Implementation of the CoreTrustSeal

The CoreTrustSeal board hereby confirms that the Trusted Digital repository The Language Bank of Finland complies with the guidelines version 2017-2019 set by the CoreTrustSeal Board.
The afore-mentioned repository has therefore acquired the CoreTrustSeal of 2016 on April 23, 2018.

The Trusted Digital repository is allowed to place an image of the CoreTrustSeal logo corresponding to the guidelines version date on their website. This image must link to this file which is hosted on the CoreTrustSeal website.

Yours sincerely,

The CoreTrustSeal Board

Assessment Information

Guidelines Version:	2017-2019 | November 10, 2016
Guidelines Information Booklet:	DSA-booklet_2017-2019.pdf
All Guidelines Documentation:	Documentation
Repository:	The Language Bank of Finland
Seal Acquiry Date:	Apr. 23, 2018
For the latest version of the awarded DSA for this repository please visit our website:	http://assessment.coretrustseal.org/seals/
Previously Acquired Seals:	Seal date: May 1, 2015 Guidelines version: 2014-2017 | July 19, 2013
This repository is owned by:	**CSC – IT Center for Science** Finland T +358 9 4572001 F +358 9 4572302 E contact@csc.fi W https://www.csc.fi/

Assessment

0. Context

Applicant Entry

Self-assessment statement:

Repository Type

National repository system

Brief Description of the Repository’s Designated Community

The Language Bank of Finland is a service for researchers in the humanities and social sciences using resources with at least some spoken or written language material. The service 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 enables access to the language resources to and by the whole CLARIN community. FIN-CLARIN collects and makes available tools and large-scale language resources through the Language Bank of Finland. Researchers and research groups can also deposit and distribute research material they have collected themselves through the Language Bank of Finland as well as make available tools they have developed as web services.

The Language Bank is legally represented by the University of Helsinki as the host of the FIN-CLARIN consortium. The development and upgrade of the service is partly funded by the Academy of Finland. The service is 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 Language Bank's infrastructure, maintenance and security, as well as services used by the Language Bank such as IDA.

The Language Bank has a wide variety of text and speech corpora and tools for studying them. The corpora can be analysed, processed and studied with the Language Bank tools 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 and students.

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/csc

Current FIN-CLARIN funding decision:

http://urn.fi/urn:nbn:lb-201804161

Corpora deposited in the Language Bank of Finland:

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

Tools installed in the Language Bank of Finland:

https://www.kielipankki.fi/tools/

Level of Curation Performed

C. Enhanced curation

Outsource Partners

n/a

Other Relevant Information

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

1. Mission/Scope

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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 whereas FIN-CLARIN's strategy defines the Language Bank's service philosophy. 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.

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.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

The evidence here does not show very clearly that the Language Bank of Finland deals with data preservation, data curation or data management. However, from the response in other requirements it becomes clear that the Language Bank of Finland actually has corpora that they try to maintain over the long-term (e.g. R7 on the "Download service (korp.csc.fi/download) with language resources that are stored in formats with inherent integrity checking" and e.g. R8 on preferred formats).

2. Licenses

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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 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.

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/

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).

- List of end-user licenses available in the Language Bank's META-SH?RE 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.

Graphical representation of license types:

http://urn.fi/urn:nbn:fi:lb-2014120238

- A chart 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

META-SHARE:

http://metashare.csc.fi/

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

Although the links are often in finnish. However, English links with the samen information do exist on the FIN-CLARIN webpage, for example https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/ClarinEULA#aca. When renewing the CoreTrustSeal certification, please refer to the English links.

3. Continuity of access

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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. In-house options for data relocation are available at CSC, for example the Digital Preservation Solution for Research Data (PAS).

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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

4. Confidentiality/Ethics

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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. Resources containing personal data are furthermore labeled +PRIV, which implies that 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/

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

Accept or send back to applicant for modification:

Accept

Comments:

5. Organizational infrastructure

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

The Language Bank of Finland is hosted by CSC – IT Center for Science, a core part of the Finnish national information technology infrastructure and a supercomputing facility. CSC was founded in 1971 and has about 320 employees (2018). 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. Research and development of the Language Bank is additionally funded by the Academy of Finland. Administration and maintenance is mainly covered by the ministry base funding. CSC provides its staff with regular training and opportunities for the experts to build and maintain networks locally as well as internationally. The core experts responsible for the Language Bank of Finland are language technologists with extensive experience of the field.

Information about CSC:

https://www.csc.fi/csc

Ministry of Education and Culture:

http://minedu.fi/en/

Academy of Finland:

http://www.aka.fi/en

Current FIN-CLARIN funding decision:

http://urn.fi/urn:nbn:lb-201804161

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

6. Expert guidance

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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.

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. In 2016, FIN-CLARIN was by far the most active national body in the CLARIN federation in user involvement. In 2018, the Language Bank also participated in the Digital History Research Method Workshop Tour organized by Aalto University.

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

Schedule of the 2016 Language Bank roadshow tour of Finnish universities and other organizations:

http://urn.fi/urn:nbn:fi:lb-201710261

CLARIN User Involvement Action Plan 2017 and overview of activities:

http://urn.fi/urn:nbn:fi:lb-201710262

- The number of UI activities by country (slide 14).

Digital History Research Method Workshop Tour:

http://urn.fi/urn:nbn:fi:lb-201802231

CLARIN knowledge centers:

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

- Information about the CLARIN Knowledge Sharing Infrastructure.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

7. Data integrity and authenticity

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

3. In progress: We are in the implementation phase.

Self-assessment statement:

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's main services are the corpus interfaces Language Archive Tools (LAT, lat.csc.fi) and Korp (korp.csc.fi), and the supercomputing cluster Taito. Taito also works as a application server (taito-shell). LAT supports MD5 checksums natively. The other services do not have consistent integrity checking at the moment. In the Language Bank's Download (korp.csc.fi/download) service, language resources are stored in formats with inherent integrity checking (e.g. 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. The original data is stored in the IDA research data storage system (ida.csc.fi).

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

Taito user guide:

http://urn.fi/urn:nbn:fi:lb-201503131

- General instructions for using the Taito super cluster at CSC.

META-SHARE:

http://metashare.csc.fi/

- The Language Bank's metadata service.

IDA:

http://openscience.fi/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

Accept or send back to applicant for modification:

Accept

Comments:

Acceptable for level 3 (implementation phase).

8. Appraisal

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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. 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.

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, including LAT and 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 selecting formats for content providers:

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 files:

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:

http://urn.fi/urn:nbn:fi:lb-201412026

- What are the 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.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

Accept. (Although most of the documents are still in Finnish, this reviewer expects increasing English documents step by step.)

9. Documented storage procedures

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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.

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://openscience.fi/ida

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

10. Preservation plan

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

3. In progress: We are in the implementation phase.

Self-assessment statement:

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 on a case-by-case basis. Archive copies contain the original data or data the regeneration of which is particularly complicated. FIN-CLARIN’s rights and responsibilities, as defined in the deposition license agreement, include the option to migrate data into new formats.

FIN-CLARIN is implicitly responsible for long-term preservation of the deposited data. A more explicit data preservation plan is being developed.

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:

Language Archive Tools (LAT)

- audiovisual corpus interface

- lat.csc.fi

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

- OAI-PMH service
- metalb.csc.fi

•

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.

FIN-CLARIN's data management plan:

http://urn.fi/urn:nbn:fi:lb-201710257

CLARIN deposition and end-user license agreements:

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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

Acceptable for level 3 (Inplementation phase). An elaborate preservation plan is still needed and, anyway, more documentation.

11. Data quality

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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 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. There are presently (2015) 29 META-SHARE nodes deployed worldwide, including the one at CSC.

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.

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.

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

Accept or send back to applicant for modification:

Accept

Comments:

12. Workflows

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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:

collecting linguistic material

- obtaining permissions from informants and IPR
- obtaining a research permit
-

finding a suitable deployment platform for the resource

- depends on data format, content and purpose

-

  depositing the resource

  - signing an agreement between the content provider and the repository
  - creating and publishing metadata
  - assigning at least one persistent identifier for the resource
  - choosing an accessibility level and a license, if applicable
  - including the the new resource in the access rights application system

-

  publishing the resource

- granting users access, if applicable
- using the resource in research and education

•

enriching the resource with

- additional depth of annotation
- publishing new versions

•

The Language Bank's users are mainly language researchers and students but the services are also available to other disciplines, especially other humanities and social sciences. 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 usable in some stage of working with language data. For some data, another repository, e.g. the Finnish Social Science Data Archive, may be more appropriate.

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.

Finnish Social Science Data Archive:

http://www.fsd.uta.fi/en/

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

13. Data discovery and identification

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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 predominantly created manually and stored on the server pid.csc.fi alongside their referent URLs. Based on the list, a Python script is used to generate an XML file that conforms to the National Library's specifications. The National Library harvests the contents of the XML file daily.

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 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 PIDs allocated by CSC:

http://urn.fi/urn:nbn:fi:lb-201802226

- A list containing each URN PID allocated by CSC for the Language Bank of Finland.

The National Library of Finland's URN information:

http://urn.fi/urn:nbn:fi:lb-2014120228

- What is a uniform resource name (URN). Instructions 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://etsin.avointiede.fi/en/

- The Finnish multi-disciplinary research data finder.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

14. Data reuse

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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. Formats are monitored and requirements and recommendations regularly updated. Data in obsolete formats is migrated into more current ones. Up-to-date metadata also contributes to retaining data usability.

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:

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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

15. Technical infrastructure

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

The Language Bank of Finland is certified as a level B infrastructure center by the CLARIN federation and adheres to the relevant standards set by CLARIN.

The Language Bank uses community software wherever possible, including:

- Korp (korp.csc.fi) originally developed by SWE-CLARIN
- Language Archive Tools (lat.csc.fi) developed originally by The Language Archive
- 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 2019.

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.

The Language Bank's systems are documented externally in the Language Bank Portal, as well as in an internal wiki.

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:

- Datacenters
- 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. The following OAIS functions are provided:

Responsibility of CSC:

- Access

- Archival Storage

- Administration

- Data Management

•

Responsibility of FIN-CLARIN:

- Ingest

•

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 ISO 27001 certification:

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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

16. Security

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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, is certified according to the ISO 27001 standard for information security. 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.

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

News about CSC's ISO certification renewal:

http://urn.fi/urn:nbn:fi:lb-201710251

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

17. Comments/feedback

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

1. No: We have not considered this yet.

Self-assessment statement:

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments: