Allegro : une plateforme « couteau suisse » pour l'exploitation des ressources textuelles

Étienne Petitjean[1] Christophe Benzitoun[2] Benjamin Husson[1] Sandrine Ollinger[1]
(1) ATILF, Université de Lorraine-CNRS, 54000 Nancy, France
(2) ATILF, CNRS-Université de Lorraine, 54000 Nancy, France
Etienne.Petitjean@atilf.fr, Christophe.Benzitoun@univ-lorraine.fr,
Benjamin.Husson@atilf.fr, Sandrine.Ollinger@atilf.fr

RÉSUMÉ
Nous nous proposons de présenter Allegro, la nouvelle plateforme pour l'exploitation de ressources textuelles développée au sein du laboratoire ATILF, à travers un inventaire rapide de ses applications actuelles et à venir, ainsi que d'une introduction à ses bases techniques. Allegro offre de nombreuses possibilités pour l'indexation et l'interrogation de données structurées, annotées et enrichies de métadonnées.

ABSTRACT
Allegro : A "Swiss knife" platform for exploitation of textual resources

We present Allegro, the new platform for the exploitation of textual resources developed at ATILF. We offer here a quick inventory of its current and future applications, before introducing its technical foundations. Allegro offers many possibilities for indexing and querying structured, annotated and metadata-enriched data.

MOTS-CLÉS : Serveur de données, Ressources linguistiques, Indexation, Interrogation.

KEYWORDS: Data server, Linguistic resources, Indexation, Querying.

1 Introduction

Depuis sa création, le laboratoire ATILF développe, exploite et met à disposition de la communauté scientifique comme du grand public de nombreuses ressources linguistiques (Bernard et al., 2002). Le moteur de recherche Stella (Dendien, 1991), ayant permis d'exploiter ces ressources jusqu'à récemment, a fait son temps. Conçu à la fin des années 80 avec les contraintes et les limitations de l'époque, il n'est pas en mesure de répondre aux nouveaux enjeux informatiques et manque d'adaptabilité. C'est pourquoi il a dû être abandonné. L'idée a donc germé de développer un nouveau moteur, qui serait évolutif et plus facilement maintenable dans le temps : Allegro [1]. Initialement pensé comme un nouveau concordancier pour la base de données textuelles Frantext (Montémont, 2014), Allegro a intégré les besoins de différents projets portés par l'ATILF au fur et à mesure de son développement.

On pourrait, à juste titre, se demander quel est l'intérêt de développer un nouvel outil pour l'exploitation des corpus, alors que certains instruments, tels que Corpus Workbench (Evert and Hardie, 2011),

1. Étienne Petitjean est à l'origine de la majeure partie des développements d'Allegro.

offrent déjà des performances tout à fait appréciables. Si nous nous sommes largement inspirés de la syntaxe de requêtes CQL, étant donné que celle-ci est transparente et largement utilisée dans la communauté des linguistes, Allegro permet une variété d'applications plus étendues. Nous avons également développé nos propres algorithmes de recherche pour optimiser les temps de réponse lors de la recherche de formes dans un lexique ou de l'exécution de requêtes sur corpus. Pour l'instant, Allegro tourne uniquement sur les serveurs de l'ATILF, mais les sources seront disponibles et publiques très bientôt. L'objectif, à moyen terme, est de proposer à la communauté scientifique un instrument gratuit, dérivé de Frantext, qui permettrait aux utilisateurs de définir leurs propres dépôts avec leurs fichiers et leurs métadonnées. Les chercheurs seraient ainsi autonomes pour créer et interroger leur corpus, choisir de les partager ou les garder privés. Un tel service pourrait trouver sa place sur la plateforme Ortolang.

Dans notre communication, nous présenterons les applications actuelles d'Allegro [2] et introduirons quelques aspects techniques de ce nouveau serveur de données pour l'exploitation de ressources textuelles et lexicales. Nous montrerons à travers ces quatre projets qu'Allegro se prête à des visées scientifiques variées, par le traitement et la visualisation de données de natures très différentes. Il offre ainsi la possibilité d'imaginer de nombreuses applications. En réduisant les efforts de transformation des données et en augmentant l'autonomie entre le cœur des applications Web et leurs interfaces, il permettra à l'avenir de concentrer nos efforts de développement informatique sur l'implémentation de nouvelles fonctionnalités, répondant au mieux aux besoins de la recherche en linguistique, en accompagnement de son évolution.

2 Ressources utilisant Allegro

Les fonctionnalités d'Allegro se sont construites à partir des besoins de cinq projets majeurs, dont quatre sont d'ores et déjà implémentés. Dans cette section, nous ferons le point sur la base Frantext et sa refonte récente, avant de présenter la base Aliento et les projets exploitant respectivement les sources du Dictionnaire de l'Académie Française et du Französiches Etymologisches Wörterbuch (FEW). Enfin, nous terminerons par évoquer le projet de refonte de portail lexical actuellement disponible sur le site du Centre National de Ressources Textuelles et Lexicales (CNRTL).

Frantext. Rendue disponible sur abonnement à travers le logiciel Stella en 1992, la base de données Frantext regroupe aujourd'hui 5 415 ouvrages, en majorité littéraire, pour plus de 250 millions de mots écrits entre 1125 et nos jours. Elle a fait l'objet de différentes évolutions au cours de ces dernières années, comme son annotation complète en parties du discours. Depuis avril 2018, c'est l'ensemble de l'interface d'interrogation et des fonctionnalités qui ont fait peau neuve en se basant désormais entièrement sur Allegro. Pour ce faire, il a fallu concilier les habitudes des utilisateurs et le besoin de nouveautés permettant des exploitations plus poussées.

Aliento. Depuis 2007, l'ATILF est partenaire du projet Aliento, qui vise la mise au point d'une méthodologie permettant l'étude de l'évolution à travers les textes, les époques et les langues, des énoncés proverbiaux et, plus spécifiquement, des énoncés sapientiels brefs (ESB). La base Aliento (Bornes-Varol et al., 2018) décrit aujourd'hui plus de dix mille ESB, répartis entre vingt-deux textes médiévaux écrits en arabe, hébreu, espagnol, catalan et latin et maintenus au format XML-TEI.

2. URL : `https://www.frantext.fr`, `https://www.aliento.eu`, `https://academie.atilf.fr`, `https://few-webapp.atilf.fr/`

Dictionnaire de l'Académie française. L'ATILF assure la conversion au format XML-TEI des articles du Dictionnaire de l'Académie française (pour les éditions 4, 7 et 9). Dans ce cadre, nous avons également réalisé une interface d'interrogation basée sur Allegro. Il s'agit d'un outil d'aide pour les lexicographes en charge de la rédaction des articles. Il permet de rechercher dans des sous-parties spécifiques des articles et de mettre en exergue des éléments de structure.

FEW rétroconverti. Le dictionnaire étymologique et historique du galloroman Französisches Etymologisches Wörterbuch (FEW) (Carles et al., 2019) est en cours de rétroconversion par traitement automatique à l'ATILF. Les trois volumes déjà rétroconvertis forment un corpus d'environ 3 000 articles. Les spécificités typographiques et structurelles de cet ouvrage le rendent délicat à manipuler et à exploiter à l'aide des instruments existants. Allegro nous a permis de répondre à ces besoins spécifiques et joue aujourd'hui un rôle dans la chaîne même de traitement de rétroconversion, en permettant d'être au plus proche de la version XML de la ressource interrogée et d'en diagnostiquer aisément les imperfections.

Portail lexical. Le portail lexical du CNRTL regroupe un ensemble de ressources linguistiques et d'outils sous la forme d'une interface de consultation simple et conviviale. Il répond à plus de 700 000 requêtes par jour. Développé en 2006, il utilise les technologies Web de l'époque (XHTML, PHP). Le poids de l'âge commence à se faire sentir, particulièrement au niveau des interfaces graphiques. Les ressources utilisées subissent un traitement spécial pour modifier leur structure afin de les insérer dans une base de données relationnelle. Chaque ressource fait l'objet d'un traitement coûteux en termes de développement et de temps. Nous envisageons une refonte complète de l'application en 2020. Elle s'accompagnera d'une mise à jour technique majeure en utilisant exclusivement Allegro pour stocker et exploiter toutes les ressources visibles sur le portail.

3 Quelques considérations techniques

Allegro est constitué de trois composants logiciels distincts : un indexeur, un environnement d'exécution et un serveur. L'indexeur prend en entrée les données et les métadonnées, il les restructure et produit un format de sortie optimisé permettant de faire des recherches efficaces aussi bien sur les données que sur leur structure. L'environnement d'exécution est le cœur du système. Il permet de définir un corpus, d'effectuer des requêtes et de récupérer les résultats dans le format de sortie choisi. Le serveur est le composant de plus haut niveau. Il encapsule l'indexeur et l'environnement d'exécution pour donner accès à toutes leurs fonctionnalités depuis un serveur Web. Ce composant gère également les autorisations d'accès aux ressources. Allegro offre une interface REST permettant de l'interroger et de l'administrer. Il est ainsi facile de l'utiliser à partir de n'importe quel langage, comme nous le faisons actuellement en TypeScript et en Go.

Les ressources au cœur des projets de la section précédente peuvent toutes être considérées comme des corpus textuels semi-structurés, de taille importante, que nous souhaitons partager avec la communauté scientifique en les accompagnant d'instruments d'exploration. La majeure partie d'entre elles sont en cours d'évolution, ce qui implique des mises à jour régulières. Bien que ces ressources disposent chacune de leurs propres modalités d'interrogation et d'affichage, elles requièrent toutes la réalisation de requêtes portant à la fois sur le contenu textuel et la structure des documents qu'elles regroupent, auxquels viennent parfois s'ajouter différentes couches d'annotation du contenu textuel.

Pour répondre à ces besoins récurrents, Allegro a été conceptualisé et développé de manière à

simplifier la mise à disposition des ressources. Il offre la possibilité de les exploiter sans transformation préalable, quel que soit leur format (texte brut, XML, CSV), pour peu qu'elles soient encodées en UTF-8. Nous limitons ainsi la multiplication de versions parallèles des ressources et nous épargnons l'écriture de chaînes de traitements spécifiques. L'ensemble des métadonnées associées aux ressources doit pour sa part être fourni dans un format JSON. Il n'y a aucun schéma de base à respecter pour chaque entrée. Le format est entièrement libre et nous pouvons donc ajouter n'importe quel type de métadonnées (chaîne de caractère, entier, flottant, booléen, null). Ces métadonnées sont ensuite directement interrogeables. Dans le cas de Frantext, chaque œuvre est ainsi associée à son année de publication, son auteur, son genre, etc., qui deviennent autant de facettes dans son interface d'interrogation utilisées pour constituer les corpus de travail.

La taille des ressources et les temps d'exécution ont également été pris en considération. Allegro permet d'exploiter des corpus textuels comportant autant de couches d'annotation que souhaité et la taille de ces corpus n'est limitée que par la RAM disponible. L'indexation de l'intégralité de la base Frantext se réalise en une minute environ.

4 Conlusion

Le logiciel Allegro, développé au départ comme un simple concordancier, a évolué pour devenir une plateforme complète destinée à simplifier la mise en ligne et l'exploitation des ressources textuelles et lexicales. Déjà dotée de riches fonctionnalités dans sa version actuelle, la plateforme, grâce à son architecture modulaire, permettra de nombreuses évolutions (nouveaux formats en entrée, nouveaux types de requêtes, etc.) lorsque de nouveaux besoins apparaîtront.

Références

Bernard, P., Lecomte, J., Dendien, J., and Pierrel, J.-M. (2002). Un ensemble de ressources informatisées et intégrées pour l'étude du fançais : FRANTEXT, TLFi, Dictionnaire de l'Académie et logiciel Stella, présentation et apprentissage de leur exploitation. In *Actes. 9ème Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN). Nancy, palais des congrès. 24-27 juin 2002*, volume 2, pages 3–36.

Bornes-Varol, M.-C., Husson, B., and Ortola, M.-S. (2018). La base de données Aliento : Bilan. *Aliento : échanges sapientiels en Méditerranée* , 10 :5–12.

Carles, H., Dallas, M., Glessgen, M., and Thibault, A. (2019). *Französiches Etymologisches Wörterbuch, Guide d'utilisation*. Bibliothèque de Linguistique Romane, Hors série 5. Éditions de linguistique et de philologie.

Dendien, J. (1991). Access to information in a textual database : access functions and optimal indexes. In *Research in Humanities Computing, Papers from the 1989 ACH-ALLC Conference, Oxford : Clarendon Press*.

Evert, S. and Hardie, A. (2011). Twenty-first century corpus workbench : Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*. University of Birmingham.

Montémont, V. (2014). Frantext, une base de données pour la recherche. In *Corpus de textes écrits et oraux : quels usages pour la recherche*, Mons, Belgium. Michel Berré.