L'identifiant pérenne, une clé pour les bases de données linguistiques dans une perspective de science ouverte

Laurent Kevers

L’identifiant pérenne, une clé pour les bases de données linguistiques dans une perspective de science ouverte

1. Résumé

L’approche proposée par la science ouverte prône une recherche plus transparente, plus solidement étayée, plus reproductible, et plus efficacement cumulative (MESRI, 2021). Ces principes sont présents dans la communauté scientifique depuis de nombreuses années et leur pertinence a déjà été soulignée, entre autres dans le contexte des technologies de la langue, et en particulier dans le cas des langues dites ‘peu dotées’ (Soria *et al.*, 2013). Leur mise en avant s'est cependant intensifiée avec, en France, les récents plans pour une science ouverte du Ministère de l’Enseignement Supérieur, de la Recherche et de l’Innovation (2018-2021 et 2021-2024). Dans ce contexte, les données de la recherche se doivent, autant que possible, de respecter les principes ‘FAIR’, c'est-à-dire Faciles à trouver ; Accessibles ; Interopérables ; Réutilisables (Wilkinson *et al.*, 2016). En linguistique plus spécifiquement, les ‘*Austin Principles of Data Citation in Linguistics*’ (Berez-Kroeker *et al.*, 2018) reprennent un ensemble similaire de bonnes pratiques vers lesquelles tendre.

Les recherches en linguistique amènent les chercheurs à récolter de nombreuses données. C’est bien entendu le cas en linguistique géographique et en dialectologie, et en particulier pour la Banque de Données Langue Corse (BDLC), que nous prenons ici comme contexte d’application. Créée en 1986, la BDLC s’adosse au Nouvel Atlas Linguistique et ethnographique de la Corse (NALC), dont le développement remonte lui aux années 1970 (Dalbera-Stefanaggi et Retali-Medori, 2015). Le programme a pour vocation de recueillir, stocker, analyser et restituer des données dialectales relatives au savoir-faire et aux traditions culturelles corses, par le biais d’enquêtes de terrain en Corse et dans le nord de la Sardaigne. Les enquêtes s’appuient sur des questionnaires thématiques constitués de listes de mots en français. Les traductions en corse sont collectées et un entretien semi-dirigé, entièrement en corse, peut s’engager. Celui-ci permet de recueillir des ethnotextes (témoignages) relatifs aux pratiques et traditions. Ces données sont ensuite analysées et intégrées sous la forme d’une base de données relationnelle.

Si une base de données est adaptée à une consultation au moyen d'une interface informatique telle qu’une application *web*, voire à la mise à disposition des données au travers d’une API (‘*Application Programming Interface*’), elle n’est souvent pas totalement satisfaisante au regard des principes de la science ouverte. En effet, l’identification d’un élément précis de la base de données, et la référence à celui-ci *via* un mécanisme de citation, ne sont en général pas choses aisées. D’autre part, une base de données est un objet dynamique par nature. En plus du caractère non stable des URL, la pérennité d’une donnée n’est pas nécessairement garantie à long terme, car des éléments de la base peuvent être

supprimés, que ce soit à dessein ou par erreur. Enfin, si les données sont généralement accessibles par l'intermédiaire d'une interface de consultation dédiée – bien qu'on puisse imaginer qu'une base de données en soit dépourvue – celle-ci n'est pas toujours conçue de manière à faciliter le référencement par les moteurs de recherche, qu'ils soient généralistes ou spécialisés. Dès lors, on peut considérer que, dans ce contexte, les principes 'FAIR' ne sont pas rencontrés de manière optimale.

Afin d'améliorer l'identification, la citation et le référencement, nous préconisons l'utilisation d'identifiants pérennes. Ceux-ci existent depuis de nombreuses années et constituent un élément important de l'infrastructure mise en place pour la science ouverte. Il s'agit de chaînes alphanumériques uniques qui permettent l'identification d'objets réels ou conceptuels et ce de manière indépendante de leur localisation. Les DOI[1] ('*Digital Object Identifier*'), en constituent une déclinaison. En pratique, un DOI peut être traduit en une URL qui permet d'accéder soit à une page proposant des métadonnées, soit directement à la donnée. Une fois émis, les DOI ne peuvent plus être supprimés ; les adresses qu'ils référencent sont en revanche modifiables. L'utilisation des DOI est déjà effective dans de nombreuses plateformes d'éditeurs, d'archives ouvertes ou d'entrepôts de données. Par exemple, les enregistrements placés sur CoCoON disposent automatiquement d'un DOI tel que <https://doi.org/10.34847/cocoon.5f2638cd-850b-3ccd-bb8b-20b3cb9d3fc9>.

Les DOI peuvent être utilisés pour identifier n'importe quel élément d'une base de données. La granularité des objets à cibler doit donc être soigneusement choisie en fonction de la nature des données et des besoins. L'objectif poursuivi est de pouvoir identifier la donnée en tant que tel, et par conséquent de rendre sa citation possible. Concrètement, en se plaçant dans le cadre de la BDLC, cette approche permet par exemple d'isoler et d'identifier chaque élément lexical individuellement. Ainsi, la 'question' relative à la « coccinelle » peut recevoir son propre DOI. Il est évidemment possible d'opter pour une granularité plus fine, et choisir de poser des DOI sur toutes les traductions associées aux différentes localités. De même, il est également envisageable d'identifier les objets annexes de la fiche lexicale tels que cartes, photos ou sons. Le choix de la fiche lexicale comme unité à cibler semble cependant être un bon compromis entre l'objectif d'identification et la complexité qui en découle, du moins dans un premier temps. Cette option dispose de plus de l'avantage de correspondre au contenu informationnel d'une carte d'atlas. À côté des fiches lexicales, il est aussi intéressant de pouvoir identifier les ethnotextes qui constituent un autre type d'information important dans la base. Si, à notre connaissance, cette approche n'est implémentée que dans quelques cas, par exemple dans le cadre de VerbaAlpina (Colcuc et Mutter, 2020), elle est *a priori* généralisable à toute base de données linguistique.

Les bénéfices sont multiples. Tout d'abord en ce qui concerne les possibilités de citation, celles-ci ne sont plus limitées à la base de données en tant qu'objet monolithique, mais

[1] Dont l'émission et la gestion est confiée en France à l'Inist-CNRS.

peuvent être réalisés à une granularité bien plus fine. Par rapport aux publications au format papier, par exemple sous la forme d'atlas linguistique pour la BDLC, les données peuvent être identifiées et citées dès leur apparition dans la base, sans attendre l'édition d'un nouveau volume. De par le caractère permanent des DOI, la pérennité de l'identification des données est assurée. Une citation effectuée à partir de ce mécanisme reste valide et identifiable, même si les données sont supprimées – intentionnellement ou non – de la base de données. L'éventuelle modification de l'URL qui permet d'accéder à la donnée n'a pas d'impact sur sa citation puisque celle-ci repose sur le DOI, qui constitue un mécanisme d'adaptation face à ce genre de changement. Enfin, l'utilisation d'un formalisme standard facilite l'interopérabilité, ainsi que les possibilités de partage et de réutilisation des données ou métadonnées. Les données sont aussi plus facilement accessibles et indexées par les moteurs de recherche, en particulier les moteurs spécialisés tels que par exemple DataCite Search. Indirectement, la fourniture d'un moyen de citation uniformisé et standardisé du contenu de la base de données est de nature à augmenter sa visibilité et sa notoriété.

Cette communication a pour objet de contribuer à un meilleur respect des principes 'FAIR' pour les ressources linguistiques, en particulier les bases de données. L'approche proposée permet une amélioration de l'identification, du référencement et de la citation des données qui y sont contenues. À notre connaissance, l'utilisation des DOI à des fins d'identification et de citation de contenu ciblé d'une base de données linguistique est peu répandue. Cette pratique est cependant souhaitable et nous désirons lui apporter plus de visibilité.

2. Références bibliographiques

Berez-Kroeker, A. L. / Andreassen, H. N. / Gawne, L. / Holton, G. / Kung, S. S. / Pulsifer, P. / Collister, L. B. / the Data Citation and Attribution in Linguistics Group / the Linguistics Data Interest Group, 2018. *The Austin Principles of Data Citation in Linguistics* (Version 1.0). <https://site.uit.no/linguisticsdatacitation/austinprinciples>

Colcuc, B. / Mutter, C., 2020. « Interopérabilité des données géolinguistiques à l'exemple du projet VerbaAlpina ». *Bien Dire et Bien Aprandre - Revue de Médiévistique* 35,131-146.

Dalbera-Stefanaggi, M.-J. / Retali-Medori, S., 2015. « Trente ans de dialectologie corse : Autour du programme Nouvel Atlas Linguistique et Ethnographique de la Corse et Banque de Données Langue Corse ». *Corse d'hier et de demain - Nouvelle série* 6, 17-25.

DataCite Search, <https://search.datacite.org>

MESRI = Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, 2021. *Deuxième Plan national pour la science ouverte. Généraliser la science ouverte en France 2021-2024*. <https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte>

Soria, C. / Mariani, J. / Zoli, C., 2013. « Dwarfs sitting on the giants' shoulders – how LTs for regional and minority languages can benefit from piggybacking major languages ». in : *Proceedings of the 17th Foundation for Endangered Languages Conference*, 73–79.

Wilkinson, M. D. / Dumontier, M. / Aalbersberg, I. J. J. / Appleton, G. / Axton, M. / Baak, A. / Blomberg, N. / Boiten, J.-W. / da Silva Santos, L. B. / Bourne, P. E. / Bouwman, J. / Brookes, A. J. / Clark, T. / Crosas, M. / Dillo, I. / Dumon, O. / Edmunds, S. / Evelo, C. T. / Finkers, R. / *et al.*, 2016. « The FAIR Guiding Principles for scientific data management and stewardship ». *Scientific Data* 3, <https://doi.org/10.1038/sdata.2016.18>.

IdP = *DORANum : Identifiants pérennes. Comment associer durablement des données à son auteur ?* <https://doranum.fr/identifiants-perennes-pid/>