Le *Dictionnaire topographique. Une API pour les toponymes anciens français*

Olivier Canteaut[1], Vincent Jolivet[2], Julien Pilla[3]

1. École nationale des chartes
65 rue de Richelieu, 75002 Paris, France
olivier.canteaut@chartes.psl.eu

2. École nationale des chartes
65 rue de Richelieu, 75002 Paris, France
vincent.jolivet@chartes.psl.eu

3. École nationale des chartes
65 rue de Richelieu, 75002 Paris, France
julien.pilla@chartes.psl.eu

RÉSUMÉ. Le Dictionnaire topographique *est une ressource de premier plan pour les historiens et les toponymistes : il compte près de 400 000 entrées et hiérarchise plus de 980 000 toponymes anciens attestés, datés et référencés. Depuis 2009, le CTHS numérise les différents volumes, pour en proposer une édition numérique. Une nouvelle application est en cours de développement. Adossée à une API documentée, elle offre un accès normalisé aux données, et tire parti du liage des données au référentiel INSEE pour localiser les toponymes. L'objectif de cette API est de favoriser les remplois de cette ressource importante, mais aussi d'en poursuivre l'enrichissement en offrant aux chercheurs une interface pour corriger et compléter le contenu au gré de leurs découvertes. Cette communication vise à faire connaître cette ressource essentielle pour l'étude toponymique : nous présenterons l'histoire de cette entreprise éditoriale hors norme, détaillerons les étapes de la numérisation, de la restructuration et de l'enrichissement des données. Nous présenterons enfin l'API et l'application associée qui rend possible l'exploitation de nouvelles relations au sein du* Dictionnaire*, et qui surtout permettra de revitaliser une entreprise éditoriale inachevée.*

ABSTRACT. The Dictionnaire topographique *is a leading resource for historians and toponymists: it has nearly 400,000 entries and ranks more than 980,000 ancient toponyms that have been documented, dated and referenced. Since 2009, the CTHS has been digitising the various volumes, in order to offer a digital edition. A new application is being developed. Its documented API provides standardized access to data, and uses data binding to the INSEE repository to locate place names. The objective of this API is to promote the re-use of this*

important resource, but also to continue to enrich it by providing researchers with an interface to correct and complete the content as they discover it. This paper aims to promote this essential resource for toponymic research: we will present the history of this extraordinary publishing initiative, detailing the steps involved in digitization, restructuring and data enrichment. Finally, we will present the API and the associated application that makes it possible to exploit new relationships within the Dictionnaire*, and above all, to revitalize an unfinished editorial initiative.*

Mots-clés : données géohistoriques, toponymes, France, API, application Web, Web de données, gazetier

Keywords: geohistorical data, toponyms, history, France, API, Web application, Linked Open Data, gazetteer

1. Introduction

Entreprise éditoriale au long cours lancée par le Comité des travaux historiques – aujourd'hui Comité des travaux historiques et scientifiques (CTHS) –, le *Dictionnaire topographique* a eu pour mission de compiler tous les toponymes anciens et modernes de la France. Au total, 35 tomes (pour 35 départements) ont été publiés de 1861 à 2008. Même si la couverture (plus du tiers du territoire métropolitain) n'est pas à la hauteur de l'ambition nationale initiale, le *Dictionnaire topographique* est une ressource de premier plan pour les historiens et les toponymistes : il compte près de 400 000 entrées et hiérarchise plus de 980 000 toponymes anciens attestés, datés et référencés. Depuis 2009, le CTHS numérise les différents volumes, pour en proposer une édition numérique, enrichie progressivement au fil des numérisations. Le corpus est complet depuis 2018, et à cette occasion, une nouvelle application est en cours de développement. Celle-ci n'est plus une simple édition numérique ; adossée à une API documentée, elle offre un accès normalisé aux données, et tire parti du liage des données au référentiel de l'INSEE pour localiser les toponymes. L'objectif de cette API est de favoriser les remplois de cette ressource importante, mais aussi d'en poursuivre l'enrichissement en offrant aux chercheurs une interface pour corriger et compléter le contenu au gré de leurs découvertes. Cette communication vise à faire connaître cette ressource essentielle pour l'étude toponymique.

2. Le *Dictionnaire topographique de la France*

2.1. Une entreprise éditoriale

Lancé en 1859 sur proposition de l'historien Léopold Delisle, le *Dictionnaire topographique de la France* avait pour ambition de doter les savants d'un dictionnaire géographique « de la France ancienne et moderne » utile à l'étude de l'histoire et de la géographie des provinces françaises. Il s'agissait de recenser les noms de lieux fournis par la géographie physique, les noms des lieux habités et ceux qui se rapportent à la « géographie historique » (anciennes circonscriptions, fiefs, abbayes, vieux chemins, etc.), en indiquant pour chacun de ces lieux sa nature (ferme, hameau, moulin, etc.), sa localisation (commune d'appartenance), diverses données historiques (ressort judiciaire, ecclésiastique) et surtout les différentes graphies de son nom au cours des siècles, dûment datées et référencées. En raison de l'ampleur de la tâche, le Comité opta rapidement pour le principe d'un volume par département, le tout devant à terme – et en théorie – être unifié par un index général.

Les débuts furent prometteurs : dix-neuf dictionnaires parurent entre 1861 et 1884, dus principalement aux archivistes départementaux, parfois à d'autres érudits, correspondants locaux du Comité. Le mouvement se poursuivit à un rythme plus modéré, à raison de deux à quatre dictionnaires par décennie jusqu'aux années 1920, les derniers volumes publiés étant ceux de la Sarthe et de la Seine-et-Marne (années

1950), de la Seine-Maritime (années 1980) et, dernier en date, celui de la Saône-et-Loire (2008). Ce sont aujourd'hui trente-cinq départements qui sont couverts, représentant plus du tiers du territoire national métropolitain.

2.2. Une base de toponymes anciens

L'appellation *Dictionnaire topographique* reflète imparfaitement le contenu des dictionnaires édités. Ils se conforment tous au modèle éditorial prescrit par L. Delisle, composé de trois parties : 1. une introduction consacrée à la géographie historique, et comprenant une description physique du département (justifiant le qualificatif « topographique ») ; 2. le corps même du dictionnaire, composé des notices historiques des toponymes (nature et localisation administrative du lieu, liste des formes anciennes référencées) ; 3. l'index général des formes anciennes.

Le cœur du dictionnaire, la donnée numérisée et (re)structurée pour construire l'application, consiste donc en une liste considérable et ordonnée de toponymes historiques attestés : cette particularité se comprend aisément dans la mesure où les principaux contributeurs ont été des archivistes et des historiens qui en travaillant à ce long recensement des formes toponymiques anciennes se conformaient à la prescription initiale de L. Delisle les enjoignant de « n'écarter aucun nom qui ait un caractère d'ancienneté et qui présente un intérêt historique ou philologique ». L'ensemble compile plus de 980 000 toponymes, typés, datés et référencés.

3. Une donnée structurée pour la recherche

C'est cette collection de toponymes qui a fait l'objet d'une campagne de numérisation depuis 2009, conduite par le CTHS, en partenariat notamment avec avec l'École des chartes, le centre d'onomastique des Archives nationales et l'UMR ARTEHIS (Dijon). L'objectif initial était de proposer une réédition numérique des dictionnaires imprimés « offrant des possibilités d'interrogation et d'exploitation nouvelles : interrogation conjointe de plusieurs dictionnaires, interrogations croisées, recherche par type de lieu... ». Ce travail de numérisation (OCR, restructuration XML des notices selon un schéma simple) a été rendu possible par la (relative) homogénéité éditoriale des différents volumes, qui ont tous respecté assez fidèlement les prescriptions éditoriales initiales. Dix unités de sens furent identifiées : article, vedette, définition, localisation, typologie, forme ancienne, commentaire, date, référence et renvoi. Les enrichissements typographiques (italique, majuscules, exposants) et la structure éditoriale (paragraphes et les numéros de pages) ont également été balisés. C'est sous cette forme que les fichiers sont distribués[1]. Une première application de consultation a été développée et est accessible sur le site du CTHS[2].

En 2018, le projet a été redéfini de manière à tirer parti du liage des données avec d'autres référentiels pour enrichir la base et rendre possible la géolocalisation

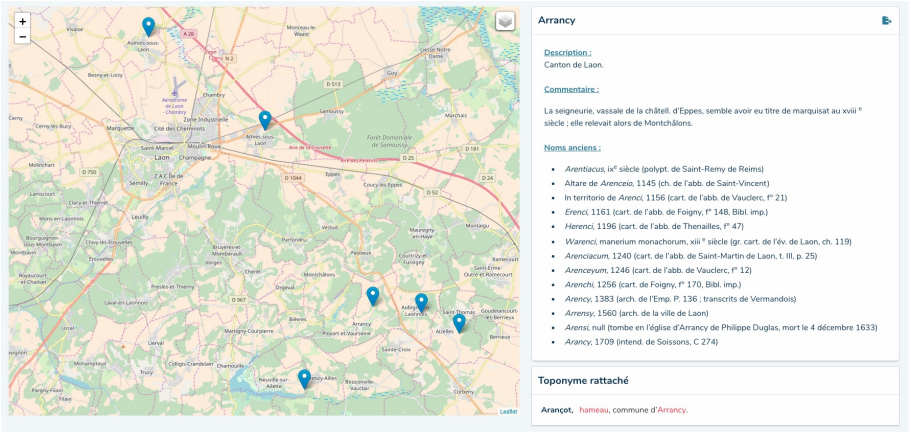
[1] https://github.com/chartes/dico-topo

[2] http://cths.fr/dico-topo/

des toponymes historiques. Une première campagne d'enrichissement a été menée avec l'aide de C. Burette, étudiante du master TNAH de l'École des chartes : 15 852 entrées du *Dictionnaire* correspondant à des communes (toponymes de type *commune*) et 283 877 entrées rattachées (des toponymes localisés dans une commune) ont été liées au référentiel de l'INSEE et ont pu être géolocalisées. Autrement dit, 75 % des noms de lieux contenus dans la base sont localisés au niveau de leur commune de rattachement. Pour les 25 % restants, un important effort d'annotation est à prévoir pour affiner leur localisation dont la granularité, du fait du découpage éditorial des dictionnaires (un tome par département), est a minima départementale.

4. Une API et une application Web

L'initiative lancée en 2009 de numérisation et de partage des données s'inscrivait dans le mouvement de l'*Open Data* : la volonté était de partager les ressources avec la communauté scientifique. La prise de conscience récente que l'enrichissement des données est conditionné à leur liage à d'autres jeux de données (INSEE, IGN, etc.) a redéfini considérablement le projet d'édition numérique initial. Il ne s'agit pas seulement de donner à lire, mais de donner accès aux données selon une méthode normalisée et documentée. L'API définie est conforme à la spécification JSON API et permet d'accéder aux données relatives à chaque toponyme (localisation, formes anciennes, coordonnées géographiques, code INSEE, etc.). L'idée n'est pas seulement de partager les données, mais de permettre leur exploitation par des applications tierces – construire par exemple un service d'identification des toponymes anciens.



Une application de consultation[3] adossée à cette API tire parti du liage des données en offrant du rebond vers d'autres référentiels et en offrant, grâce au service géoportail de l'IGN, une carte des toponymes localisés. Sur le modèle de Pleiades, l'application permet à un utilisateur authentifié de corriger une entrée ou de saisir de nouvelles formes anciennes attestées des toponymes. Notre ambition à travers ces fonctionnalités est de relancer une initiative éditoriale centenaire et malheureusement interrompue depuis 2008 et d'envisager – à terme – une couverture plus exhaustive du territoire métropolitain.

5. Conclusion

La communication présentera donc un projet éditorial centenaire, une initiative numérique d'une décennie et un projet de *Linked Open Data* récent. La première version de l'application présentée est une véritable preuve de concept. Le travail à entreprendre reste considérable : améliorer le liage des données avec d'autres référentiels (notamment GeoNames, Pleiades, WHG) et surtout fédérer la communauté des chercheurs, historiens et archivistes qui pourra contribuer à l'enrichissement de cette base précieuse pour étudier la toponymie ancienne.

[3] https://github.com/chartes/dico-topo-app