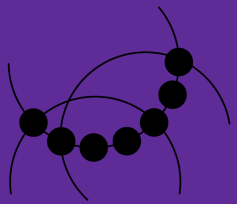
A data repository for the management of dynamic linguistic datasets

Thomas Gaillat[1], Leonardo Contreras Roa[2], Juvénal Attoumbré[1]
[1]Université de Rennes, [2]Université de Picardie Jules Verne

CLARIN Annual Conference 2021



1. Introduction

Most available language corpora are static and risk obsolescence.

Proposal: using a dynamic database to provide a seamless workflow from data ingestion to data querying and data set creation

2. Current Practices

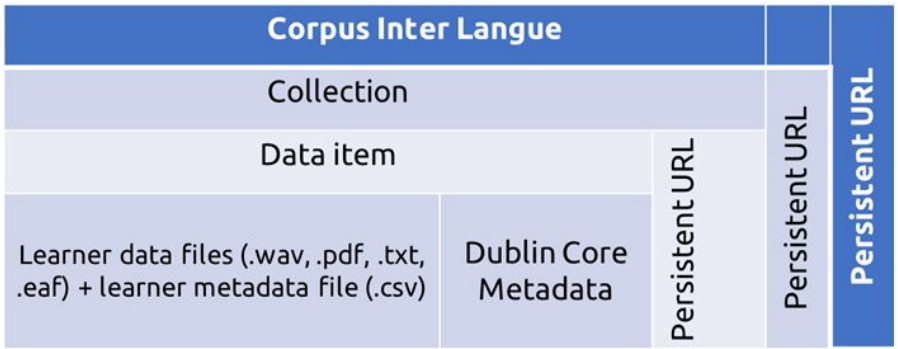
- Use of single persistent URLs for entire datasets
  - Full downloads from repositories - Ex: Ortolang (Huma-Num)

- A move towards interoperability
  - NLP pipelines
  - Interoperable tools and datasets - Ex: European Language Grid

3. Use case: A data repository for a learner corpus

Corpus CIL

- English and French L2
- 115 speakers but 84 recordings in DB
- Spoken and written modules

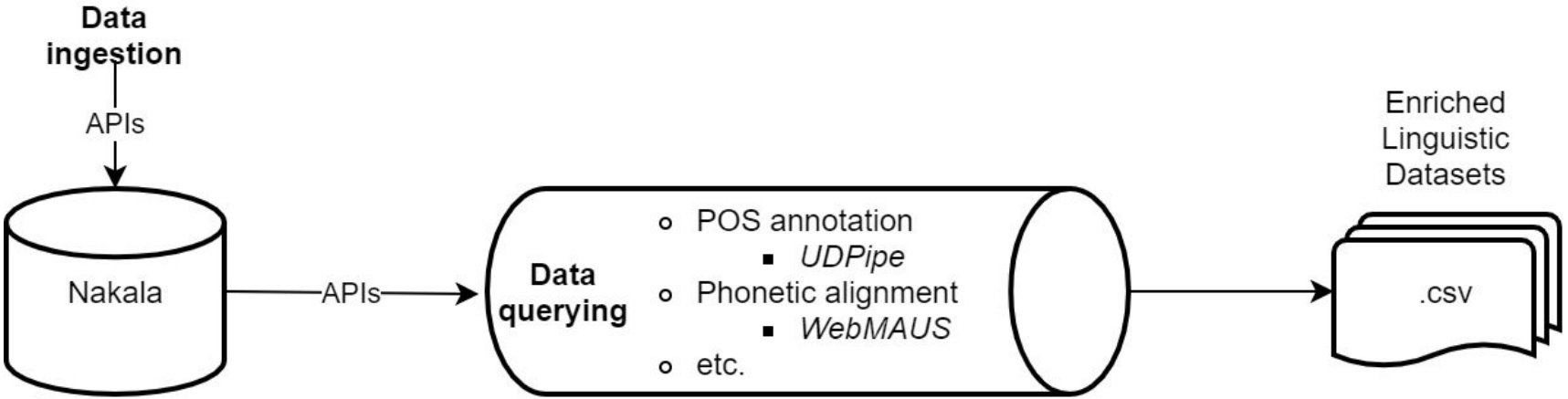


Repository:
Data are organised in collections of data items
Each data item is assigned a persistent URL
Data items are grouped in collections
Each collection is assigned a persistent URL as well

Structure of CIL in Nakala:
One learner = one data item
One data item <- Dublin Core metadata + data files (incl. learner metadata file)

3. Use case: A data repository for a learner corpus

Generating linguistic datasets



4. Perspectives

- development of customised scripts (R) for data annotation
  - semantic, morphosyntactic, phonological and phonetic
- using annotated data for didactic purposes
  - predicting learner proficiency level, determining weaknesses and strengths
- need for global state versioning of corpus
  - more precise version bookkeeping
  - longitudinal comparisons within one same corpus

References

Arbach, N. (2015). *Constitution d'un corpus oral deFLE : enjeux théoriques et méthodologiques* [Phd thesis, Université Rennes 2].

Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouyé, M., & Zarrouk, M. (in press). Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, *34*(1).

TGIR Huma-Num, (2021). *Documentation NAKALA. Documentation des services de la TGIR Huma-Num*.

Thomas Kisler, Uwe Reichel, and Florian Schiel (2017). *Multilingual processing of speech via web services*. Computer Speech & Language, 45:326 – 347.

Jan Wijffels (2020). *Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. R package version 0.8.5.

Pierrel, J.-M. (2014). Ortolang. Une infrastructure de mutualisation de ressources linguistiques écrites et orales. *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle*, *11*(11–1), Article 1. https://doi.org/10.4000/rdlc.1724

Rehm, G., Bontcheva, K., Choukri, K., Hajič, J., Piperidis, S., & Vasiļjevs, A. (Eds.). (2020). *Proceedings of the 1st International Workshop on Language Technology Platforms*. European Language Resources Association. https://www.aclweb.org/anthology/2020.iwltp-1.0

Sérasset, G., Witt, A., Heid, U., & Sasaki, F. (2009). Multilingual language resources and interoperability. *Language Resources and Evaluation*, *43*(1), 1–14.