The Language Resource Life Cycle: Towards a Generic Model for Creating, Maintaining, Using and Distributing Language Resources

Georg Rehm
DFKI GmbH
Berlin, Germany
georg.rehm@dfki.de

Abstract

Language Resources (LRs) are an essential ingredient of current approaches in Linguistics, Computational Linguistics, Language Technology and related fields. LRs are collections of spoken or written language data, typically annotated with linguistic analysis information. Different types of LRs exist, for example, corpora, ontologies, lexicons, collections of spoken language data (audio), or collections that also include video (multimedia, multimodal). Often, LRs are distributed with specific tools, documentation, manuals or research publications. The different phases that involve creating and distributing an LR can be conceptualised as a *life cycle*. While the idea of handling the LR production and maintenance process in terms of a life cycle has been brought up quite some time ago, a best practice model or common approach can still be considered a research gap. This article wants to help fill this gap by proposing an initial version of a generic *Language Resource Life Cycle* that can be used to inform, direct, control and evaluate LR research and development activities (including description, management, production, validation and evaluation workflows).

Keywords: LR infrastructures and architectures, Metadata, Standards for LRs, Life cycle

1. Introduction

Language Resources are an essential ingredient of current research and development approaches in Linguistics, Computational Linguistics, Language Technology and related fields. The respective methodologies usually involve the application of statistical machine learning techniques to vast amounts of data in order to train engines, generate models, create tools or to address novel, data-driven research questions, among others.

Generally speaking, Language Resources are collections of spoken or written language data (including audio and video), typically, but not necessarily, annotated with linguistic analysis information. There are different types of Language Resources such as corpora, ontologies, lexicons, collections of spoken language data (audio), or collections that also include video (multimedia, multimodal). Often, Language Resources are distributed with specific tools, documentation, manuals or research publications.

In order to ensure a certain level of sustainability and interoperability, standards and best practice approaches are used for many resources – either as is or including modifications – for the annotation of various linguistic analysis levels or metadata concerning the formal description of the whole resource. There are many cases in which resources have become popular and proven valuable for the research community. In these cases a resource often serves as the basis of additional research that explores related or maybe even completely different areas by extending the resource with additional language data or with additional linguistic analysis levels that have not been in scope with regard to the research question the resource was originally created for.

The different phases that involve, among others, creating and distributing a Language Resource can be conceptualised as a *life cycle*. While the idea of handling the Language Resource production and maintenance process in terms of a life cycle has been brought up quite some time ago, a best practice model or common approach can still be considered a research gap. According to the FLaReNet Strategic Language Resource Agenda and several other recent reports and assessments, the Computational Linguistics and Language Technology communities have a demand for a reference model for resource development, "including the language resource life cycle" based on the observation that the "management of the life cycle of language resource creation has been largely overlooked in our community" (Calzolari et al., 2011) (p. 15).

This contribution wants to help fill this gap by proposing an initial version of a generic *Language Resource Life Cycle* that can be used to inform, direct, control and evaluate research and development work (including description, management, production, validation and evaluation workflows) in the fields of Computational Linguistics, Language Technologies, Digital Humanities and others. The proposed Language Resource Life Cycle is based on the analysis and generalisation of relevant publications as well as the author's own previous research activities and experience. Section 2. discusses related work. Section 3. provides a brief description of the seven phases of the initial version of the Language Resource Life Cycle. The following Section 4. briefly discusses several potential benefits of the life cycle. Section 5. concludes the article with an outlook on future work.

2. Related Work

Aspects of life cycles of Language Resources have been proposed in previous work by language resource distribution institutions and agencies, platform and infrastructure initiatives, as well as research and development projects carried out in academic institutions, research centres or companies. The notion of resource life cycles has also found its way into textbooks such as, for example, *Natural Language Processing with Python* (Bird et al., 2009) (Chapter 11, "Managing Linguistic Data"), where the authors explain technical and practical aspects of the concept based on the phases *data collection*, *annotation*, *quality control*, and *publication*. The

cycle can continue after the publication of a corpus because it is *modified/enriched during the course of research*. Additionally, a language resource is to be described using standardised metadata such as, for example, OLAC.

More detailed processes and additional phases are provided by (van Veenendaal et al., 2013) who discuss the approach developed in the STEVIN programme. The life cycle described there is mostly concerned with the requirements and processes of the large-scale joint research and development effort STEVIN, including the centrally organised distribution of Language Resources. The five phases used in this life cycle are *acquisition*, *management*, *maintenance*, *distribution* and *support services*. The specific focus – management of many resources instead of the creation of one specific resource – of this life cycle becomes apparent already in the first phase: *acquisition* does not refer to language data but to the actual ownership and intellectual property rights (IPR) of a resource, which is, of course, vital to ensure its long-term accessibility. This phase also includes the *evaluation and validation* of a resource through external parties to gauge a resource's quality and sustainability; here, data is validated against XML schemas, documentation is checked and software is tested. The other phases refer to storing and backing up resources on servers (*management*), preparing distribution versions and regularly performing checks if a resource needs *maintenance*, making resources available through a web shop as a downloadable file or through other media based on open or commercial license agreements (*distribution*), as well as *support services*.

In META-SHARE (Piperidis et al., 2014) a similar yet different scenario with regard to the description of resources with metadata is used: META-SHARE functions as an open resource exchange infrastructure, i. e., when making existing language resources available through META-SHARE, the resources are being described with the built-in metadata editor, which implements the metadata schema that was prepared for META-SHARE specifically. Several automatic procedures exist to transform existing metadata descriptions based on commonly used formats into the META-SHARE format. META-SHARE does not support a full language resource life cycle yet.

(Wittenburg et al., 2010) concentrate on technical requirements that focus upon e-science functionalities and the long-term preservation of resources. Among them are the proper definition of the objects that are the basic units of management including PIDs, checksums and high-quality metadata as well as the aggregation of such objects in arbitrary ways to enable researchers to combine multiple resources in experiments. Data survival and authenticity are to be provided through proper bitstream management and safe replication as well as the use of open standards, especially with regard to the syntax and semantics of the used annotation formats, which are to be registered in schema and concept repositories.

Yet another set of aspects is discussed by (Nicolas et al., 2010) who describe the main guidelines of the Victoria project. The authors concentrate on aspects such as improving the efficiency of collaborative annotation work, using established frameworks and resources as well as licenses that are fit for purpose.

In any life cycle of digital resources, the application of standards on as many levels as possible is an essential aspect (licensing, IPR, data formats, annotation formats, metadata, querying, storage etc.). With regard to the description of resources with metadata, ISOcat CMDI foresees the data element "LifeCycleStatus"[1] that provides an "indication of the status in the life cycle of a resource", for example, "planned, development, released, production, withdrawn, retired, superseded, archived". As the explanation of this data element shows, the scope of this element is beyond language resources proper: "Tools are often released according to different status, development versions and productive versions are common [...], for other language resources the release is the end of the development process. Hence, the status corresponds to the life cycle model of the resource type. The examples selected should be general enough to be usable in various contexts."

3. The Language Resource Life Cycle

Based on the analysis of these relevant publications and previous research and development experience by the author, e. g., (Wörner et al., 2006; Schmidt et al., 2006; Rehm et al., 2008b; Rehm et al., 2008c; Lehmberg et al., 2007; Rehm et al., 2008a; Rehm et al., 2009; Piperidis et al., 2014; Rehm et al., 2014), a comprehensive *Language Resource Life Cycle* is proposed. The author's own contributions highlight best practice insights and specific aspects of selected phases of the proposed life cycle, the initial version of which is shown in Figure 1.[2] The life cycle consists of a set of external factors and forces, the internal project context (i. e., the goal and objective of the Language Resource development project) and seven individual phases:

- External Context and LR/LT Ecosystem and Landscape
- Internal Project Context – Start of the Language Resource Life Cycle
- Phase 1: Data Acquisition and Data Collection
- Phase 2: Data Curation and Data Annotation
- Phase 3: Linguistic Analysis and Research
- Phase 4: Evaluation and Quality Control
- Phase 5: Description
- Phase 6: Packaging
- Phase 7: Distribution and Publication of the Langauge Resource

The start of the life cycle is marked or initiated by a specific *linguistic research question or project* (including areas such as language documentation and language preservation) or a

[1] `http://www.isocat.org/rest/dc/3818`

[2] The author would like to thank the anonymous reviewer who pointed out a similar life cycle for research data: `http://www.data-archive.ac.uk/create-manage/life-cycle`. The LR Life Cycle can be conceptualised as a specialised and more concrete incarnation of the research data life cycle.

Heterogeneous situation outside the life cycle dictates and provides external factors and forces:

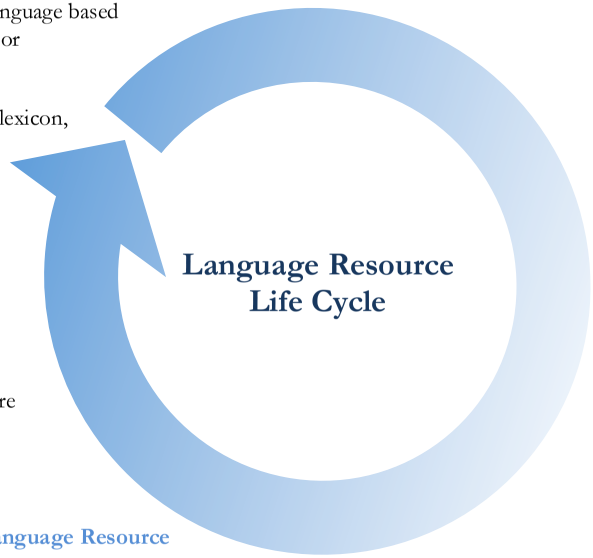
- LR/LT ecosystem and landscape with available resources and technologies for several languages.
- Gaps in terms of resources, technologies, languages.
- Regional, national, international data creation strategies, initiatives and goals to fill gaps and to foster research, innovation and development.

Start of the Life Cycle

- *Linguistic research question* (incl. language documentation and preservation, creation of a resource in a new language based on a popular resource in some other language) or
- *Technology development goal*
- Project goal dictates type of LR to be created (lexicon, corpus, ontology etc.)
- Project goal also dictates some of the subsequent phases of the life cycle.

Life Cycle continues …

- Maintenance checks or user feedback: any errors in the LR? Corrections or data format upgrades needed? Release patch or update.
- New version or new research: acquisition of more data and subsequent annotation and/or annotation of additional linguistic analysis levels (sharing and re-use of language resources).



Phase 1: Data Acquisition and Data Collection

- Data source:
  - *New data* (harvested from the web, acquired using specialised tools, outsourced to service provider etc.)
  - *Existing (legacy) data* in arbitrary research-oriented formats (alternative **start** or **continuation** of cycle)
- IPR and legal issues: who is the rightful owner or copyright holder of the primary data?

Phase 2: Data Curation and Data Annotation

- Selection of linguistic analysis levels
- Formats, tag sets, schemas
  - Standards (SGML, XML, TEI, CES, LAF etc.)
  - Syntax and semantics registered in recognised repository
- Tools (GUIs etc.; software has its own life cycle)
- Annotation: manual, semi-automatic, fully automatic
- Definition/identification of objects through PIDs
- Data management (XML databases, editors, revision control systems etc.)

Phase 3: Linguistic Analysis and Research

- Querying
  - Metadata
  - Data and their annotations
- Analysis: manual, semi-automatic, fully automatic
- Visualisation
- Combining data and annotations from various different sources
  - Interoperability standards

Phase 4: Evaluation and Quality Control

- LR evaluation, quality evaluation
- Validation (schema compliance etc.)
- Reviews
- Independent assessments
- Publications
- Experiments

Phase 5: Description

- Metadata records
  - Standards (OLAC, IMDI etc.)
- IPR and legal issues: who owns the annotations?
- Documentation, manuals, scientific publications
- Software used or implemented to create the LR

Phase 6: Packaging

- Any constraints on the data and annotations based on IPR and legal aspects?
- Example: sentence shuffling

Phase 7: Distribution and Publication of the Language Resource

- IPR and legal issues: license agreement between user/OEM and distributor
- Distribution through web interface
  - Querying: metadata, data annotations, both
  - Visualisation
  - Optional: distribution as stand-alone package
- Distribution through APIs (Linked data, REST etc.)
- Distribution as stand-alone package
  - LR provider distributes through own mechanism/website
  - Through a centralised or decentralised distribution infrastructure; these have their own life cycles: acquisition of LR, management of LR, maintenance, distribution, support

Figure 1: Language Resource Life Cycle

technology development goal; the linguistic research question also includes the possible creation of a resource in a new language based on a popular resource in some other language (such as, for example, WordNet spawning cousins in other languages). A focus on the linguistic research aspect would entail, naturally, a focus on Phase 3 (the actual Linguistic Analysis and Research with aspects such as, for instance, querying and visualisation), while a focus on technology development would concentrate, rather, on the fully automatic annotation of vast amounts of language data (including a qualitative evaluation of the created annotations and a validation of the created markup structures with regard to one or more schemas); this also includes the fully automated generation of language resources with the help of Language Resource factories. An alternative starting point is foreseen in Phase 1 for the specific situation when existing data or legacy data that might be years or even decades old, is to be processed and transformed into a sustainable language resource. The life cycle typically ends with the distribution and publication of a language resource; this includes making a resource available by download or, for example, as Linked Open Data through an API. However, the life cycle can continue if maintenance checks or user feedback result in minor updates or when a new version of an already existing resource needs to be prepared. Such an extension process would necessitate another iteration through the life cycle.

In addition to minor updates, there can be another reason for another iteration: in order to add a new annotation layer, i. e., to extend an existing language resource with new, additional analysis information by adding one or more levels of linguistic analysis or description. Such an extension or evolution of a language resource is also tightly intertwined with the concepts of recycling and reuse. Connected to this issue is the aspect of provenance of the added analysis information (which organisation added the information with the help of which tool or service and based on which scientific approach etc.) and properly documenting these aspects, i. e., the resource's processing and analysis history, in corresponding metadata descriptions. These must also include, for all annotation layers, notes on the legal situation and availability of the annotation layers (including the primary research data, of course).

This conceptualisation into different phases provides a generalised view. It is evident that different individual instantiations of actually preparing, curating and analysing a language resource call for slightly different organisations into phases and sub-phases. For example, persistent identification and data management (Phase 2) could also be in Phase 1 in some concrete life cycles; likewise, metadata-based documentation (Phase 5) could also be needed already in Phase 2.

This initial version of the Language Resource Life Cycle can be conceptualised as both a generic and idealised conceptualisation: working on different types of language resources and from different perspectives, having specific goals and objectives in mind, each with their own respective requirements, results in either slightly or significantly different organisations or instantiations of the life cycle in phases or sub-phases. For example, a resource does not necessarily have to be described with metadata in Phase 5 – this process could also be initiated in an earlier phase, however, it should be finished before packaging (Phase 6) and distributing a resource (Phase 7).

Furthermore, this is not the only life cycle to be considered when discussing the preparation of language resources because there is a close connection between the general scientific evolution of the whole field and the conceptualisation as depicted in the Language Resource Life Cycle. Similar life cycles exist with regard to the technological progression of basically all areas that relate to phases of the life cycle: there is research (and also more and more standardisation activities) around the topics of markup languages, metadata descriptions, querying engines, visualisations, IPR and copyright legislation and so on. If their individual pull forces – created through new versions of standards or software – are strong enough, these external factors are able to trigger a new life cycle iteration for a certain language resource: if a new major version of a tag-set has been released by the organisation that maintains it, this new version should, ideally, also be applied to all resources annotated using this tag-set so that they are not considered obsolete for making use of a deprecated annotation format.

Research and development activities on and around language resources cannot be discussed and evaluated in isolation. This is why the Language Resource Life Cycle also references, in the upper left hand corner of Figure 1, the highly heterogeneous context and situation outside the life cycle proper that exerts forces and provides factors to the actual initiation and implementation of the life cycle in concrete projects. Among these are the LR/LT ecosystem and the overall landscape of available resources and technologies for one or more languages. Highly relevant are gaps in terms of these resources and technologies as well as regional, national and international resource, technology and data creation strategies, initiatives and goals to fill these gaps and to foster research, innovation and development, often related to one or more specific languages.

4. Benefits of a Life-Cycle-Based Approach

The approach of making language resources publicly available for sharing purposes can be considered established best practice by now. This way it is possible, for example, to reduce or maybe share the costs needed to build a language resource or to invite the community to extend a language resource with one or more additional levels of linguistic analysis or by adding more annotated data. In that regard, the dynamic nature of language resources can correspond to the evolutionary nature of language but in order to be able to reflect these processes properly, a clear consensus needs to emerge in terms of describing, operationalising and maybe partially automating the different phases and changes a language resource undergoes. The author predicts that the clear need for such a general life cycle-based approach will, in the medium to long run, result in some kind of standard that will contribute to the important aspect of increased interoperability between language technologies and language resources. An important prerequisite for increased interoperability, sharing and also reproducibility of resources is some sort of mutual understanding regarding the question

how language resources are created, produced, described, annotated, evaluated, extended, distributed (including the concrete phases, sub-phases and names used to refer to these different phases).

5. Summary and Conclusions

This article presents a generic, common approach of a Language Resource Life Cycle, which conceptualises the different phases that involve creating, maintaining and distributing a language resource. According to the FLaReNet Strategic Language Resource Agenda and other recent reports, our community has a strong demand for a language resource life cycle and reference model, based on the observation that the "management of the life cycle of language resource creation has been largely overlooked in our community" (Calzolari et al., 2011) (p. 15).

Future work will include a more thorough specification of the Language Resource Life Cycle, including the preparation of an abstract and generic formal model of the life cycle (including a machine- and human-readable XML-based serialisation format) that takes all major metadata schemes used in the community into account. We will also take a closer look at which factors can cause a language resource slowly to decay or to become obsolete over the years and how to address these issues (for example, through standardised metadata and annotation formats, by making them highly visible and building up communities around language resources etc.).

Acknowledgments

The author would like to thank the two anonymous reviewers for their helpful comments. The author would also like to thank Bettina Kluge (University of Hildesheim, Germany) and Stelios Piperidis (R. C. Athena, Greece) for valuable comments on an earlier version of this paper.

Bibliographical References

Victoria Arranz et al., editors. (2010). *Proceedings of the LREC 2010 Workshop Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*, Valletta, Malta, May. `http://workshops.elda.org/lrslm2010/`.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*. O'Reilly. `http://www.nltk.org/book3/`.

Calzolari, N., Bel, N., Choukri, K., Mariani, J., Monachini, M., Odijk, J., Piperidis, S., Quochi, V., and Soria, C. (2011). Final FLaReNet Deliverable: Language Resources for the Future – The Future of Language Resources: The Strategic Language Resource Agenda, September. `http://www.flarenet.eu/sites/default/files/FLaReNet_Book.pdf`.

Lehmberg, T., Chiarcos, C., Rehm, G., and Witt, A. (2007). Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten. In Georg Rehm, et al., editors, *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen – Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*, pages 93–102. Gunter Narr, Tübingen.

Nicolas, L., Molinero, M. A., Sagot, B., Formoso, N. F., and Castro, V. V. (2010). Creating and maintaining language resources: the main guidelines of the Victoria project. In Arranz and van Eerten (Arranz and van Eerten, 2010), pages 6–9. `http://workshops.elda.org/lrslm2010/`.

Piperidis, S., Papageorgiou, H., Spurk, C., Rehm, G., Choukri, K., Hamon, O., Calzolari, N., del Gratta, R., Magnini, B., and Girardi, C. (2014). META-SHARE: One year after. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, May.

Rehm, G., Eckart, R., Chiarcos, C., and Dellert, J. (2008a). Ontology-Based XQuery'ing of XML-Encoded Language Resources on Multiple Annotation Layers. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco, May.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A., Symonenko, S., Tavosanis, M., and Vidulin, V. (2008b). Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco, May.

Rehm, G., Schonefeld, O., Witt, A., Lehmberg, T., Chiarcos, C., Bechara, H., Eishold, F., Evang, K., Leshtanska, M., Savkov, A., and Stark, M. (2008c). The Metadata-Database of a Next Generation Sustainability Web-Platform for Language Resources. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco, May.

Rehm, G., Schonefeld, O., Witt, A., Hinrichs, E., and Reis, M. (2009). Sustainability of Annotated Resources in Linguistics: A Web-Platform for Exploring, Querying and Distributing Linguistic Corpora and Other Resources. *Literary and Linguistic Computing*, 24(2):193–210. Selected papers from Digital Humanities 2008.

Rehm, G., Uszkoreit, H., Dagan, I., Goetcherian, V., Dogan, M. U., Mermer, C., Váradi, T., Kirchmeier-Andersen, S., Stickel, G., Jones, M. P., Oeter, S., and Gramstad, S. (2014). An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age". In *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014)*, Reykjavik, Iceland, May.

Schmidt, T., Chiarcos, C., Lehmberg, T., Rehm, G., Witt, A., and Hinrichs, E. (2006). Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources. In *Proceedings of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art*, East Lansing, Michigan, June.

van Veenendaal, R., van Eerten, L., Cucchiarini, C., and Spyns, P. (2013). The Dutch-Flemish HLT Agency: Managing the Lifecycle of STEVIN's Language Resources. In Peter Spyns et al., editors, *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, Theory and Applications of Natural Language Processing, pages 381–394. Springer.

Wittenburg, P., Ringersma, J., Trilsbeek, P., Elbers, W., and Broeder, D. (2010). Resource Lifecycle Management: Changing Cultures. In Arranz and van Eerten (Arranz and van Eerten, 2010), pages 14–18. `http://workshops.elda.org/lrslm2010/`.

Wörner, K., Witt, A., Rehm, G., and Dipper, S. (2006). Modelling Linguistic Data Structures. In B. Tommie Usdin, editor, *Proceedings of Extreme Markup Languages 2006*, Montréal, Canada, August.