI (*Corporation for National Research Initiatives*), est un dispositif qui comprend une suite logicielle de gestion d'identifiants, une organisation et des services. Il fournit également un mécanisme de résolution des identifiants. (C'est sur ce même système que fonctionne le DOI, *Digital Object Identifier*, dont il sera abondamment question plus bas.)

En 2001, un autre type d'identifiant généraliste a été lancé, le ARK (*Archival Resource Key*), à l'initiative de *California Digital Library*, qui attribue des identifiants d'autorités nommantes uniques et fournit des règles de constitution des identifiants (BERMES 2006). À la différence du *Handle* ou du DOI qui requièrent le paiement de frais de gestion et maintenance, l'identifiant ARK est gratuit. En revanche, hormis l'attribution et la gestion des numéros pour l'autorité nommante (le NAAN : *Name Assigning Authority Number*) qui est assuré par la *California Digital Library*, la responsabilité de la création, de la structuration et de la résolution de l'identifiant incombe au producteur de la donnée. L'une des particularités propres à

l'identifiant ARK est la possibilité de gérer (par le biais du *qualifier*) les différents niveaux de granularité d'une ressource : par exemple les pages d'un livre, ou les phrases d'un texte.

Ces quatre identifiants (PURL, Handle, ARK, DOI) sont les plus répandus dans les établissements en charge de la gestion des ressources patrimoniales, les centres de recherche, et les maisons d'éditions. Toutefois, il existe pléthore d'autres identifiants numériques relevant de domaines plus spécifiques (médecine, sciences, biologie, agriculture...), recensés par des services d'enregistrement et validation tels que *identifiers.org*.

Conservateurs, documentalistes et informaticiens soulignent les précautions à prendre à l'égard de la promesse que suggère l'emploi du terme « pérenne ». En dépit de l'appui sur une conception technique et sur des autorités internationales de nommage et de gestion qui confèrent en effet certaines garanties de durabilité, ces identifiants ne sauraient être dits pérennes au sens du temps long des bibliothèques et des archives. D'une part, la responsabilité du maintien pérenne de l'accès aux données repose sur le producteur de données (KUNZE 2003, 6) ; d'autre part, la pérennité des organisations sur lesquelles reposent les identifiants n'est pas garantie.

De la sorte, il est impropre de placer les identifiants sur une unique échelle de valeurs. L'identifiant parfait n'existe pas, chacun ayant ses avantages et ses limites. Le choix d'identifiants est conditionné par plusieurs paramètres, dont les pratiques d'identification liées aux métiers ainsi que les capacités techniques d'une institution à gérer son propre système de nommage et à assurer la résolution des identifiants.

En comparaison des identifiants PURL, Handle et ARK, qu'est-ce qui fait la spécificité du DOI, et peut justifier son déploiement ?

3.2 L'identifiant numérique d'objet (DOI) : repères historiques

Dans le monde des identifiants, le DOI a acquis une notoriété auprès des chercheurs qui lui vaut à l'heure actuelle de faire autorité dans le système international d'identification des publications et données scientifiques. A titre d'exemple l'INRA recommande spécifiquement

le DOI comme modalité principale pour identifier la donnée (INRA 2019). De même, le SCD de l'Université de Bordeaux Montaigne a fait le choix du DOI dès 2016. (BAUDRY ET MACHEFERT 2016).

Le système DOI est né d'une initiative du monde de l'édition, qui a éprouvé le besoin d'un équivalent numérique aux ISSN (*International Standard Serial Number*) et ISBN (*International Standard Book Number*), mais aussi d'une méthode pour identifier les différents niveaux de granularité de leurs ressources (par exemple, les articles parus successivement dans une même revue). La décennie 1990 connaît « une croissance soudaine de l'intérêt pour les méthodes d'identification d'un contenu dans l'environnement numérique » (GREEN et BIDE 1997).

Le DOI est un d'identifiant dont le mécanisme technique de résolution est basé sur le système Handle (PASKIN 1999). Le DOI est lancé en 1997 (PASKIN 2015). Au même moment est créé l'*International DOI Foundation* (IDF) avec pour but de développer ce système et d'en assurer la maintenance. Depuis 2012, le DOI est devenu la norme ISO 26324:2012. La Fondation DOI est désignée comme l'autorité garantissant le respect de la norme et l'unicité des identifiants. La partie opérationnelle du système est assurée par des agences d'enregistrement dont Crossref, qui gère les DOI pour les publications scientifiques, et DataCite, qui se consacre à l'enregistrement des DOI pour les données de la recherche.

La croissance exponentielle du volume des données numériques de la recherche, et leur diffusion sur le web, suscite en effet chez certains chercheurs un intérêt pour la question des identifiants. La nécessité de vérifier les sources et la perspective de produire de nouvelles recherches à partir des données existantes, a inspiré le projet d'étendre le système DOI aux données de la recherche. DataCite, qui détient le rôle d'agence d'enregistrement et de maintenance pour les DOI des données de recherche, a été créée en 2009 sous la forme d'un consortium de plusieurs établissements producteurs de données. La France a intégré le consortium par l'Institut de l'information scientifique et technique (InIST) du CNRS, ce qui offre la possibilité à toutes les unités CNRS de bénéficier des services de DataCite. Il est à noter que le DOI est un système payant, et que l'appartenance au CNRS n'exempte pas l'unité de recherche d'un versement financier pour bénéficier de ces services, actuellement de 180

euros par an. (En revanche, l'abonnement permet la création d'un nombre illimité de DOI : le paiement se fait par abonnement annuel fixe, et non au nombre de DOI créés.)

Le DOI suit les spécifications pour la structuration d'une URI (*International DOI Foundation*, 2029) et se construit comme dans la Figure 1 :

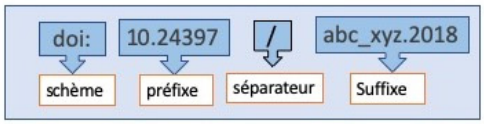

Figure 1. Nomenclature DOI

- un schème qui précise le contexte dans lequel l'identifiant est défini (doi:);
- un préfixe numérique désignant l'autorité nommante (dans notre cas, le laboratoire LACITO) assigné par l'organisme qui assure au niveau mondial la gestion des identifiants (DataCite);
- un suffixe alphanumérique, qui peut être opaque ou signifiant, défini par l'autorité nommante. Cet identifiant peut exister déjà et identifier la ressource de manière unique au niveau local ou il peut être nouveau.

La manière dont l'identifiant est utilisé dans les habitudes de citation appelle quelques commentaires. Cette pratique a évolué au fil du temps. Ainsi, dans les premières années suivant la création du DOI, les agences d'enregistrement recommandaient la formule *doi:préfixe/suffixe* partant du principe qu'un jour le protocole *doi:* serait intégré dans les navigateurs et que la résolution des DOI se ferait automatiquement (Crossref 2017). De nombreuses publications contiennent des exemples de cette pratique d'identification, qui reste recommandée dans les articles méthodologiques (*Digital Object Identifier*, 2011). Le protocole *doi:* n'ayant pas été intégré aux navigateurs, les habitudes de citation des DOI par les utilisateurs ont privilégié les bénéfices immédiats du lien actionnable, lequel mène en un clic à la ressource.

Aujourd'hui, cette pratique s'est imposée même au niveau des agences d'enregistrement, qui la recommandent exclusivement (DataCite 2019), au rebours des précautions formulées par la Fondation DOI (laquelle nuance le propos en évoquant une évolution des ressources et des services sur des décennies voire des centaines d'années[4]). Les normes de citation APA, qui

4. *International DOI Foundation*, *Numbering*, partie 2.6.5 : Principles. https://www.doi.org/10.1000/182.

constituent la référence principale dans le monde scientifique, suit le mouvement et recommande une seule manière de citer le DOI, à savoir, celle qui permet la localisation de la ressource.

Dans l'état, un DOI n'est pas actionnable. Afin d'obtenir un lien cliquable, l'identifiant (composé du préfixe, séparateur et suffixe) (par exemple 10.24397/pangloss-0004335) doit être attaché au nom du serveur, appelé résolveur ou proxy, qui permet de localiser sur le web la ressource identifiée par le DOI. Actuellement c'est l'adresse web de la Fondation DOI qui remplit le rôle de résolveur, doi.org, remplaçant depuis 2017 le résolveur initial, dx.doi.org. La forme finale du DOI devient ainsi, dans cet exemple, https://doi.org/10.24397/pangloss-0004335.

4. Enjeux de la création des DOI pour la Collection Pangloss

Les ressources de la Collection Pangloss bénéficient de plusieurs identifiants non payants, chacun accordé par une institution différente : *Archival Resource Key* (ARK), délivré par le Centre informatique national de l'enseignement supérieur (CINES) qui archive les données de la collection, *Handle* (hdl) délivré par le moteur de recherche Isidore de la Très Grande Infrastructure de recherche Huma-Num lors du moissonnage de la plate-forme Cocoon, et *Persistent Uniform Resource Locator* (PURL) implémenté par la plate-forme Cocoon elle-même. Dans ce contexte, le choix d'attribuer de surcroît un identifiant DOI à chaque ressource peut surprendre : n'est-ce pas un excès de zèle, dans un contexte de budgets tendus et de ressources humaines limitées ?

Si, pour l'équipe de la Collection Pangloss, le DOI est apparu comme une nécessité, c'est du fait d'une demande des utilisateurs. L'adoption d'un type d'identifiant est aujourd'hui une question de choix institutionnel et d'usage métier. La communauté des linguistes aurait également pu s'orienter vers l'identifiant ISLRN (*International Standard Language Resource Number*), conçu spécialement pour les ressources linguistiques au début des années 2010. Toutefois, l'ISLRN, non actionnable, n'a pas été adopté par la communauté scientifique en dépit du soutien que lui apporte le consortium ELRA/ELDA. Le choix en faveur du DOI paraît s'être joué entre 2015 et 2016, à mesure qu'un nombre croissant de linguistes et d'éditeurs

de contenus linguistiques recommandaient l'utilisation du DOI en raison de sa généralisation dans les pratiques de citation scientifique. Le DOI jouit d'un vif prestige auprès des chercheurs, qui en sont familiers par le biais des publications scientifiques : les maisons d'édition les plus en vue, qu'elles soient d'orientation « science ouverte » (comme OpenEdition Books, ou la maison d'édition de linguistique *Language Science Press*) ou commerciales, recourent aux identifiants DOI. Cette perception des DOI est illustrée de façon anecdotique (mais, à nos yeux, significative) par le raccourci par lequel Mark DINGEMANSE, chercheur en linguistique, présente le DOI comme « l'identifiant stable (pérenne) universellement adopté pour référencer le contenu scientifique »[5]. Ce point de vue, répandu parmi les chercheurs en Sciences humaines et sociales, constitue un argument de poids en faveur du DOI.

En outre, les services mis en œuvre par DataCite et Crossref pour tisser des liens entre données et publications constituent un avantage clair du DOI. Grâce aux services mis en place par DataCite (notamment les outils de recherche, l'exposition, les statistiques, et *Event Data* qui relie les données aux publications (GARZA 2010), les métadonnées de la Collection Pangloss sont davantage visibles et circulent plus largement depuis le déploiement des identifiants DOI. Elles sont intégrées à un entrepôt qui contient plus de 18 millions de métadonnées descriptives[6] accompagnant les données produites par des établissements du monde entier.

Le besoin des déposants et usagers de la Collection Pangloss est de localiser la ressource en un clic et le système DOI Pangloss utilise la version actionnable de l'identifiant. Il incombe dès lors aux gestionnaires de la collection de s'assurer régulièrement que les liens URL sont à jour auprès de DataCite.

5. https://twitter.com/dingemansemark/status/1150847590759620610. Consulté le 2 décembre 2019.
6. DataCite, *DataCite Statistics*, https://stats.datacite.org/. Consulté le 3 décembre 2019.

5. Implémentation

L'opération d'attribution de DOI implique, pour l'établissement demandeur d'identifiants, la transmission au service d'enregistrement, *DataCite Metadata Store* (DMS), des informations suivantes :

- l'identifiant (le DOI proprement dit)
- un fichier de métadonnées descriptives pour la ressource
- l'URL permettant d'accéder à la ressource

La ressource identifiée par un DOI est décrite par un ensemble de métadonnées. Le fichier exigé pour l'enregistrement d'un DOI, appelé *kernel metadata* ou *DOI kernel,* est encodé en XML et respecte un schéma défini par chaque agence d'enregistrement. DataCite a fait évoluer son schéma (la première datant de 2011 et la dernière - la version 4.3 - d'août 2019) apportant régulièrement des compléments et modifications au niveaux des champs descriptif à la demande des utilisateurs.

Enfin, l'URL de la ressource est de la responsabilité de l'établissement qui souhaite s'engager dans ce processus. Celui-ci doit fournir au service d'enregistrement de DataCite des URL stables, s'assurer de la préservation de l'objet à long terme et de l'accessibilité des objets. Il est évident que la garantie de localiser un identifiant sur le long terme est étroitement liée à l'effort de l'institution productrice de données à maintenir des liens stables. DataCite a pour mission d'assurer la gestion et la maintenance des identifiants de manière pérenne ; en revanche, elle n'est pas en mesure de garantir la localisation de la ressource si le fournisseur de données n'assure pas la mise à jour des liens URL.

Le cheminement technique qui a permis de porter le projet au stade de la réalisation pratique ne fera ici l'objet que d'une brève description de procédure et on se permettra de renvoyer, pour tous détails, à la documentation technique qui accompagne les scripts, disponibles en ligne (VASILE 2018).

L'attribution des DOI aux ressources de la Collection Pangloss s'est faite par la mise en place d'une chaîne de traitement développée à l'aide du langage informatique Python qui comprend trois étapes : la création du fichier de métadonnées descriptives selon le modèle DataCite, la

déclaration du DOI pour chaque ressource et enfin l'appel du web service *DataCite MDS API* pour l'envoi en masse de ces informations afin d'obtenir l'enregistrement des DOI. L'ensemble du processus d'enregistrement des DOI est schématisé par la Figure 2.

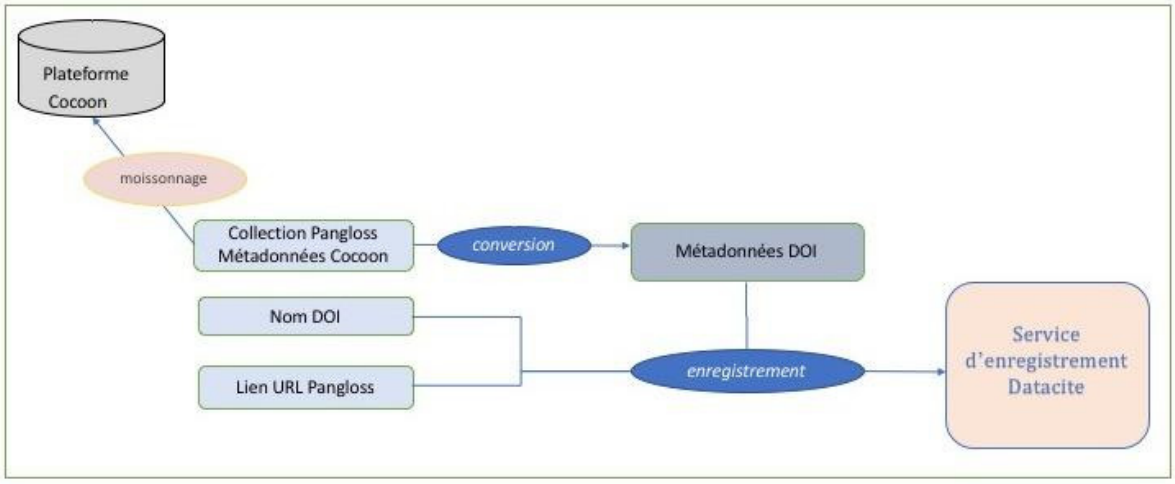

Figure 2. Chaîne de traitement dans la création des DOI pour la Collection Pangloss.

Les ressources de la collection possèdent déjà des métadonnées descriptives (encodées en XML) qui leur sont associées dans la plate-forme Cocoon. Le processus de création des fichiers de métadonnées pour l'enregistrement des DOI passe donc par une opération de conversion.

Nous avons créé un tableau de correspondance entre les éléments des deux formats (voir https://gitlab.huma-num.fr/avasile/mapping-datacite-dc) avec pour objectif de générer de fichiers de métadonnées DOI les plus complets possibles à partir des métadonnées Cocoon, sans les recréer manuellement. À la différence des métadonnées de la plate-forme Cocoon, conçues pour le domaine linguistique (selon les spécifications OLAC), le modèle DOI est fonctionnel et répond aux besoins descriptifs d'une large typologie de ressources. Par conséquent, toutes les métadonnées d'origine d'une ressource Pangloss n'ont pas une correspondance dans le modèle DOI.

A titre d'exemple, l'élément obligatoire *creator* dans le modèle DOI n'a pas son équivalent exact dans le modèle Cocoon. Ce dernier emploie l'élément *contributor* avec différents rôles (*researcher*, *performer*, *depositor*, *editor*, etc.). Ainsi, afin de générer l'élément DOI *creator* nous avons récupéré le contenu de l'élément *contributor* avec le rôle *researcher.* Toutefois, cette connexion ne fonctionne pas toujours car le modèle Cocoon ne repose pas systématiquement sur un contributeur de type *researcher* ni ne rend pas obligatoire l'élément

contributor. Nous avons été amenés alors à créer la balise *creator* à partir des autres rôles de contributeurs. Quant à l'absence totale de contributeur, l'application que nous avons développée contraint les producteurs des données à générer un créateur pour répondre aux exigences du modèle DOI.

La procédure mise en œuvre permet la création de DOI pour chaque ressource de la collection. En revanche, il paraissant important de pouvoir créer un DOI de granularité fine, pour faire référence à une phrase précise au sein d'un texte, ou un mot précis au sein d'une liste de vocabulaire. En effet, lorsqu'une source est citée dans une publication de linguistique, il est généralement fait référence à un énoncé précis (voir par exemple MICHAUD 2017, 196). Si la Collection Pangloss possédait déjà des identifiants pour citer les ressources dans leur intégralité (grâce aux identifiants PURL fournis par Cocoon), ces identifiants n'atteignaient la granularité désirée. L'objectif des chercheurs n'était pas uniquement d'identifier les ressources, mais aussi de pouvoir citer un passage précis, par le biais d'un identifiant actionnable qui conduise le lecteur directement à la phrase citée.

Les identifiants existants auraient pu répondre à ces besoins : ARK par la gestion de la granularité, et PURL par la gestion de la redirection du lien URL. À cette fin, nous avons réfléchi dans un premier temps à une solution pour attribuer un DOI à chacune de ces sous-parties. Mais une difficulté de taille est apparue : ces sous-parties (les phrases d'un texte, les mots d'une liste de vocabulaire) ne constituent pas des entités autonomes. De ce fait, elles sont dépourvues de métadonnées spécifiques. Au plan technique, des tests nous ont prouvé que les DOI pouvait être créés pour des phrases ou mots au sein d'un texte, mais le choix d'attribuer un DOI à chaque élément de contenu des fichiers numériques aurait été contestable : le DOI décrit et identifie un document, pas chaque élément de contenu composant un document. Le choix final pour la mise en production a été de mettre en œuvre un système de renvoi par ancre sur la page web, à partir du DOI de la ressource parente[7].

Voici à titre d'exemple, un DOI cité dans un article de linguistique (MICHAUD et al. 2019) qui évoque la phrase numéro 13 d'une ressource textuelle, intitulée *Buried alive (version 2)*:

7. Technique mise en œuvre dans la version de production du code : https://github.com/CNRS/DoiPangloss/releases/tag/V1.1.

https://doi.org/10.24397/pangloss-0004537#S13 Le lien est construit à partir du DOI de la ressource (10.24397/pangloss-0004537) attaché au résolveur DOI (https://doi.org), auquel a été rajouté l'ancre identifiant le numéro de la phrase (#S13 ; « S » désigne ici « Sentence », le premier niveau de granularité des textes).

On relèvera que cette solution comporte un risque de brouiller aux yeux des utilisateurs l'importante distinction entre localisateurs *(locators)* et identifiants *(identifiers)*. Dans quelle mesure la syntaxe de l'identifiant fourni pour les phrases d'un texte sera-t-elle intelligible aux utilisateurs ? Il faut une certaine qualité d'attention pour déceler, dans la chaîne https://doi.org/10.24397/pangloss-0004537#S13, les trois éléments qui viennent d'être décrits : l'adresse du résolveur, le DOI proprement dit, et la précision concernant une phrase en particulier (S13). Nous plaçons ici nos espoirs dans des progrès des connaissances en informatique, qui sont nécessaires à l'échelle de la société (HENRY et al. 2018) et particulièrement importants pour les chercheurs dans le domaine des humanités et des sciences sociales, dont les sciences du langage (FILLMORE 1992 ; COLLINS et al. 2015, 10).

Il est également permis d'espérer que la présentation des DOI sur l'interface en ligne de la Collection Pangloss facilite un apprentissage de la syntaxe des DOI proposés, comme l'illustre la Figure 3. Les icônes DOI sont cliquables et ouvrent une fenêtre contenant le DOI : celle à droite du titre donne le DOI du document ; celle à droite de chaque numéro de phrase comporte une ancre identique à ce numéro de phrase (ici, la première).

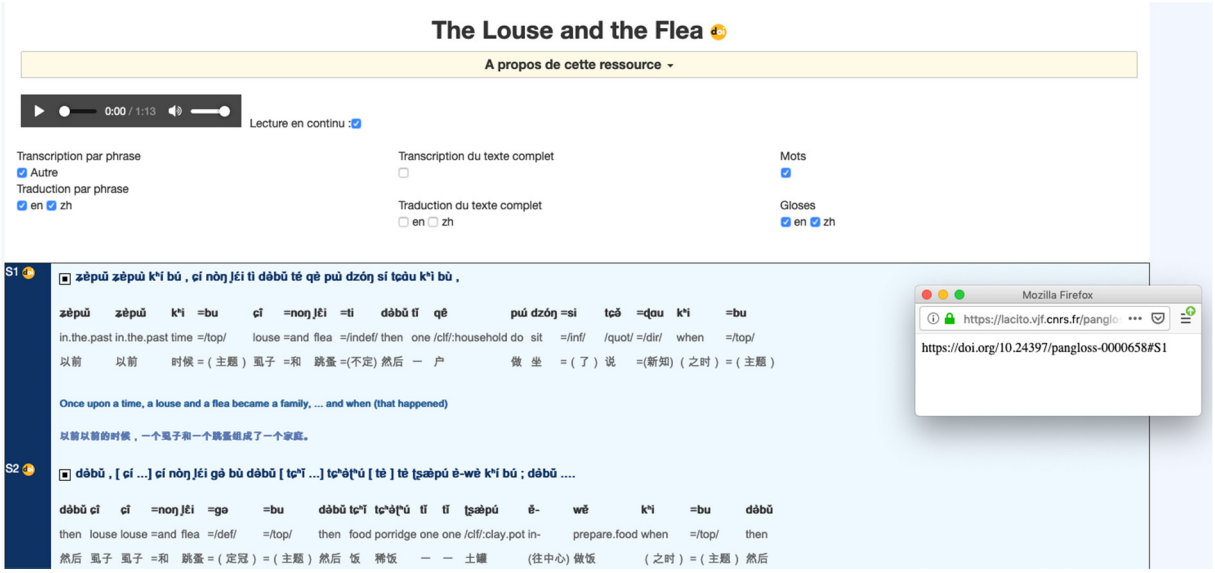

Figure 3. Affichage d'un texte sur l'interface en ligne de la Collection Pangloss

Plus encore, répondre au besoin d'associer les phrases avec un DOI actionnable a eu des conséquences sur le respect des bonnes pratiques d'utilisation du DOI, car ce choix vient à l'encontre de la recommandation de DataCite concernant la page de résolution d'un DOI (la *landing page*)[8]. Il aurait été plus logique de se servir de la page de métadonnées de la plateforme Cocoon comme *landing page*. Or l'équipe Pangloss a privilégié le besoin utilisateur pour qui le DOI actionnable était la véritable plus-value.

Dans un environnement numérique, l'utilisation du DOI permet aussi d'établir un lien entre l'article scientifique et la donnée citée *via* les métadonnées de l'article. Ainsi, le portail d'archives ouvertes HAL sur lequel est déposé l'article cité ci-dessus affiche sur la page de métadonnées complètes, à la rubrique « Données associées », le numéro DOI de la donnée citée : voir un exemple en Figure 4. Cette information facilite la reconnaissance du lien publication-donnée par des algorithmes, et, partant, l'avènement d'environnements dans lesquels soit atteinte une meilleure intégration des publications, données et outils.

8. DataCite, *Best Practices for DOI Landing Pages*, https://support.datacite.org/docs/landing-pages. Consulté le 15 avril 2020.

Date de publication	2019
Domaine	• **Sciences de l'Homme et Société/Linguistique**
Mots-clés	en speech recognition, machine learning, error analysis, interdisciplinarity, Computational Language Documentation
Langue du document	anglais
Données associées	• 10.24397/pangloss-0004537#S13
Date de production/ écriture	2019
Licence	Paternité - Pas d'utilisation commerciale - Partage selon les Conditions Initiales

Figure 4. Extrait de la page du dépôt halshs-02059313 (MICHAUD et al. 2019), contenant un lien (actionnable) via identifiant DOI vers des données hébergées dans la Collection Pangloss

6. Premier bilan et perspectives

L'application que nous avons développée a permis, à l'été 2020, la création de DOI pour les 5794 ressources de la Collection Pangloss. Les nouveaux dépôts bénéficient eux aussi d'un DOI pour chaque document. S'il est encore tôt pour dresser un bilan, on soulignera quelques perspectives ouvertes par ce déploiement.

Le premier enjeu était de satisfaire la demande de chercheurs convaincus que des identifiants DOI constituaient une impérieuse nécessité pour attirer déposants et utilisateurs. Par souci de placer l'usager au centre de la réflexion (CHAUDIRON et IHADJADENE 2002 ; WASSON, HOLTON, et ROTH 2016), il paraissait indispensable de répondre à cette attente. La nouvelle s'est répandue via les carnets de recherche[9] et les réseaux sociaux, et ce premier objectif est atteint.

Le déploiement d'identifiants DOI a en outre pour effet de mettre en lumière la nécessité de nouveaux développements logiciels pour la réalisation pratique d'une association plus étroite entre publications, données et outils. Ainsi, Zotero, outil de gestion bibliographique libre et gratuit, permet d'ajouter un document par son identifiant DOI, mais cela ne fonctionne pas à l'heure actuelle pour les DOI de DataCite, dont ceux de la Collection Pangloss. Il importe donc de participer à un effort concerté afin que cette passerelle soit ouverte.

9. Voir notamment : Guillaume JACQUES, « La Collection Pangloss : des DOI pour toutes les ressources », Carnet de recherche *Panchronica*, https://panchr.hypotheses.org/2931. Consulté le 12 février 2019.

Au plan institutionnel, le travail de déploiement d'identifiants DOI suggère le projet d'étendre l'application du script générateur de DOI à d'autres collections hébergées par la plate-forme Cocoon (nous l'avons testé pour la collection du laboratoire LLACAN : Langage, langues et cultures d'Afrique). Si la plate-forme Cocoon possédait un système d'attribution de DOI pour les documents de toutes les collections hébergées, cela constituerait un attrait supplémentaire pour les déposants et utilisateurs.

Conclusion

Le projet d'attribution des noms DOI aux ressources linguistiques de la Collection Pangloss a fourni l'occasion d'une réflexion sur le rôle des identifiants numériques dans l'identification et la citation des données collectées lors des missions de terrain. Ce travail confirme que le choix d'un identifiant plutôt qu'un autre est davantage une question de pratique institutionnelle et de métier qu'un choix établi sur des critères techniques. Il nous a paru évident que si les chercheurs connaissaient l'existence et l'utilité des principaux identifiants numériques (PURL, ARK, DOI, Handle), le dévolu jeté sur le DOI s'expliquait par l'influence du modèle de l'édition scientifique. Les maisons d'édition, à l'origine du système d'identification DOI, apportent visibilité et reconnaissance au chercheur qui publie dans les revues les plus cotées. Par association, les identifiants DOI se trouvent parés d'un grand prestige aux yeux des chercheurs. Aujourd'hui, les nouveaux services mis en place par les agences d'attribution des DOI renforcent le statut de cet identifiant notamment par la mise en relation entre les publications et les données, ce qui en fait – non sans quelque paradoxe – un élément important du paysage mouvant de la science ouverte.

Références citées

BAUDRY, J., et MACHEFERT, S. « Retour d'expérience DOI - Université Bordeaux-Montaigne », *Bibliopedia*. http://bibliopedia.fr/w/index.php?title=Retour_d%27exp%C3%A9rience_DOI_-_Universit%C3%A9_Bordeaux_Montaigne. Consulté le 2 décembre 2019.

BERMES, E. 2006. « Des identifiants pérennes pour les ressources numériques : l'expérience de la BnF ». Bibliothèque nationale de France. http://bibnum.bnf.fr/identifiants/identifiants-200605.pdf.

BERMES, E., et POUPEAU G. 2012. « Les technologies du web appliquées aux données structurées ». Dans *Séminaire IST Inria : le document numérique à l'heure du web de données*. Sous la direction de Lisette Calderan, Pascale Laurent, Hélène Lowinger et Jacques Millet, 41-84. Sciences et techniques de l'information. Carnac, France : ADBS. https://hal.inria.fr/hal-00843775.

CHAUDIRON, S., et IHADJADENE M. 2002. « Quelle place pour l'usager dans l'évaluation des Systèmes de recherche d'information (SRI) ? » Dans *Recherches récentes en sciences de l'information : Convergences et dynamiques, actes du colloque MICS-LERASS*., 211-31. Paris : ADBS Éditions.

CHOUKRI, K., PARK J., HAMON O., et ARRANZ V. 2011. "Proposal for the international standard language resource number". Dans *Proceedings of the Workshop on Language Resources, Technology and Services in the Sharing Paradigm*., 75-83.

COLLINS, S., HARROWER N., HAUG D., IMMENHAUSER B., LAUER G., ORLANDI T., ROMARY L., et Wandl-Vogt E. 2015. "Going digital: creating change in the Humanities". ALLEA.

CROSSREF, "Crossref Display Guidelines (March 2017)". https://doi.org/10.13003/5jchdy. Consulté le 29 avril 2020.

DataCite, *DOI Display Guidelines*, 2019. https://support.datacite.org/docs/datacite-doi-display-guidelines. Consulté le 15 avril 2020.

« Digital Object Identifier (DOI) », *La maison des revues*, Des outils pour Journals, Référencement et interopérabilité, 16 décembre 2011. http://www.maisondesrevues.org/253. Consulté le 29 avril 2020.

FILLMORE, C. J. 1992. « Corpus linguistics or computer-aided armchair linguistics ». Dans *Directions in corpus linguistics. Proceedings of Nobel Symposium*, 82 : 35-60.

GARZA, K. "Datacite Citation Display: Unlocking Data Citations", *DataCite Blog*, 7 janvier 2020. https://www.doi.org/10.5438/1843-k679.

GREEN, B., et BIDE M. 1997. "Unique Identifiers: a brief introduction". *Book Industry Communication*. https://bic.org.uk/files/pdfs/uniquid.pdf.

HENRY, J., HERNALESTEEN A., DUMAS B., et COLLARD A-S. 2018. « Que signifie éduquer au numérique ? Pour une approche interdisciplinaire ».

HILSeE, H-W, et KOTHE J. 2006. *Implementing persistent identifiers*. Göttingen : Consortium of European Research Libraries (CERL).

HOCQUET, A. 2019. « Reprenez ce texte sur la science ouverte et transformez-le ». Dans *S.I.Lex, le blog revisité : parcours de lectures dans le carnet d'un juriste et bibliothécaire*. Sous la direction de Sarah Clément et Mélanie Leroy-Terquem, 684. Villeurbanne : Presses de l'Enssib.

INRA. IST-Données de la Recherche, « Datapartage - Findable ». https://www6.inra.fr/datapartage/Produire-des-donnees-FAIR/Findable].

International DOI Foundation, "Numbering". 19 décembre 2019. https://www.doi.org/10.1000/182.

KUNZE, J. 2003. "Towards electronic persistence using ARK identifiers". *California Digital Library*.

MASINTER, L., BERNERS-LEE T., et FIELDING R. T. 2005. "Uniform resource identifier (URI): Generic syntax". *Network Working Group*.

MAUREL, L. 2019. « L'ouverture des données de recherche : un retour aux sources de l'Ethos de la Science ? » *S.I. Lex*. https://scinfolex.com/2019/06/05/louverture-des-donnees-de-recherche-un-retour-aux-sources-de-lethos-de-la-science/.

MICHAILOVSKY, B., MAZAUDON M., MICHAUD A., . GUILLAUME S., FRANCOIS A., et ADAMOU E. 2014. "Documenting and researching endangered languages: the Pangloss Collection". *Language Documentation and Conservation* 8 : 119-35.

MICHAUD, A. 2017. *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, A., ADAMS O., COX C., et GUILLAUME S. 2019. "Phonetic lessons from automatic phonemic transcription: preliminary reflections on Na (Sino-Tibetan) and Tsuut'ina (Dene) data". Dans *Proceedings of ICPhS XIX (19th International Congress of Phonetic Sciences)*. Melbourne. https://halshs.archives-ouvertes.fr/halshs-02059313.

Norman Paskin, « DOI: Current Status and Outlook », OCDE (2007), *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*, Éditions OCDE, Paris. https://doi.org/10.1787/9789264034020-en-fr.

PASKIN, N. 2015. « The digital object identifier: From ad hoc to national to international ». Dans *The critical component: standards in the information exchange environment*. Sous la direction de Todd A. Carpenter. Atlanta : American Library Association Publishing.

PASKIN N., « DOI: Current Status and Outlook », *D-Lib Magazine*, nr.5, mai 1999 https://www.doi.org/10.1045/may99-paskin.

RENTIER, B. 2018. *Science ouverte, le défi de la transparence*. L'Académie en poche 114. Bruxelles : Académie royale de Belgique.

SUBER, P. 2016. *Qu'est-ce que l'accès ouvert ?* Marseille : OpenEdition Press.

VASILE, A. 2018. *Spécifications techniques de l'application DoiPangloss*. https://github.com/vasaura/DoiPangloss/releases/tag/v1.0.

WASSON, C., HOLTON G., et ROTH H. S. 2016. "Bringing user-centered design to the field of language archives". *Language Documentation and Conservation* 10 : 641-81.