The Language Bank of Finland

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Domain or subject-based repository, National repository system; including governmental

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

Brief Description of Repository

The Language Bank of Finland is a service for researchers in the digital humanities and social sciences using written, spoken or visual language resources. The Language Bank is coordinated by the national FIN-CLARIN consortium, an umbrella organization formed by Finnish universities, CSC – IT Center for Science, and the Institute for the Languages of Finland. FIN-CLARIN is a part of the international CLARIN ERIC research infrastructure. FIN-CLARIN forms the FIN-CLARIAH consortium together with DARIAH-FI.

The Language Bank has a wide variety of textual and audiovisual corpora and services for studying them. The corpora can be analyzed, processed and studied in the Language Bank or they can be downloaded. Many corpora are publicly accessible, some require log-in. The rights to use restricted resources can be applied for electronically. Using the Language Bank of Finland is free for researchers, teachers and students.

The Language Bank is legally represented by the University of Helsinki as the host of the FIN-CLARIN and FIN-CLARIAH consortia. The maintenance of the Language Bank is funded by the Ministry of Education and Culture. The development of the Language Bank is additionally partly funded by the Academy of Finland. The service is technically located at CSC, a non-profit, state-owned company administered by the Ministry of Education and Culture. CSC maintains and develops the centralized national IT infrastructure and uses it to provide nationwide IT services for research, libraries, archives, museums and culture as well as information, education and research management. CSC is responsible for the infrastructure, maintenance and security, as well as services used by the Language Bank, such as IDA. The University of Helsinki is responsible for the administration and content of the Language Bank.

The Language Bank Portal:
https://www.kielipankki.fi/language-bank/

Overview of the FIN-CLARIN organization:
https://www.kielipankki.fi/organization/

FIN-CLARIN introduction:
http://urn.fi/urn:nbn:fi:lb-201710211

Information about CSC:

https://www.csc.fi/en/csc

Corpora deposited in the Language Bank of Finland:
https://www.kielipankki.fi/corpora/

Web services provided by tools installed in the Language Bank of Finland:
https://www.kielipankki.fi/tools/

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

Brief Description of the Repository's Designated Community.

The Language Bank's users are mainly language researchers and students. The formation of the FIN-CLARIAH consortium in 2022 further broadens the user community to digital humanities and social sciences and more actively offer tools, text and speech data relevant for those communities. In line with this principle, the Language Bank accepts resource depositions as long as the data is stored as text or speech and the tools are usable at some stage of working with language data. For other data, like e.g. survey results, another repository, such as the Finnish Social Science Data Archive, may be more appropriate.

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

Level of Curation Performed. Select all relevant types from:

A. Content distributed as deposited, B. Basic curation – e.g. brief checking; addition of basic metadata or documentation, C. Enhanced curation – e.g. conversion to new formats; enhancement of documentation

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

Comments

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

Insource/Outsource Partners. If applicable, please list them.

The Faculty of Arts at the University of Helsinki (UHEL) carries the overall responsibility for the FIN-CLARIAH RI as the host of the Director and Vice Director. UHEL will in particular be responsible for curating, preparing and updating language-based materials into installable databases in the Language Bank according to CLARIN and DARIAH standards. Curating material includes informing the depositors and helping them to enforce formatting standards and licensing conventions for language tools and materials. In addition, UHEL will further develop or acquire tools for processing and automatically annotating language-based materials to be provided through the Language Bank and CSC. UHEL will also have coordinating responsibility for information dissemination for using tools and materials as well as FIN-CLARIAH researcher training through online courses and websites.

CSC provides and is responsible for the technical infrastructure of FIN-CLARIAH as well as providing technological expertise for developing the services. CSC’s role includes installation, maintenance, development and tailoring of the language-based resource services available for the Finnish research community, as well as their technical user support. CSC offers the computational resources, secure storage and other platforms for the Language Bank, installs new versions of databases and software as they become available, and assists in tailoring and developing processing and annotation tools and researcher workflows. CSC is responsible for operating the authentication and authorization infrastructure, allocates user space and provides access. CSC participates in the integration and harmonization of CLARIN and DARIAH services. In particular, CSC provides services as a CLARIN B Service Centre. CSC provides training for its services and assists UHEL in providing training for the Language Bank services to the FIN-CLARIAH members.

Other FIN-CLARIAH members develop tools and workflows as well as curate materials to be tailored to the CLARIN and DARIAH standards and centrally provided through the Language Bank and the CSC services. All FIN-CLARIAH members provide advice locally on how to use the FIN-CLARIAH services in their research and education.

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

Summary of Significant Changes Since Last Application (if applicable).

New supercomputers have been introduced at CSC. The research infrastructure services (RI) are offered through remote access, and the RI takes the necessary changes brought about by the growth in digitalization and data intensity into account by providing access to the new LUMI supercomputer environment offered by CSC for further processing of the accruing digital SSH big data material.

The Language Archive Tools (LAT) service was decommissioned. The documents and links have been included to match the new guidelines.

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

Other Relevant Information.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

ORGANIZATIONAL INFRASTRUCTURE

1. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The vision, mission and strategy of CSC – IT Center for Science form the basis of the Language Bank of Finland's operation. In addition, the Language Bank of Finland is a part of the national FIN-CLARIN consortium's strategy.

CSC's strategy covers data and preservation. The main targets of the CSC strategy:

1. To establish an internationally competitive ecosystem of scientific computing service of the whole Finnish research community.
2. To make digital data available and easy to use, securely, internationally, now and forever.
3. To establish an internationally recognized Finnish data analytics hub for research, education and public sector.
4. To make digital education and learning service into an interoperable ecosystem for all education levels.

FIN-CLARIN's strategy defines the Language Bank's service philosophy: The FIN-CLARIN consortium is the Finnish part of the European CLARIN collaboration building a research infrastructure for language-related resources in Humanities and Social Sciences. The goal of FIN-CLARIN is to provide access for all researchers in Finland to CLARIN compatible language resources. Likewise, the researchers in other countries will have access to the resources in Finland.

In order for the language resources to be found in scientific resource inventories, all resources must be equipped with compatible metadata descriptions. Their access policies must be stated clearly in a common format, and it should be easy to apply for and to grant permissions to use them. The resources must also be provided in a well-known or standardized format. In Finland, the related services will be located at the Language Bank of Finland. FIN-CLARIN is currently developing the assortment of materials available in the Language Bank as well as the instructions and support for using these resources.

CLARIN (Common Language Resources and Technology Infrastructure) is part of the ESFRI (European Strategy Forum on Research Infrastructures) roadmap. FIN-CLARIN is part of the Finnish national roadmap.

Finland is a full member of the European research infrastructure consortium CLARIN ERIC, whose task it is to ensure that the language processing services are coordinated and compatible in its member countries. Meanwhile, FIN-CLARIN has continued implementing the recommendations produced by CLARIN.

The Academy of Finland, one of the main funders of FIN-CLARIN, mentions The Language Bank of Finland as one of their recommended data management services in their Open data management plan.

A study conducted in 2020 by the Helsinki Institute for Social Sciences and Humanities measuring the use of and need for research infrastructures refers to both CLARIN and CSC multiple times.

The Language Bank of Finland's mission:
https://www.kielipankki.fi/organization/

CSC – IT Center for Science's strategy:
http://urn.fi/urn:nbn:fi:lb-2014120213
• CSC's strategy, consisting of mission, vision and values.

FIN-CLARIN's strategy:
http://urn.fi/urn:nbn:fi:lb-201710211
• Strategy, mission and vision of the FIN-CLARIN consortium.

The Language Bank of Finland's data management plan:
http://urn.fi/urn:nbn:fi:lb-201710257

The Academy of Finland's "Open data: data management plan":
http://urn.fi/urn:nbn:fi:lb-202103032

The Need for Research Infrastructures in the Social Sciences and Humanities at the University of Helsinki - Results of a survey for principal investigators
http://urn.fi/urn:nbn:fi:lb-202103031

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

2. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The Language Bank of Finland is not a legal entity; it is a service at CSC – IT Center for Science. The national FIN-CLARIN consortium coordinating the Language Bank is juridically represented by the University of Helsinki. Each deposited language resource is covered by a deposit agreement between FIN-CLARIN and the content provider. Each user is required to accept general terms and conditions as well as resource-specific licenses.

FIN-CLARIN ensures that publicly available content is accompanied by the appropriate licenses and agreements. By signing the agreement, the content provider asserts the resource's authenticity. FIN-CLARIN’s experts check the metadata provided by the content provider. FIN-CLARIN assumes ultimate responsibility for metadata correctness. The Language Bank produces the metadata entries in the META-SHARE service that are then approved and possibly amended by the content providers depositing the data.

The Language Bank of Finland uses predominantly CLARIN and Creative Commons licenses. By approving a license, the user agrees to follow the terms of use of the applicable resources. User access can be terminated or suspended by CSC without notice in the event of any unauthorized use of the services or if CSC has a justified reason to suspect that the services are used contrary to the terms and conditions.

Details on service access and availability are described in the Language Bank's Terms of use.

In addition to the general terms and conditions, many resources are associated with specific licenses. Access rights are only granted to applicants who have accepted all applicable terms and licenses. The licenses can be classified into three main groups: public (PUB), academic (ACA) and restricted (RES). Public resources can be accessed openly. Academic resources require the applicant to possess an academic status, such as researcher or student. Restricted resources require a personal permission granted by the content provider or their delegated contact person, who, in many cases, are the administrators of the Language Bank.

FIN-CLARIN experts aid content providers in choosing appropriate licenses for their language resources. There is also an online tool for assisting the process.

How to access language resources:
https://www.kielipankki.fi/access/

Terms of use of the Language Bank of Finland:
https://www.kielipankki.fi/language-bank/terms-of-use/

Information about agreements and licenses:
http://urn.fi/urn:nbn:fi:lb-2014120215

- Instructions for content providers for preparing language resources for publication.

1. Acquiring permissions from informants
2. Choosing the license class and end-user licenses
3. Deposition agreement

- What kinds of license classes and settings are available and how to choose the correct one. The main categories are public (PUB), academic (ACA), and restricted (RES).
- A list of end-user licenses available in the Language Bank's META-SHARE metadata service.

CLARIN license categories:
http://urn.fi/urn:nbn:fi:lb-2014120233

- CLARIN licenses explained in more detail.
- What kind of categories and traits are included in the different licenses that can be applied to CLARIN-compatible language resources in the Language Bank.
- PUB, ACA and RES licenses and the attributes they can be accompanied with.

General public (PUB) license:
http://urn.fi/urn:nbn:fi:lb-201802221

General academic (ACA) license:
http://urn.fi/urn:nbn:fi:lb-201802222

General restricted (RES) license:
http://urn.fi/urn:nbn:fi:lb-201802223

License selection helper tool:
http://urn.fi/urn:nbn:fi:lb-2014120237

- An application for identifying what kind of a license is suitable for a given language resource.

Instructions for creating language resources:

http://urn.fi/urn:nbn:fi:lb-2014120229

- Comprehensive instructions and a checklist for content providers for creating language resources.

1. Collecting the corpus
2. Scheduling and budgeting
3. Assistance from FIN-CLARIN
4. Preliminarily choosing a license class
5. Permissions required in the collecting stage
6. Finding a suitable deposition platform
7. Technical format and compatibility
8. Literation and annotation
9. Assembling and publishing metadata
10. Determining a suitable license
11. Deposition agreement with FIN-CLARIN
12. Transferring the finished language resource to the Language Bank

Deposition agreement templates:

http://urn.fi/urn:nbn:fi:lb-2021090221

META-SHARE:

http://metashare.csc.fi/

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

3. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase
accept

Reviewer 2

Comments:
3 – The repository is in the implementation phase
In the absence of a documented and shared succession plan that continues services including preservation this complies at level 3.

Response:

FIN-CLARIN became part of CLARIN ERIC through an Act of Parliament, committing Finland to permanently support CLARIN through the FIN-CLARIN consortium. Terminating the CLARIN membership would require another Act of Parliament, which would be highly exceptional and incur a one-year transition period. This is similar to Finland's membership in other organizations, e.g. CERN (European Organization for Nuclear Research) since 1991.

Therefore, no concrete plans have been made to relocate the present infrastructure off-premise. In-house options for data relocation are available at CSC, for example the Digital Preservation Solution for Research Data (PAS).

The dismantling of the RI requires that all datasets be transferred to long-term storage to be accessible for the verifiability of conducted research. Computing power and memory hardware can be reallocated and operational staff can be terminated or reassigned to other duties.

Each of the Language Bank of Finland's public services has a disaster recovery and business continuity plan. The plans are updated annually and approved by CSC's Head of Security. The documents comply with CSC's internal and external quality and security requirements.

The guaranteed preservation period of each language resource is defined by the resource's preservation category. The preservation categories and service levels are described in the the Language Bank's long-term preservation plan.

The Language Bank considers itself the primary custodian of the deposited data.

The Act of Parliament that made Finland a part of CLARIN ERIC:
http://urn.fi/urn:nbn:fi:lb-201802224

Digital Preservation Solution for Research Data (PAS):
http://urn.fi/urn:nbn:fi:lb-201802225

Long-term preservation plan of the Language Bank of Finland:

http://urn.fi/urn:nbn:fi:lb-202110101

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

4. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The Language Bank of Finland provides extensive information about the permissions required by resource deposition, including personal data considerations. This documentation is used as guidance for selecting the appropriate licensing conditions together with the the intellectual property holders. The CLARIN deposition license agreements (DELA) require that the data provider have the necessary rights to deposit the resource, including the right to publish data with disclosure risk. A breach of the DELA leads to termination of any agreements if no corrective action has been taken within 30 days.

Language resources with disclosure risk are published either using the CLARIN ACA license (requires an academic user to log in using their institutional credentials) or the CLARIN RES license (requires a personal permission). Such resources are thus not publicly available without at least personal user identification. The level of the disclosure risk is determined by the data owner. The Language Bank has procedures to encrypt data with disclosure risk. The Language Bank has no data anonymization service.

Resources containing personal data are furthermore labeled +PRIV, which means access can only be granted if the applicant presents reasonable grounds for using the resource. Access to resources with restricted content is handled via the Language Bank Rights system (LBR). LBR retains an electronic trail of the application and approval process, which includes the licenses approved by the user. A breach of a license lead to termination of access.

As a a general rule, intellectual property holders decide according to their own principles who is allowed to access their data in accordance to relevant data protection legislation. Each resource's IP holder has the right to handle the application process themselves via the Language Bank Rights system or to delegate the process to the Language Bank's administrators at CSC. Content providers can also log in to LBR to manage and monitor existing access rights concerning their resources.

Information about permission, agreements and licenses:
http://urn.fi/urn:nbn:fi:lb-2014120215
- The data collector is instructed to make sure to have appropriate permissions from the data subject(s) to publish the resource.
- The document also covers intellectual property rights, personal data considerations and ethical review boards as well as license classes and attributes. The main license classes are the CLARIN categories Public (PUB), Academic (ACA), and Restricted (RES), of which ACA and RES are the most relevant in this context.

Language Bank Rights:
https://lbr.csc.fi/
- The Language Bank's resource access application system.

How to access the language resources:
https://www.kielipankki.fi/access/

Encrypting data at the Language Bank:
https://www.kielipankki.fi/development/encryption/

CLARIN legal information:
http://urn.fi/urn:nbn:fi:lb-2014120216
- Definitions of licenses in the international CLARIN federation. When possible, corpora in the Language Bank are given licenses that are compatible with CLARIN in order to guarantee maximal interoperability.

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

5. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The host of FIN-CLARIAH is UHEL and the main RI computing facilities are hosted by CSC. The Faculty of Arts at UHEL currently carries the overall responsibility for the RI as the host of the Director and Vice Director of FIN-CLARIAH at the University of Helsinki. The FIN-CLARIAH consortium consists of two national RI components FIN-CLARIN and DARIAH-FI. The FIN-CLARIN members have in 2016 signed a Consortium Agreement. The DARIAH-FI members have signed letters of intent to participate in a consortium. The long-term commitment of the host organizations to developing the RI and maintaining its services is demonstrated by five permanent employees at UHEL and two at CSC, as well as through long-term funding for the Language Bank services by the Ministry of Education and Culture.

CSC – IT Center for Science is a core part of the Finnish national information technology infrastructure and a supercomputing facility. CSC was founded in 1971 and has about 500 employees (2022). It is a non-profit company that provides services for research, education, culture, public administration and enterprises.

CSC is funded by the Ministry of Education and Culture. The Language Bank's administration and maintenance is covered by the ministry base funding. Research and development of the Language Bank is additionally funded by the Academy of Finland. The Academy is presently funding FIN-CLARIAH by 4,6 million € in 2022–2023. This additional funding is typically renewed every two or three years.

CSC provides its staff with regular training and opportunities for the experts to build and maintain networks locally as well as internationally.

The two permanent core experts responsible for the Language Bank of Finland at CSC are language technologists with extensive experience in the field. At the University of Helsinki, there are five persons working primarily for the Language Bank. The total number of staff of the repository is seven (2022).

Information about CSC:
https://www.csc.fi/en/csc

Ministry of Education and Culture:
http://minedu.fi/en/

Academy of Finland:
http://www.aka.fi/en

The most recent funding decision of the Academy of Finland:
http://urn.fi/urn:nbn:fi:lb-2022011310

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

6. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The Language Bank of Finland has permanent experts both at CSC – IT Center for Science and University of Helsinki. At CSC, the experts are predominantly focused on technical maintenance and development of services and other resources. The University of Helsinki, in turn, has experts whose tasks include supporting researchers in producing resources, negotiating contracts with linguistic content providers, disseminating information about the Language Bank, teaching courses in using language resources, developing language tools, and producing content for the Language Bank Portal. All of the Language Bank's core services are operated by the Language Bank's internal experts. External experts are employed for the maintenance of certain specific tools such as Nimiarkisto (nimiarkisto.fi) and Signbank (signbank.csc.fi).

The Language Bank regularly visits Finnish universities and research groups on its "roadshow" tours. There are also courses, both physical and online, aimed at the Language Bank's users. FIN-CLARIN is one of the most active national bodies in the CLARIN federation in user involvement. The Language Bank has also participated in similar activities organized by other organizations.

Users can contact the Language Bank directly via e-mail or phone, both at CSC for technical issues related to the infrastructure and the University for questions about the content of the repository. The Language Bank Portal has a variety of manuals and instructions, mostly in Finnish but also partly in English.

Internationally, the CLARIN federation has knowledge centers around the world that also provide data, tools and knowledge.

The Language Bank Portal:
https://www.kielipankki.fi/language-bank/

Manuals and instructions at the Portal:
https://www.kielipankki.fi/tuki/
https://www.kielipankki.fi/support/

- Support and guidance documents for the corpora and tools at the Language Bank

CLARIN knowledge centers:

http://urn.fi/urn:nbn:fi:lb-201710213

- Information about the CLARIN Knowledge Sharing Infrastructure.

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

DIGITAL OBJECT MANAGEMENT

7. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

FIN-CLARIN ensures that publicly available content is accompanied by appropriate licenses and agreements. By signing the agreement, the content provider asserts the resource's authenticity. FIN-CLARIN’s experts check the metadata provided by the content provider. FIN-CLARIN assumes ultimate responsibility for metadata correctness. The Language Bank produces the metadata entries in the META-SHARE service that are then approved and possibly amended by the content providers depositing the data.

The Language Bank’s main services are the corpus interface Korp (korp.csc.fi), and the supercomputing cluster Puhti. Puhti also works as an application server. In the Language Bank's Download service (www.kielipankki.fi/download), language resources are stored in formats with inherent integrity checking (predominantly zip).

Every dataset is described in and linked to the Language Bank's metadata service META-SHARE (metashare.csc.fi). Changes in the data are logged in META-SHARE. Minor updates to the data that do not warrant a new version of the data set are logged in the metadata description of the data set, e.g. fixes to individual tokens or annotations. The original data is stored in the IDA research data storage system (ida.csc.fi). The Language Bank has a data life cycle management process that aims to strike a balance between persistently offering data for easy reproducibility of research and keeping flexibility to account for the need of adjustments, e.g. correcting metadata.

In all services, write access to data is tightly controlled. Shared services are developed using configuration management tools, such as Ansible.

CSC user guides:
http://urn.fi/urn:nbn:fi:lb-201503131
• General instructions for using the super cluster at CSC.

META-SHARE:
http://metashare.csc.fi/
• The Language Bank's metadata service.

IDA:
https://www.fairdata.fi/en/ida/
• The Finnish research data storage system, hosted by CSC – IT Center for Science.

Life cycle and metadata model of language resources:
http://urn.fi/urn:nbn:fi:lb-201710212
• Instructions how to manage versions of language resources in the Language Bank. Instructions for deciding whether a change in a file requires creating a new version or not.

Information about Ansible:
http://urn.fi/urn:nbn:fi:lb-201710263

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

8. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The Language Bank accepts and makes available resources with a natural language component in the data or the metadata that have been produced by researchers in Finland or by researchers of Finnish or Fenno-Ugric languages.

FIN-CLARIN, the national consortium coordinating the Language Bank of Finland, employs a corpus production line framework for preparing, depositing and enriching language resources. FIN-CLARIN's experts provide depositors with support and consultation during the different phases of a language resource's lifespan, including:

• preparing for collecting language data
• writing contracts with informants so that the resulting corpora will be as accessible as possible, and other legal advice
• tools for producing, maintaining and developing corpora
• file formats
• metadata
• collaboration between national and international bodies, projects and initiatives

The supported and recommended formats vary from service to service and are listed in the Language Bank Portal with more specific instructions. The Language Bank of Finland provides general instructions as well as platform-specific detailed ones. Besides the recommended formats, versions of the resources in other formats may additionally also be deposited, if considered advantageous. Suboptimal formats may also be accepted in case of particularly endangered material. When needed, files can also be converted.

The Language Bank has a three-tiered categorization of the service level provided for each deposited resource:

A. The resource is under active development. The Language Bank of Finland fixes any issues as soon as possible.
B. The resource is developed only upon user request. The Language Bank of Finland aims to fix issues concerning the resource, but external contribution may be required.
C. The resource is available ”as is”. The Language Bank of Finland does not fix nor develop the resource.

The national coordinator of FIN-CLARIN makes the ultimate decision to deposit a new resource and allocate its service level. The legal team for research support at the University of Helsinki participates in preparing new agreements and allocating appropriate license categories.

The Language Bank's users are mainly language researchers and students. The formation of the FIN-CLARIAH consortium in 2022 further broadens the user community to digital humanities and social sciences and more actively offer tools, text and speech data relevant for those communities. In line with this principle, the Language Bank accepts resource depositions as long as the data is stored as text or speech and the tools are usable at some stage of working with language data. For other data, like e.g. survey results, another repository, such as the Finnish Social Science Data Archive, may be more appropriate.

Instructions for creating language resources:
http://urn.fi/urn:nbn:fi:lb-2014120229
• Comprehensive instructions and a checklist for content providers for creating language resources.
1. Collecting the corpus
2. Scheduling and budgeting
3. Assistance from FIN-CLARIN
4. Preliminarily choosing a license class
5. Permissions required in the collecting stage
6. Finding a suitable deposition platform
7. Technical format and compatibility
8. Literation and annotation
9. Assembling and publishing metadata
10. Determining a suitable license
11. Deposition agreement with FIN-CLARIN
12. Transferring the finished language resource to the Language Bank

Overview of the corpus production line:
http://urn.fi/urn:nbn:fi:lb-201412022
• A visual representation of the life cycle of a corpus in the Language Bank of Finland. Produced by the University of

Helsinki for instructing existing and future users and content providers. The corpora produced according to the production line are deposited in the Language Bank's repositories, such as Korp. A more detailed and practical version of the instructions described in the instructions above.

Instructions for annotating corpora:
http://urn.fi/urn:nbn:fi:lb-201412023
• How to properly annotate text, audio and multimedia corpora. Information about relevant formats and software. These instructions aim at assuring the Language Bank's content's high quality and usability.

Instructions for content providers for selecting file formats:
http://urn.fi/urn:nbn:fi:lb-201412024
• What are the preferred file formats in the Language Bank. The requirements are service-specific. These instructions aim at assuring the files deposited in the Language Bank are of the highest possible quality and usability.

More detailed instructions about multimedia file formats:
http://urn.fi/urn:nbn:fi:lb-201412025
• How to convert audio and video files into the preferred formats. These instructions aim at assuring the files deposited in the Language Bank are of the highest possible quality and usability.

More detailed instructions concerning the Korp corpus interface's file formats:
http://urn.fi/urn:nbn:fi:lb-201412026
• The file format requirements for the Korp interface in more detail. These instructions aim at assuring the files deposited in Korp are of the highest possible quality and usability.

Instructions for transferring language resources to the Language Bank of Finland:
http://urn.fi/urn:nbn:fi:lb-201412027
• How to deliver a language resource to the Language Bank. These instructions assist the content providers in publishing their data.

The service levels of the Language Bank's corpora:
https://www.kielipankki.fi/corpora/

Finnish Social Science Data Archive:
http://www.fsd.uta.fi/en/

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

Accept

9. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The Language Bank is physically hosted at one location: CSC – IT Center for Science. There are detailed instructions for the language resource creation process, including the ingestion phase. Most of the Language Bank's data is publicly available, but data requiring authorization can also be distributed. Users can be authorized either via affiliation to an academic institution or via a personal application. Archive copies of data are stored in the IDA secure data service provided by CSC. The IDA service is funded by the Ministry of Education.

FIN-CLARIN, the federation coordinating the Language Bank, makes the deposition and long-term preservation decisions of the Language Bank. The technical branch of the Language Bank (CSC) is responsible for storing, operating and distributing the deposited language resources (corpora and software), except for resources and services maintained and operated by partners either on their own or CSC's virtual servers.

The servers of the Language Bank are backed up daily by CSC's backup team. The backup team has its own business continuity and disaster recovery plans. Backups are protected by a backup policy with a minimum of 21-day retention on disk and 90 day retention on tape. In order to verify the integrity of data transferred over the network, Cyclic Redundancy Check checksums (CRC) are generated at the respective clients before the data transfer. These checksums are verified with the CRC checksums generated by the backup agent, as soon as the data transfer is complete, and vice versa. This verification is done for all backup and restore operations for all data traveling to and from the backup agents.

Data recovery takes place on multiple levels: virtual machines are backed up as a whole and can be restored quickly, large filesystem partitions are backed up on the file level. The Language Bank administrators receive and review regular reports about the status of the backups. The backup team performs random file level restore tests on different platforms. Existing jobs on disk and the deduplication database are verified on a weekly basis. This ensures that existing backup jobs are restorable.

Instructions for language resource producers:
http://urn.fi/urn:nbn:fi:lb-201412021

- Instructions for the whole process of producing language resources, from gathering material to publishing.

The ingest process:
http://urn.fi/urn:nbn:fi:lb-201710253

- Instructions how to publish a language resource in the Language Bank.

Applying for access to the Language Bank's resources:
https://www.kielipankki.fi/support/access/

The IDA secure data service:
https://www.fairdata.fi/en/ida/

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

10. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

FIN-CLARIN is responsible for the long-term preservation of the deposited data. FIN-CLARIN has a data management plan that refers to individual plans of all member universities and other organizations. The plan covers the following topics:

- Division of responsibilities
- Data management principles, policies and guidelines
- Recommended and supported data management solutions
- Data management tools
- Data management training and instruction
- Internal data management
- Partner activities

Archive copies of data are distinguished from distribution copies. Archive copies contain the original deposited data or derived data that is particularly complicated to regenerate. FIN-CLARIN’s rights and responsibilities, as defined in the deposition license agreement, include the option to migrate data into new formats.

In order to mitigate the effect of file format obsolescence, depositing data in widely used, raw and open formats is encouraged. Keeping several formats in parallel may also be justified. The recommendations favor foreseeably longevous formats. The formats are monitored and requirements and recommendations regularly updated. Datasets curated by the Language Bank are migrated from one format to another when the original format has become obsolete. Up-to-date metadata also contributes to retaining data usability.

Each of the Language Bank of Finland's public services has a Disaster Recovery and Business Continuity Plan. The plans are updated annually and approved by CSC's Head of Security. The documents comply with CSC's internal and external quality and security requirements. These services include:

- Korp
■ concordance search service for text corpora
■ korp.csc.fi
- META-SHARE
■ metadata repository
■ metashare.csc.fi
- OAI-PMH service

■ https://kielipankki.fi/md_api/que?verb=Identify

• Persistent identifiers (PID)

■ persistent identifier system

■ uses the Universal Resource Name (URN) technology

■ pid.csc.fi

The systems are backed up daily, and backups are retained for three months. The backup process is internally documented. The backup team has its own business continuity and disaster recovery plans.

Long-term preservation plan of the Language Bank of Finland:
http://urn.fi/urn:nbn:fi:lb-202110101

FIN-CLARIN's data management policy:
http://urn.fi/urn:nbn:fi:lb-201710257

CLARIN deposition and end-user license agreements:
http://urn.fi/urn:nbn:fi:lb-2014120216

The preferred file formats in the Language Bank:
http://urn.fi/urn:nbn:fi:lb-201412024
• The requirements are service-specific. These instructions aim at assuring the files deposited in the Language Bank are of the highest possible quality and usability.

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

11. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The metadata of the language resources in the Language Bank of Finland are stored in the META-SHARE service. META-SHARE is an open, integrated, secure and interoperable sharing and exchange facility devoted to the sustainable sharing and dissemination of language resources and designed as a network of distributed repositories. META-SHARE has a metadata schema with obligatory fields that each entry must contain in order to be accepted. Some of the fields offer a set of fixed values. META-SHARE is developed by the META-NET consortium. University of Helsinki, CSC – IT Center for Science's partner in the Finnish FIN-CLARIN consortium, is a member of META-NET.

The Language Bank has no automatic deposition service. Each new resource is handled and verified manually. Metadata is entered and curated by the Language Bank staff in close collaboration with the data owners. Users can give feedback on the data and metadata via the Language Bank's communication channels provided in the Language Bank Portal.

Reference instructions are created for each corpus in the Language Bank. The instructions contain information about the intellectual property holder of the corpus, the year it was created, its name, and a persistent link to its metadata.

The Language Bank of Finland's META-SHARE node:
http://metashare.csc.fi
• The national META-SHARE node deployed in the Language Bank of Finland, serving as the central metadata repository of the Language Bank. CSC is responsible for the Finnish META-SHARE node's operation. The metadata is entered, curated and maintained in collaboration with FIN-CLARIN.

META-SHARE user manual:
http://urn.fi/urn:nbn:fi:lb-201412028
• How to use the current version of META-SHARE.

Reference instructions for the corpora deposited in the Language Bank:
https://www.kielipankki.fi/corpora/
• The quote links in the corpus table lead to the reference instructions of each corpus.

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

12. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The Language Bank of Finland predominantly collects language data of the languages of Finland as well as other language research material produced or enriched in Finnish universities and research institutions.

The data life cycle in the Language Bank is presented in FIN-CLARIN's corpus production line procedure. Phases of the life cycle (and the parties responsible for each phase):

• collecting linguistic material (data owner)

■ obtaining permissions from informants and IPR

■ obtaining a research permit

• finding a suitable deployment platform for the resource (data owner)

■ depends on data format, content and purpose

• depositing the resource (Language Bank)

■ signing an agreement between the content provider and the repository

■ creating and publishing metadata

■ assigning at least one persistent identifier for the resource

■ assigning an accessibility level and a license, if applicable

■ including the new resource in the access rights application system, if applicable

• publishing the resource (Language Bank)

■ managing access to the resource according to the license

• enriching the resource with additional depth of annotation (data owner, other researchers, or the Language Bank)

• publishing new versions (Language Bank)

Life cycle and metadata model of language resources:
http://urn.fi/urn:nbn:fi:lb-201710212
• Instructions how to manage versions of language resources in the Language Bank. Instructions for deciding whether a change in a file requires creating a new version or not.

Instructions for creating language resources:
http://urn.fi/urn:nbn:fi:lb-2014120229
• Comprehensive instructions and a checklist for content providers for creating language resources.

The Language Bank's corpus pipeline description:
http://urn.fi/urn:nbn:fi:lb-2022011311

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

13. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Each corpus in the Language Bank of Finland has a Universal Resource Name (URN). The URN system is provided by the National Library of Finland who has allocated the Language Bank its own namespace (urn:nbn:fi:lb). The URNs are created manually and harvested and published daily by the National Library of Finland.

The Language Bank is also interoperable with the Handle PID system. A round trip conversion principle exists between the two systems, so that either form of PID can always be derived from the other. CSC – IT Center for Science, the organization hosting the Language Bank, is also a member of the European Persistent Identifier Consortium (EPIC).

All resources, corpora and tools, in the Language Bank are listed in the Language Bank Portal. The corpus table is searchable, and each item is linked to the META-SHARE metadata repository. All resources are also searchable and browsable in the META-SHARE service. Metadata is also available in machine-readable form as stated in CLARIN's center requirements.

The citation information for each corpus is directly accessible via the Language Bank Portal's corpus list (https://www.kielipankki.fi/corpora/).

The data in META-SHARE is harvested via an OAI-PHM interface into the CLARIN federation's Virtual Language Observatory repository where, in turn, language resources all around the world are featured. In a scientifically wider context, information about the Language Bank's resources is also harvested into the Finnish Etsin research data finder.

The National Library of Finland's URN information:
http://urn.fi/URN:NBN:fi-fe2018093036991
• What is a uniform resource name (URN). Overview provided by the National Library of Finland, the provider of the URN service used at the Language Bank.

The Handle system:
http://www.handle.net
• The website of the commercial PID system used in the European Persistent Identifier Consortium (EPIC) with which the Language Bank's URN system has been made compatible.

European Persistent Identifier Consortium:

http://www.pidconsortium.eu

- The website of the PID consortium CSC is a member of.

The Language Bank Portal:

https://www.kielipankki.fi/language-bank

META-SHARE:

http://metashare.csc.fi

- The Language Bank's metadata repository.

CLARIN Level B center requirements:

http://hdl.handle.net/11372/DOC-78

- Requirements concerning machine-readable metadata (section 7).

CLARIN Virtual Language Observatory:

https://vlo.clarin.eu

- The CLARIN federation’s centralized metadata repository.

Etsin:

https://www.fairdata.fi/en/etsin/

- The Finnish multi-disciplinary research data finder.

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

14. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The metadata of the language resources in the Language Bank of Finland are stored in the META-SHARE service. FIN-CLARIN takes care of entering the metadata of new resources into META-SHARE as well as maintaining them up to date. Content providers can also create META-SHARE accounts and edit the metadata themselves. META-SHARE produces Dublin Core format metadata that is also enriched as CLARIN CMDI.

The Language Bank has a three-tiered categorization of the service level provided for each deposited resource. Service level A data is curated by the Language Bank and migrated to new formats when necessary. The service levels of each resource are listed in the corpus list (www.kielipankki.fi/corpora).

META-SHARE:
http://metashare.csc.fi/
• The Language Bank's metadata repository.

Instructions for creating and providing metadata:
http://urn.fi/urn:nbn:fi:lb-201412029
• How to compile and present metadata for language resources deposited in the Language Bank and what kind of assistance is available for content providers.

Information about CLARIN component metadata:
http://urn.fi/urn:nbn:fi:lb-2014120210
• Information about the metadata infrastructure of the international CLARIN consortium. This is relevant because the Language Bank wants to be compatible with other centers within the consortium.

Text corpus XML annotation specification:
http://urn.fi/urn:nbn:fi:lb-2014120211
• How to present metadata in XML files deposited in the Language Bank and instructions for properly encoding the corpora.

CLARIN deposition and end-user license agreements:

http://urn.fi/urn:nbn:fi:lb-2014120216

The service levels of the Language Bank's corpora:

https://www.kielipankki.fi/corpora/

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

TECHNOLOGY

15. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The Language Bank of Finland is certified as a level B infrastructure center by the CLARIN federation and adheres to the relevant standards outlined in the CLARIN Level B center requirements:

- Federated identity management (single sign-on)
- Support for PIDs (Handle)
- Metadata export in the CMDI format to the VLO via OAI-PMH
- TLS certificates for servers

The Language Bank uses community software wherever possible, including:

- Korp (korp.csc.fi) originally developed by SWE-CLARIN
- META-SHARE (metashare.csc.fi) developed by META-NET
- Helsinki Finite State Transducer (HFST) developed by the University of Helsinki
- GitHub for version control

The Language Bank of Finland's current infrastructure development project is funded by the Academy of Finland until the end of 2023. The core services of the Language Bank are funded by the Finnish Ministry of Education and CSC.

Most of the Language Bank's service portfolio is maintained on GitHub. Certain services, including the Language Bank Portal, are maintained using a private GitHub repository. Other service setups are already publicly visible, e.g. the Sanat lexicon database (sanat.csc.fi) and the Helsinki Finite State Transducer tools that are installed in CSC's supercomputing environment for the Language Bank's users.

All services of the Language Bank and the underlying infrastructure are continuously monitored and the administrators are alerted in case of abnormalities, such as high CPU load or memory usage.

The Language Bank's systems are documented externally in the Language Bank Portal, as well as in CSC's internal wiki that contains service descriptions and security documentation.

The Language Bank services are connected to the internet with a connection speed of 1 gigabit per second using the Finnish University Research Network, Funet.

The Language Bank has a policy for dealing with different versions and instances of the deposited language resources. For long term preservation, the Language Bank has a life cycle and metadata model of language resources. FIN-CLARIN has a data management plan that also refers to individual plans of all member universities and other organizations.

CSC is certified according to the standard ISO/IEC 27001:2013. The certification covers operations, development and management of CSC's ICT-platforms. The following functions related to the Language Bank are also covered:

- Data centers
- Networks
- Virtualization platform

- Storage and backup
- Operations systems
- Information security and physical security
- Change, incident and capacity management

CSC's ISO/IEC 27001:2013 certification does not cover application level functions of the Language Bank (e.g. the Korp user interface). The following OAIS functions are provided:

- Responsibility of CSC:
  - Access
  - Archival Storage
  - Administration
  - Data Management
- Responsibility of FIN-CLARIN:
  - Ingest

CSC is responsible for preserving and operating the language resources (corpora and software) deposited in the Language Bank, except for resources and services maintained and operated by partners either on their own or CSC's virtual servers.

CLARIN Level B center requirements:
http://hdl.handle.net/11372/DOC-78

The Language Bank's GitHub collaborative public repository:
http://urn.fi/urn:nbn:fi:lb-201710255

The Sanat lexicon database:
http://sanat.csc.fi/

Information about the Helsinki Finite State Transducer technology:
http://urn.fi/urn:nbn:fi:lb-20140730183

Information about the Funet network:
http://urn.fi/urn:nbn:fi:lb-201710256

Overview of the CSC computing environment:
http://urn.fi/urn:nbn:fi:lb-2014120218

- Information about the computing resources at CSC. This includes the computing environment of the Language Bank.

Overview of the CSC storage environment:

http://urn.fi/urn:nbn:fi:lb-2014120217

- Information about the data systems at CSC. This includes the data environment of the Language Bank.

About the ISO 27001 certification:

http://urn.fi/urn:nbn:fi:lb-201710252

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

16. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase
accept

Reviewer 2

Comments:
3 – The repository is in the implementation phase

Response:

The Language Bank has taken steps to minimize the risk of data loss due to technical or human error or even malice. CSC, the hosting institution of the Language Bank and consortium partner of FIN-CLARIN, is certified according to the ISO 27001 standard for information security. Application level development is tracked using version control (GitHub) and automated installation procedures. The Language Bank's business continuity and disaster recovery plans are reviewed annually by the Language Bank's experts and approved by CSC's security officer. All services run on regularly backed up virtual machines and can be restored quickly by CSC's specialists.

In addition to continuous monitoring, operating systems and the software stack are kept up to date and tested for vulnerabilities four times a year. The tests are performed both manually and automatically. This proactive approach helps to minimize attack vectors to the Language Bank's services.

Information about security at CSC:
http://urn.fi/urn:nbn:fi:lb-201710264

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept at level three noting that the evidence relates to the service aspects outsourced to CSC rather than those from the applicant themselves.

APPLICANT FEEDBACK

Comments/feedback

These Requirements are not seen as final, and we value your input to improve the CoreTrustSeal certification procedure. Any comments on the quality of the Requirements, their relevance to your organization, or any other contribution, will be considered as part of future iterations.

Response:

Reviewer Entry

Reviewer 1

Comments:
The applicant has provided clarification on the funding scheme. They have funding for 2022-2023 and typically renew it every 2-3 years. I would recommend approving this application.

Reviewer 2

Comments:
Thank you for your revised application. When you renew in three years we would suggest that you more clearly separate responsibilities between yourselves and outsource partners, including CSD. We would also suggest a greater degree of

detailed supporting evidence is made public.