Managing Fine-grained Metadata for Text Bases in Extremely Low Resource Languages: the Cases of Two Regional Languages of France

Marianne Vergez-Couret, Delphine Bernhard, Michael Nauge, Myriam Bras, Pablo Ruiz, Carole Werner

Managing Fine-grained Metadata for Text Bases in Extremely Low Resource Languages: the Cases of Two Regional Languages of France

Marianne Vergez-Couret[♣], Delphine Bernhard[△], Michael Nauge[♣], Myriam Bras[♢], Pablo Ruiz Fabo[△], Carole Werner[△]
[♣] Université de Poitiers, FoReLLIS UR 15076, F-86000 Poitiers, France
[△] Université de Strasbourg, LiLPa UR 1339, F-67000 Strasbourg, France
[♢] Université de Toulouse, CLLE UMR 5263, F-31000 Toulouse, France
marianne.vergez.couret@univ-poitiers.fr, dbernhard@unistra.fr, michael.nauge@univ-poitiers.fr
bras@univ-tlse2.fr, ruizfabo@unistra.fr, wernerc@unistra.fr

Abstract

Metadata are key components of language resources and facilitate their exploitation and re-use. Their creation is a labour intensive process and requires a modeling step, which identifies resource-specific information as well as standards and controlled vocabularies that can be reused. In this article, we focus on metadata for documenting text bases for regional languages of France characterised by several levels of variation (space, time, usage, social status), based on a survey of existing metadata schema. Moreover, we implement our metadata model as a database structure for the Heurist data management system, which combines both the ease of use of spreadsheets and the ability to model complex relationships between entities of relational databases. The Heurist template is made freely available and was used to describe metadata for text bases in Alsatian and Poitevin-Santongeais. We also propose tools to automatically generate XML metadata headers files from the database.

Keywords: Text bases, Metadata, Text typology, Variation, Regional languages of France

1. Introduction and Objectives

Metadata for text bases are important for describing, querying, filtering, analysing, visualising and sharing corpora. The critical role of metadata has been acknowledged since the beginnings of corpus linguistics and in pioneering works such as the British National Corpus (BNC). In the BNC (Leech, 1992), criteria used for designing a balanced corpus (subject field / domain, genre, level, date, demographics, discourse type, etc.) are detailed in the *header*[1] and may thus be used to perform precise analyses of language facts observed in the corpus.

Yet, Soria and Mariani (2013) observed the following: "The majority of language resources is still poorly documented or not documented at all, and use of metadata elements to describe and document resources is still uncommon and often inconsistent. [...] Single authors can find it difficult to mention their own resources, simply because they can have a hard time deciding the relevant set of metadata elements to be used. Moreover, there is no sufficient awareness about the importance of documentation, which is often disregarded as a useless burden."

While the situation has improved since then, in line with the FAIR principles (Wilkinson et al., 2016), inconsistent, missing or inadequate metadata are still an issue. For instance, extremely large text bases collected from the web for the purpose of training large language models, e.g. CommonCrawl, are much less documented than carefully crafted balanced corpora. While recent endeavours aim at better documenting, filtering and cleaning those data sets, such as OSCAR (Abadji et al., 2022), it is still mostly infeasible to automatically classify documents into precise categories and detailed metadata. As a consequence, information about each source is usually limited to its URL, its date of collection, its language and some simple annotations. This leads to limitations in the possibility of using only some relevant subparts of the data set for specific tasks or explaining systems trained on these data.

In this paper, we argue that providing high-quality and precise metadata is even more crucial for text bases in extremely low-resource languages with several levels of variation: variation in space (diatopic), time (diachronic), usage (diaphasic) and social status (diastratic). Corpora in these languages are often small and the scarcity of data may amplify the impact of biases present in the corpus. This is because there is little chance that they will be smoothed out by other data, as may be the case in larger, more varied corpora. Metadata databases are particularly efficient in helping corpus builders identify potential biases.

[1] http://www.natcorp.ox.ac.uk/docs/URG/cdifhd.html

SIGUL2024 Workshop, pages 212–221

Working on low-resource languages also comes with its own set of constraints, often related to the lack of human and financial resources. This leads to a need for greater efficiency, so that the human and financial resources available are used as productively as possible. Soria et al. (2013) detail several practical recommendations for the development of language resources for lower-resource languages, stressing the need for accurate and reliable documentation, thus guaranteeing the reusability and discoverability of the language resources. Metadata also facilitate the monitoring of digital language support and language resource representativeness by language planning specialists (Giagkou et al., 2022).

However, providing metadata for texts from "minority" literary traditions presents specific challenges. These texts are often understudied, resulting in a lack of information about the authors, including their biography, date and place of birth and about their literary characteristics, such as genre, register, or type of discourse. Work on metadata databases enables this information to be collected and made available, laying the foundations for a more inclusive literary history and preservation of cultural heritage.

Characterising the language variety of a document is another of these challenges. Firstly, there is a lack of language codes (see Section 2.2), which creates tension between adherence to international standards and the need for detailed language characterisation. Secondly, retrieving this information may be difficult when the biography of the author is unknown, as previously mentioned. Additionally, filling in this type of information requires specialists who know the language well enough to be able to recognise its varieties.

Finally, the "burden" of metadata documentation is also related to the lack of appropriate tools to assist in this task, beyond simple spreadsheets.

In this research, we first perform an in-depth survey of metadata for text corpora with a special focus on several levels of variation (Section 2). We also analyse tools which can be used for describing metadata (Section 3). Based on this survey, we propose a metadata model tailored to the specific properties of small-scale corpora collected to represent variation in low-resource languages: here we focus on two regional languages of France, Poitevin-Santongeais and Alsatian (Section 4). We implement this model as a Heurist (Johnson, 2008) database structure and make the model available as a Heurist template, for use in other similar projects. We use the model to manually describe text bases for Alsatian and Poitevin-Santongeais (Section 5). Finally we present tools for automatically generating XML metadata files out of CSV files exported from the database in Section 6.

2. Overview of Metadata in Existing Text Repositories

Table 1 summarizes the available metadata for a selection of representative online text bases for French and several moderate or low resource regional languages of France (Alsatian, Basque, Catalan, Corsican, Occitan, Picard, Poitevin-Saintongeais). The metadata used to search these databases include diatopic variation for languages of France other than French, the date of publication (and the date of creation for BaTelÒc), sometimes diatopic or generational information about the writers, as well as information about the type of text (usually the genre, but also domain or derivation: original or translation). Besides, the metadata available for searching the corpus do not exclude the existence of more extensive metadata to describe the data sets, which is often the case.

Our analysis of these text bases shows that, while dialects are usually described, information about the biography of the authors/speakers is not always available. Filtering based on text type is usually possible, but the categories used across the different text bases are not consistent and do not refer to a standard controlled vocabulary.

In the rest of the section, we survey and detail metadata in existing text repositories for a wider array of languages. Following Menzel et al. (2021), we distinguish between 'descriptive metadata' (minimal metadata, language and script) and 'derived metadata' (biographical information about authors and speakers, document curation, text typology). The first type serves "identification and discovery" purposes, while the second type "enhance[s] the '(re)usability' of a corpus for an intended user community" (Menzel et al., 2021).

2.1. Minimal Descriptive Metadata

Minimal metadata concerns descriptive elements which can be found in generic resource description schemas. We plan to deposit documents from our text base on the Nakala data repository[2] maintained by the French Huma-Num research infrastructure. Nakala assigns permanent DOI to resources and provides an API as well as an OAI-MPH endpoint to harvest resources and their metadata. Nakala has a set of 5 compulsory (data type, title, authors, creation date, license) and 3 recommended (description, keywords, language) metadata.[3] These are inspired from the DublinCore,

[2] https://nakala.fr/

[3] https://documentation.huma-num.fr/nakala-guide-de-description/

Text bases	Languages, dialects and spelling conventions	Date	Authors and Speakers	Text typology
BaTelÒc[a]	Various dialects and spelling conventions for Occitan	Edition date and Creation date	Date of birth	Genre
Corpus Textual Informaticat de la Llengua Catalana[b]	Various dialects for Catalan	Edition date		Derivation (original or translation), text type
Frantext[c]	French	Edition date	French and francophone	Genre, Domain, Channel (book or manuscript)
ParCoLab[d]	Alsatian, Corsican, French, Occitan, Poitevin-Santongeais[e]			Domain, Derivation (original or translation)
PicarText[f]	Various dialects for Picard		Date of birth, "reference" location	Genre
XX Mendeko Euskararen COrpus[g]	Various dialects for Basque	Edition date		Genre

Table 1: Metadata for text bases for languages of France.

[a] http://redac.univ-tlse2.fr/bateloc/
[b] https://ctilc.iec.cat/scripts/
[c] http://www.frantext.fr
[d] http://parcolab.univ-tlse2.fr/
[e] ParCoLab also includes Serbian, English and Spanish documents, it was originally designed as a Serbian/French/English parallel corpus.
[f] https://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/
[g] http://xxmendea.euskaltzaindia.eus/Corpus/

and it is possible to additionally include elements from the qualified DublinCore.[4]

2.2. Language and Script

Languages can be described using several language codes: ISO 639-3, Glottolog (Hammarström et al., 2023) or WALS (Dryer and Haspelmath, 2013). The writing system (or script) is also worth documenting, using the ISO 15924 four letter code.[5] All three language code categorisations as well as the writing system are documented in the TeDDi sample corpus (Moran et al., 2022).

In our text bases for Alsatian and Poitevin-Saintongeais, only two scripts are represented: Latin (`Latn`) and Latin Fraktur (`Latf`). Latin Fraktur is only used in older Alsatian documents. But, even though Alsatian and Poitevin-Saintongeais are recognised as "languages of France",[6] existing language codes and classifications are incomplete or lack precision for both languages. `gsw` is the ISO 639-3 code for Alemannic, which encompasses both Alsatian and Swiss German (codes such as `gsw-FR` to specify that the language is spoken in France, or `gsw-u-sd-fr67`[7] to iden-

[4] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[5] https://www.unicode.org/iso15924/codelists.html
[6] https://www.culture.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France/Agir-pour-les-langues/Promouvoir-les-langues-de-France/Langues-regionales
[7] `u` refers to the Unicode locale extension subtag, `sd` to regional subdivision and `fr67` to the Bas-Rhin department. See https://en.

tify the variant spoken in Bas-Rhin, could be used, following BCP-47). Poitevin-Santongeais has no ISO 639-3 code, the `fra` code for French would have to be used which is absolutely unsatisfactory. Glottolog provides codes for Poitevin (`poit1240`), Santongeais (`sant1407`) and Low Alemannic Alsatian (`alsa1241`), but they are classified in a way that is not entirely appropriate: Alsatian as a dialect of Central Alemannic (Alsatian is an ambiguous umbrella term, which actually includes non-Alemannic Franconian dialects spoken in the Alsace region), and Poitevin and Saintongeais as dialects of French. WALS has a code for Alsatian (`alt`) but none for Poitevin-Saintongeais. Steps are currently being taken with SIL International to provide Alsatian and Poitevin-Saintongeais with an ISO code. Although the creation of these language codes will be a major step forward for both languages, they will not even be sufficient to document our target languages efficiently, and ad hoc classifications will have to be used for diatopic variants. This choice has also been made by Pettersson and Borin (2019) who describe the specific language variety in addition to the ISO 639-3 code. We will also be approaching Glottolog to harmonize our language/dialect classifications with theirs.

2.3. Biographical Information about Authors or Speakers

With the development of oral corpora, metadata describing speakers (age, gender, occupation, etc.) began to appear. But metadata about authors for databases of written texts are just as relevant, given the intra-individual variation depending in particular on age and geographical origin (Combettes, 2022). In the case of text databases for minority languages, speakers' linguistic skills vary according to their date of birth, which is a relevant metadata implemented in text bases for minority languages such as Occitan and Picard, respectively BaTelÒc (Bras and Vergez-Couret, 2016) and PicarText (Eloy et al., 2015). BaTelòc metadata also include additional information on the author, such a his/her date of death and the localisation of his/her language, although not used as criteria to select texts up to now. As mentioned previously, the collection of biographical information can however be difficult, if not impossible, for lesser known authors from the past.

Metadata about authors can also be connected to Wikidata[8] and other linked data repositories, through a unique identifier. Ruiz Fabo et al. (2020) describe the MeThAl project which aims at building a diachronic corpus of Alsatian theatre plays. Authors and theatre plays are associated to their Wikidata identifiers and new Wikidata identifiers were created if needed. Publisher locations for each play were also collected, although their relation to authors' and characters' language varieties is of course very indirect. Another example of documenting metadata potentially indicative of authors' or speakers' biographical information is found in Pettersson and Borin (2019), who recorded the location where the text was produced.

2.4. Document Curation

We use the term "document curation" to describe the procedures applied to the original text (be it printed or digital) in order to obtain the final digital document included in our text bases. These procedures include digitisation, OCR, correction of the OCR, manual transcription, alignment of parallel texts, etc. They are carried out within a project by identified personnel whose contribution must be acknowledged.

The metadata of BaTelÒc (Bras and Vergez-Couret, 2016) document the person responsible for acquiring the text and its rights, the organisation that publishes the TEI XML document and the person responsible for creating the TEI XML file. They also document the people involved in entering metadata in the metadata database, and in the TEI XML encoding process, as well as editing decisions or modifications such as typing error correction on the text. Pettersson and Borin (2019) include metadata about the digitisation method, the transcription principles and the name of the transcriber. Kevers (2022) document the person who is primarily responsible for creating the TEI XML document, the organisation that publishes the TEI XML document, the people involved in the TEI XML compilation and encoding process, the main software used for conversion to text, as well as editing decisions (standardisation, definition of text units, etc.).

2.5. Text Typology

Text typology refers to information about texts based on the communication goals of the author, which lead to the adoption of specific discursive (e.g., genre, register) and text formatting (e.g., layout, organisation, channel) norms. These metadata require an analysis of the texts and lead to their classification into pre-defined categories.

Information on text typology is documented in a very heterogeneous way, depending on the tools or description models used. A simple classification of document types is usually provided in ref-

wikipedia.org/wiki/IETF_language_tag and http://www.unicode.org/reports/tr35/#Locale_Extension_Key_and_Type_Data.

[8] https://www.wikidata.org

erence management tools such as Zotero.[9] In the TEI P5 (TEI Consortium, 2023), the `textDesc` elements describes the channel, constitution, derivation, domain, factuality, interaction, preparedness and purpose of a text. A taxonomy of web registers is proposed by Egbert et al. (2015), with 8 main registers. This categorisation is used by Laippala et al. (2023) for automatically classifying English web documents into registers and by Laippala et al. (2022) for 14 languages, based on the OSCAR corpus. BaTelÒc (Bras and Vergez-Couret, 2016) uses 16 genre categories: novel, literary tale, memoir and chronicle, short-story, essay, poetry, play, song, correspondence, speech, treaty, traditional oral storytelling, scientific text, press, oral text, other. Moran et al. (2022) describe 6 broad and 25 narrow genre categories used to organise their collection of text samples for typologically diverse languages. In a similar way, Pettersson and Borin (2019) use a two-level taxonomy of genres for describing historical corpora. The CAHIER text typology thesaurus (Galleron et al., 2021)[10] describes a very detailed taxonomy with 368 concepts and 9 broad categories: domain, factuality, form, genre, contents layout, origin, target audience, channel, discourse type. Each concept is identified with a persistent identifier in the form of a Handle URI. To the best of our knowledge, this thesaurus provides the most complete typology for literary texts.

Overall, there is no standard textual typology that covers all possible types. The CAHIER typology is mainly oriented towards literary texts and therefore lacks descriptors for other texts, while the taxonomy of web registers by Egbert et al. (2015) is naturally oriented towards web content and does not deal with printed literary works. Furthermore, the typologies are not always based on clearly established criteria, which leads to some confusion between different notions and terms such as genre, register or domain.

We argue that it is important to refer to existing typologies for comparability and interoperability (in accordance with the FAIR principles), rather than creating a new typology. In addition, the use of multiple vocabularies reduces the risk of documents not being described or being assigned to an inappropriate category. The use of controlled vocabularies also ensures consistency through the use of standardised terminology.

[9] https://www.zotero.org/support/kb/item_types_and_fields

[10] https://opentheso.huma-num.fr/opentheso/?idt=43

3. Tools for Describing Metadata

Metadata for text bases need to be handled using appropriate tools, to prevent errors and facilitate the metadata collection and structuring process. Unfortunately, these tools are often not described in research papers, only the resulting metadata.

Spreadsheet software seems to be the simplest and most straightforward solution. The metadata for the ParCoLab and MeThAI projects are managed using Google Sheets. For ParCoLab, the spreadsheet can be filled in via an online form (Stosic et al., 2024).

Relational databases are a more flexible option, in particular for modelling complex metadata. In the BaTelÒc project, a Microsoft Access database has been used to manage five relational tables (source text, author, publisher, document curation, data curator) with a user-friendly interface to enter metadata and a Visual Basic script for the automatic generation of the TEI header of the target XML file. In the TeDDi project, metadata is described in four relational tables, implemented in SQLite (Moran et al., 2022). However, designing and implementing relational databases can be a daunting task for non specialists.

The Heurist data management system (Johnson, 2008; Heurist Team, 2023) combines both the ease of use of spreadsheets and the ability to model complex relationships between entities of relational databases. It is particularly used for digital humanities projects and still unfamiliar to the NLP and langage resources communities. Heurist proposes a no code interface to a relational database, which is well suited for people who are not computer scientists and yet wish to design complex data collections for their research data. Heurist proposes a list of predefined entities that users can choose from and new entity types can be defined. Bulk modifications can be easily performed to change metadata properties for several entities at the same time. In addition, controlled vocabularies can be used to describe entities and new controlled vocabularies can be added to the existing ones. Databases created with Heurist can also be published as websites and complex filters can be built to export parts of the database as CSV or JSON files. In this project, we chose to use Heurist, as it was meeting our needs.

4. Proposed Metadata Model

The proposed metadata model is described in Figure 1.[11] It is based on the metadata used for other text repositories described in the previous section and addresses some of the limitations identified in

[11] The diagram has been generated using the Mermaid tool: https://mermaid.live.

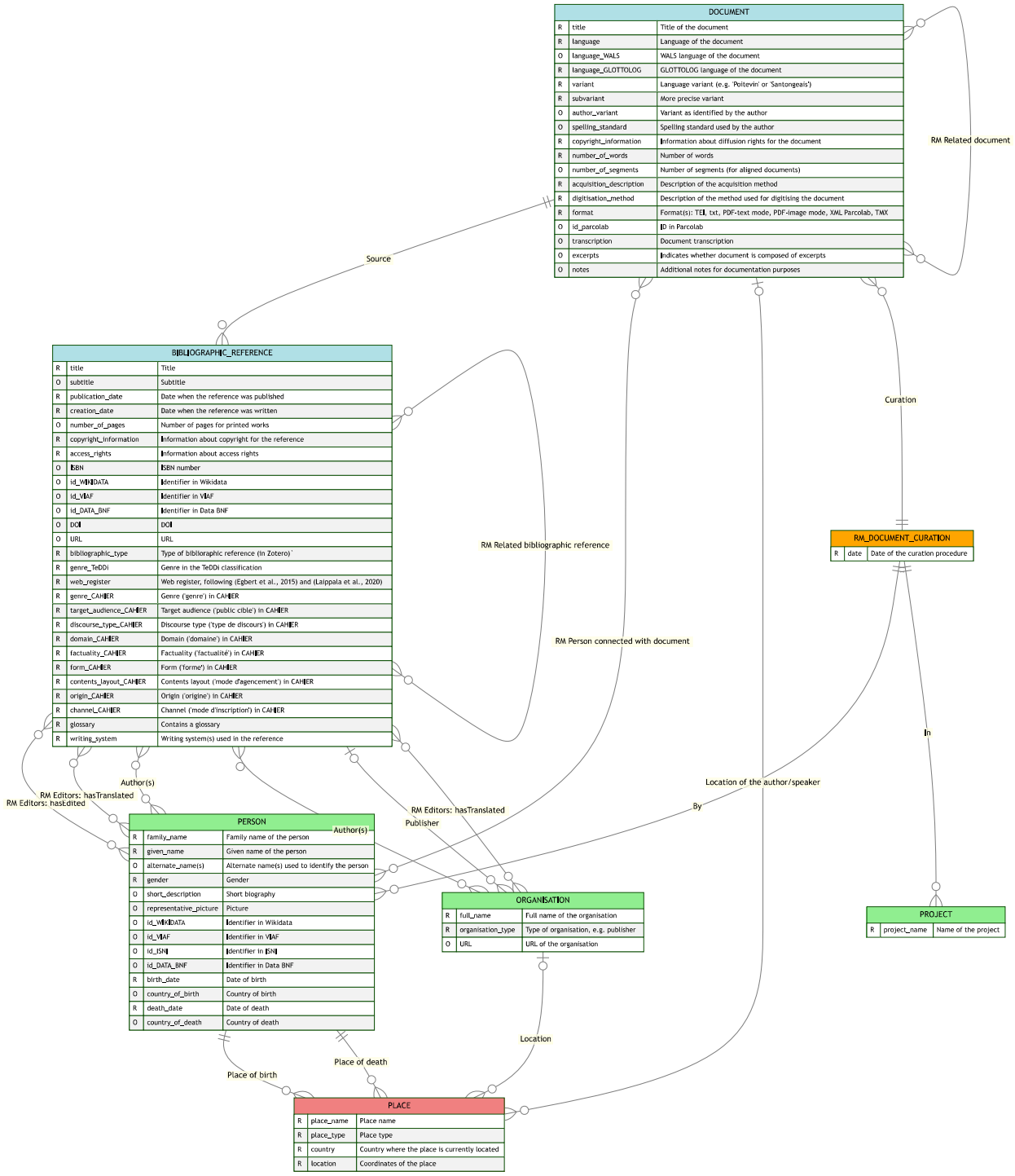

Figure 1: Entity relationship diagram. For the attributes, 'R' indicates that it is required or recommended, 'O' that it is optional. 'RM' indicates a relationship marker, where the relationship is typed with a constrained vocabulary.

our review. We thus propose a unique database model for a variety of texts (literary or web-based), built from a compilation of existing metadata and controlled vocabularies with the aim of providing a model of fine-grained metadata for reusability and interoperability. Data integrity is maintained by using different tables and relationships between tables. This ensures that information is not duplicated unnecessarily. When the database schema was created, every entity and field was described in Heurist to ensure that the database schema was well-documented.

An important distinction is made between the bibliographic reference, which contains all information relevant for a printed or online reference, and the electronic document which is part of the text

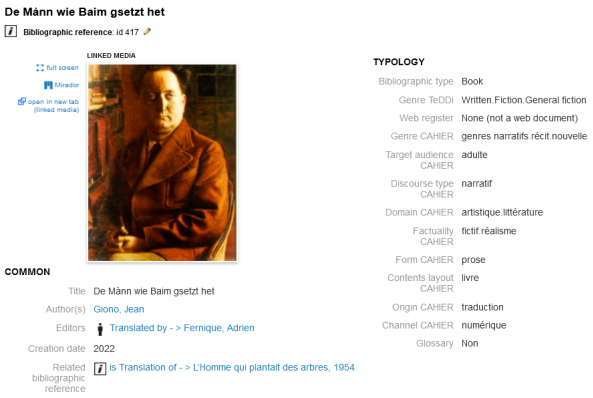

Figure 2: Example bibliographic reference in the Alsatian Heurist database.

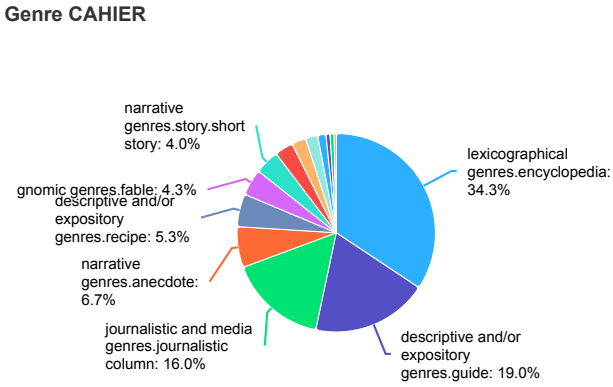

Figure 3: Document genres in the Alsatian database.

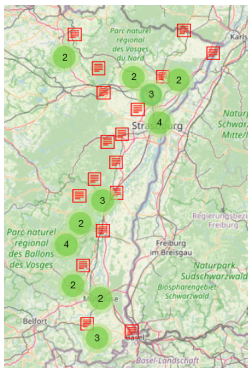

Figure 4: Distribution of Alsatian speakers/authors in the Heurist database.

base (both in blue in Figure 1). This distinction is made because a document can be only a part of a larger reference, e.g. a text in a given language in a multilingual reference. The bibliographic reference contains information about text typology according to the CAHIER, TedDDi and Zotero classifications, as well as web register for web based references.

Information about languages, spelling standards and digitisation are attached to a document.

Bibliographic references can be related using a typed relationship marker: “derivation” (translation, adaptation, subtitling, spelling variant) or “part of” another reference (extract, chapter, preface). The same relationship marker can also be applied to documents.

There is a specific relationship marker for document curation (marked in orange), which indicates who did the curation, within which project and when.

The other entities describe people (authors, curators, translators), organisations (editors, associations) and projects. There is also an entity type for places (birth / death places, editors’ location).

Both Heurist databases for Poitevin-Saintongeais and Alsatian have been registered as Heurist templates, with the following IDs: 1471 (Poitevin-Saintongeais) and 1564 (Alsatian). Record types can thus be imported in new Heurist databases by interested users.[12]

5. Text Bases for Alsatian and Poitevin-Santongeais

The metadata model described in Figure 1 was thoroughly tested and refined by inserting hundreds of representative records to describe metadata for text bases for Alsatian and Poitevin-

[12]For help, see https://int-heuristweb-prod.intersect.org.au/heurist/?db=Heurist_Help_System&website&id=39&pageid=627

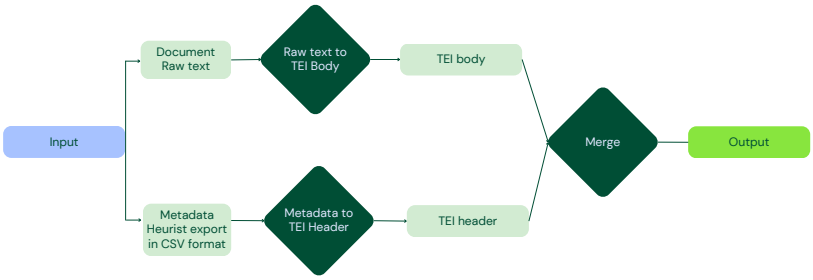

Figure 5: Generation of TEI files from the Heurist database.

Santongeais. This allowed problems to be identified and resolved. We also verified that all entities and relationships were captured in the model. The collection of metadata is still ongoing, and the databases will continue to grow in the coming months.

Currently, the text base for Alsatian contains 115 bibliographic references and 301 documents, along with 53 persons, 52 places, 27 organisations, 2 projects. Figure 2 shows a screenshot of a Heurist record for a bibliographic reference. The reference is related to other entities in the database: persons (author, translator), related bibliographic reference (original reference in French). Figure 3 shows the distribution of the genres of the documents according to two different genre typologies: the Heurist CSV export function generates files which are easy to process with data analysis and visualisation programs. The CAHIER typology is more fine-grained than the TeDDi typology and both allow for a complementary description of the resources. Some visualisations are also directly available within Heurist, such as the map which shows the locations of authors/speakers of documents (see Figure 4).

The text base for Poitevin-Saintongeais has originally been designed by Liliane Jagueneau for literary texts. Currently, it contains 150 bibliographic references and 31 documents, along with 114 persons, 122 places, 94 organisations and 2 projects. The texts are only literary texts but we intend to diversify with various genres such as web documents and newspaper articles.

6. Automatic Generation of XML-TEI files

At the same time, tools have been designed to automatically generate XML-TEI format files with metadata headers,[13] since documents described in the Heurist database will be made available, in particular on the ParCoLab platform. More specifically, a set of scripts create XML-TEI files in the expected format for corpus repositories in the ParCoLab aligned text library, from CSV files containing metadata extracted from the Heurist database and plain-text documents. The general process of the scripts can be described as follows, see Figure 5:

1. Generating XML-TEI headers files from CSV files containing metadata extracted from the Heurist database;
2. Generating XML-TEI body files from plain-text documents;
3. Assembling XML-TEI header and body pairs.

The scripts are based on a more generic tool for converting a metadata file (CSV) to XML header.[14] This generic tool uses a simple mapping file giving correspondences from a column in the CSV file to an element in the target XML tree. For instance, `Subject` is mapped to the TEI `<keywords type="subject">` element.

7. Conclusion and Perspectives

Metadata are key components of language resources and facilitate their exploitation and re-use. In this article, we addressed the management of metadata for two regional languages of France and proposed a metadata model based on a survey of metadata in existing text repositories. We showed that the Heurist data management system presents several advantages for this task: ease of use, modelling of complex relationships between entities, controlled vocabularies, bulk modifications.

The metadata model proposed for Poitevin-Santongeais and Alsatian texts in the Heurist system may benefit other regional languages of France. For instance, Occitan metadata of BaTelÒc could be managed by the open Heurist system rather than by a commercial application. In the future, we would like to develop tools to evaluate the quality of our metadata, following the characteristics proposed by Bruce and Hillmann

[13] https://gitlab.huma-num.fr/mshs-poitiers/forellis/parcolab_tools

[14] XMLify: https://gitlab.huma-num.fr/mshs-poitiers/plateforme/xmlify

(2004), in particular completeness, accuracy, logical consistency and coherence.

8. Ethics Statement

We only include publicly available information about persons in the text bases. Copyright information is detailed in the metadata for bibliographic references.

9. Acknowledgements

This work has been carried out within the framework of the ANR-21-CE27-0004 DIVITAL project supported by the French National Research Agency. We would like to thank Régis Witz (MISHA, Strasbourg) and Gaëlle Coz (MSHS, Poitiers) for their support in designing the database.

10. Bibliographical References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Myriam Bras and Marianne Vergez-Couret. 2016. BatelÒc: A text base for the Occitan language. *Language Documentation and Conservation in Europe*, Special Publication No. 9:133–149.

Thomas R Bruce and Dianne I Hillmann. 2004. The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In Dianne I Hillmann and DL Westbrooks, editors, *Metadata in Practice*, pages 238–256. ALA editions, London.

Bernard Combettes. 2022. Suggestions for a diachronic text linguistics. In D. Ablali and G. Achard-Bayle, editors, *French theories on text and discourse*, pages 169–183. De Gruyter, Berlin.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Jean-Michel Eloy, Fanny Martin, and Christophe Rey. 2015. PICARTEXT : Une ressource informatisée pour la langue picarde. In *Actes de TALaRE 2015 - Traitement Automatique des Langues Régionales de France et d'Europe, Atelier associé à TALN - 2015 22e conférence sur le Traitement Automatique des Langues Naturelles*, Caen, France.

Ioana Galleron, Fatiha Idmhand, Alexei Lavrentiev, Marie-Luce Demonet, and Anne Réach-Ngô. 2021. Décrire les textes dans le cadre d'une édition numérique.

Maria Giagkou, Stelios Piperidis, Penny Labropoulou, Miltos Deligiannis, Athanasia Kolovou, and Leon Voukoutis. 2022. Collaborative metadata aggregation and curation in support of digital language equality monitoring. In *Proceedings of the Workshop towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 27–35, Marseille, France. European Language Resources Association.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. Glottolog 4.8.

Heurist Team. 2023. HEURIST: A unique solution to the data management needs of Humanities researchers. https://heuristnetwork.org/.

Ian Johnson. 2008. Heurist: A Web 2.0 Approach to Integrating Research, Teaching and Web Publishing. In *Proceedings of the 36th CAA Conference*, volume 2, pages 291–297.

Laurent Kevers. 2022. CCdC - Le Corpus Canopé de Corse. Technical report, UMR 6240 CNRS LISA - Université de Corse.

Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki-Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. Register identification from the unrestricted open Web using the Corpus of Online Registers of English. *Language Resources and Evaluation*, 57(3):1045–1079.

Veronika Laippala, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, Valtteri Skantsi, Lintang Sutawika, and Sampo Pyysalo. 2022. Towards better structured and less noisy Web data: Oscar with Register annotations. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 215–221, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Geoffrey Leech. 1992. 100 million words of English: The British National Corpus (BNC). *Language research*, 28(1):1–13.

Katrin Menzel, Jörg Knappen, and Elke Teich. 2021. Generating linguistically relevant metadata for the Royal Society Corpus. *Research in Corpus Linguistics*, 9:1–18.

Steven Moran, Christian Bentz, Ximena Gutierrez-Vasques, Olga Sozinova, and Tanja Samardzic. 2022. TeDDi sample: Text data diversity sample for language comparison and multilingual NLP. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1150–1158, Marseille, France. European Language Resources Association.

Eva Pettersson and Lars Borin. 2019. Towards a Swedish diachronic corpus: Intended content, structure and format of version 1.0. Technical Report SCR-03-2019.

Pablo Ruiz Fabo, Delphine Bernhard, and Carole Werner. 2020. Création d'un corpus FAIR de théâtre en alsacien et normalisation de variétés non-contemporaines. In *2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT)*, pages 34–43, Montrouge, France. CNRS.

Claudia Soria and Joseph Mariani. 2013. Searching LTs for minority languages. In *Actes de TALaRE Traitement Automatique des Langues Régionales de France et d'Europe*, Les Sable d'Olonne, France.

Claudia Soria, Joseph Mariani, and Carlo Zoli. 2013. Dwarfs sitting on the giants' shoulders–how LTs for regional and minority languages can benefit from piggybacking major languages. In *Proceedings of XVII FEL Conference*, pages 73–79.

Dejan Stosic, Saša Marjanović, Delphine Bernhard, Myriam Bras, Laurent Kevers, Stella Medori, Marianne Vergez-Couret, and Carole Werner. 2024. Extending a parallel corpus and platform with four regional languages of france. In *Proceedings of LREC-COLING 2024*.

TEI Consortium. 2023. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.6.0. Last updated on 4th April 2023, revision f18deffba.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):1–9.