1. General information

Brief description of the work in general terms

On one hand, resources such as data, text, and multimedia are essential for the long-term preservation of languages and for their communities to use and develop their language. On the other hand, major tech companies like Amazon, Google, Microsoft, and others have been harvesting all the data and knowledge from the web, structuring them in the form of knowledge graphs, and using them to support their applications and services. This has worked for dominant languages, but there is little or no support for indigenous languages that are affected by language policies and have an endangered status. We enable under-resourced language communities to represent their knowledge and language as knowledge graphs, which will act as a common sense layer for AI applications, facilitating AI reasoning to operate more responsibly and transparently.

2. Results

Description of the results achieved

- A schema for modelling and integrating linguistic data as knowledge graphs.
- Collecting and modelling resources in a structured format. E.g., for the Puno Quechua language are stored on Mozilla Data Collective (https://mozilladatacollective.com)
- The hosting and storing of under-resourced language knowledge graphs based on FAIR and LOD principles using an Open Source and Community-drive approach: Wikibase (http://wikibase.cloud)

3. Planned follow-up activities

List of remaining activities (timetable)

February to May 2026: User testing phase
- Linking Lexemes with Pronunciation Audio files based on the Lexeme model
- User testing of the final platform that provides access to the knowledge graph
June to August 2026:
- Feedback analysis and platform adjustments
September and October 2026
- Finalizing thesis draft that includes the additional results and insights
November and December 2026
- Preparing for defense

4. Suggestions for continuation by third parties

What continuations for third parties result from your work?

- Representation of words as lexemes for under-resourced languages of their interests.
- Creation of knowledge graphs for under-resourced languages across Europe and the world.
- Linking to a bigger knowledge ecosystem where all words are interlinked between knowledge graphs.