KIELIPANKKI

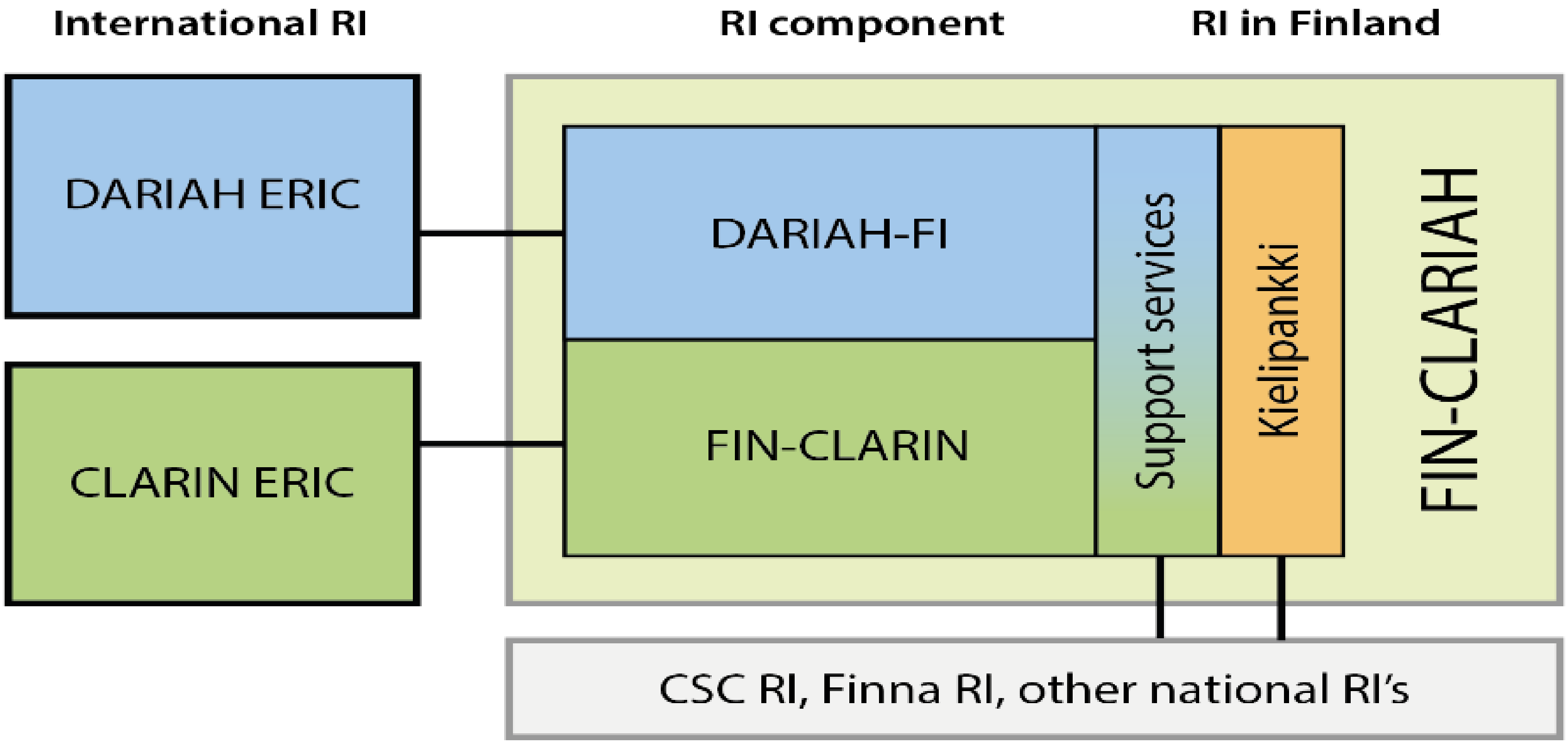
The Language Bank of Finland

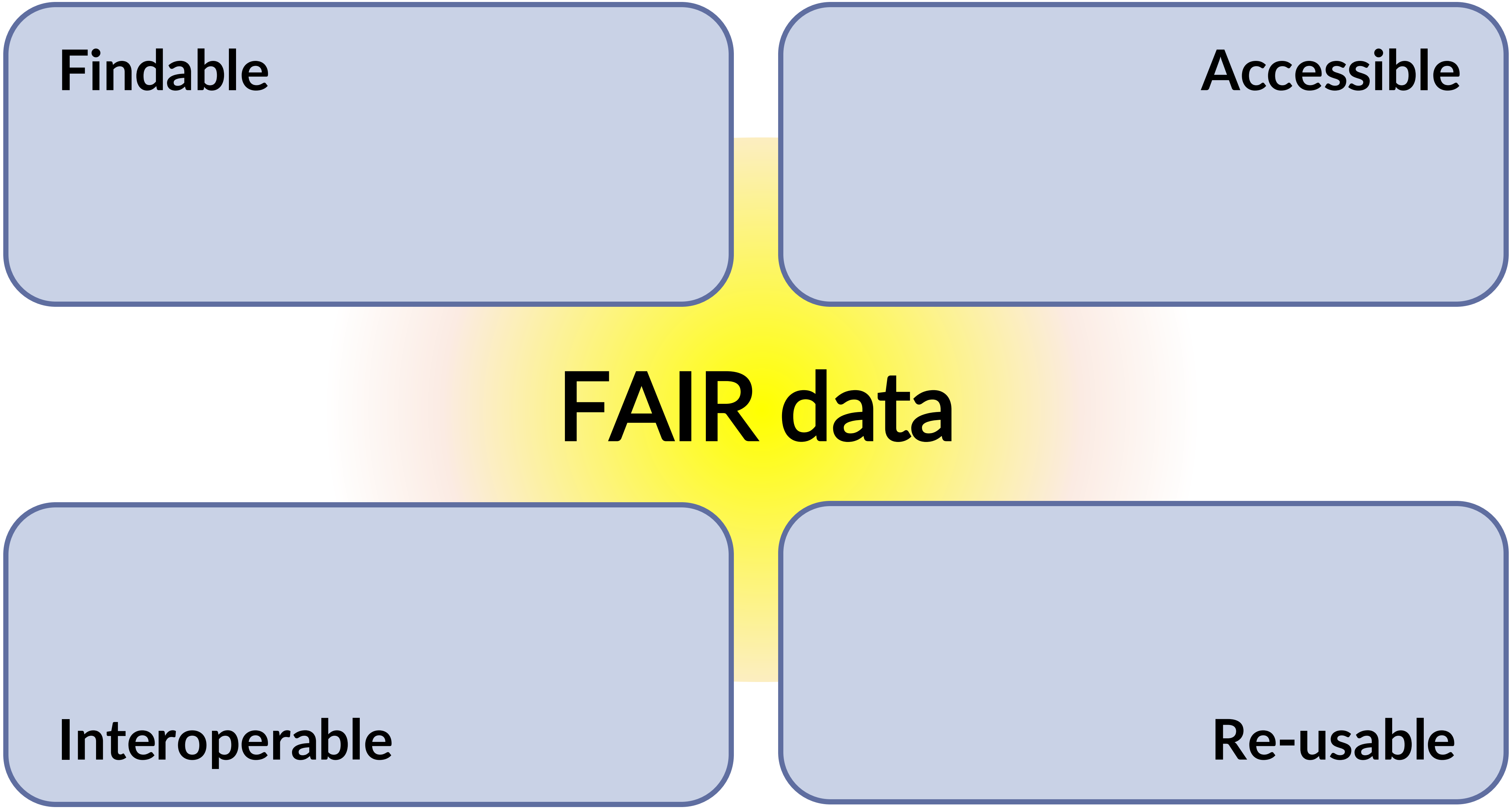
Agreements for the reuse of social media and interview data

Mietta Lennes

mietta.lennes@helsinki.fi

FIN-CLARIN


International RI
RI component
RI in Finland
DARIAH ERIC
CLARIN ERIC
DARIAH-FI
FIN-CLARIN
Support services
Kielipankki
FIN-CLARIAH
CSC RI, Finna RI, other national RI’s



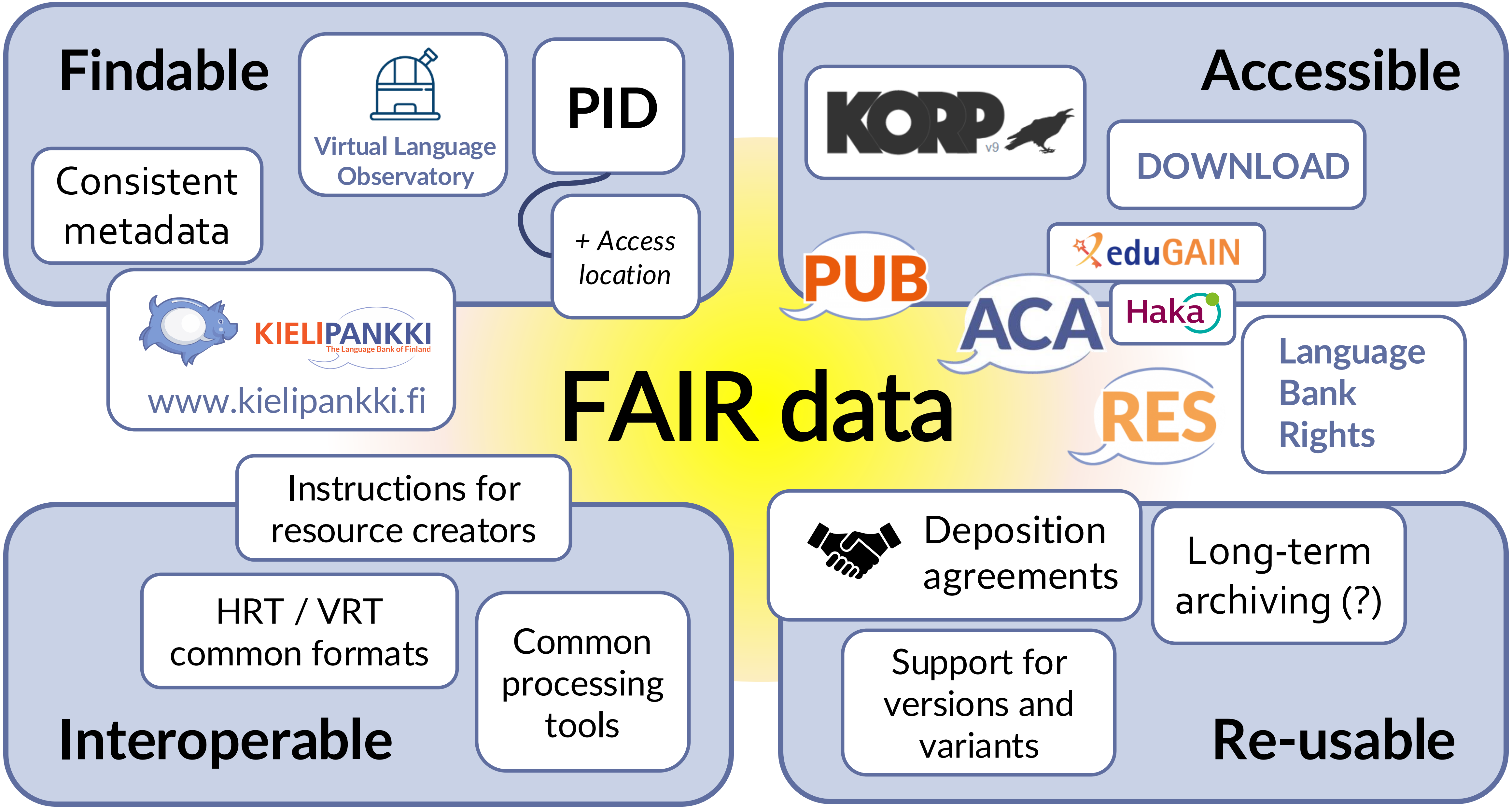
Findable

Accessible

FAIR data

Interoperable

Re-usable



FAIR data

Findable

Consistent metadata

Virtual Language Observatory

PID

+ *Access location*

KIELIPANKKI The Language Bank of Finland

www.kielipankki.fi

Accessible

KORP v9

DOWNLOAD

eduGAIN

Haka

PUB

ACA

RES

Language Bank Rights

Interoperable

Instructions for resource creators

HRT / VRT common formats

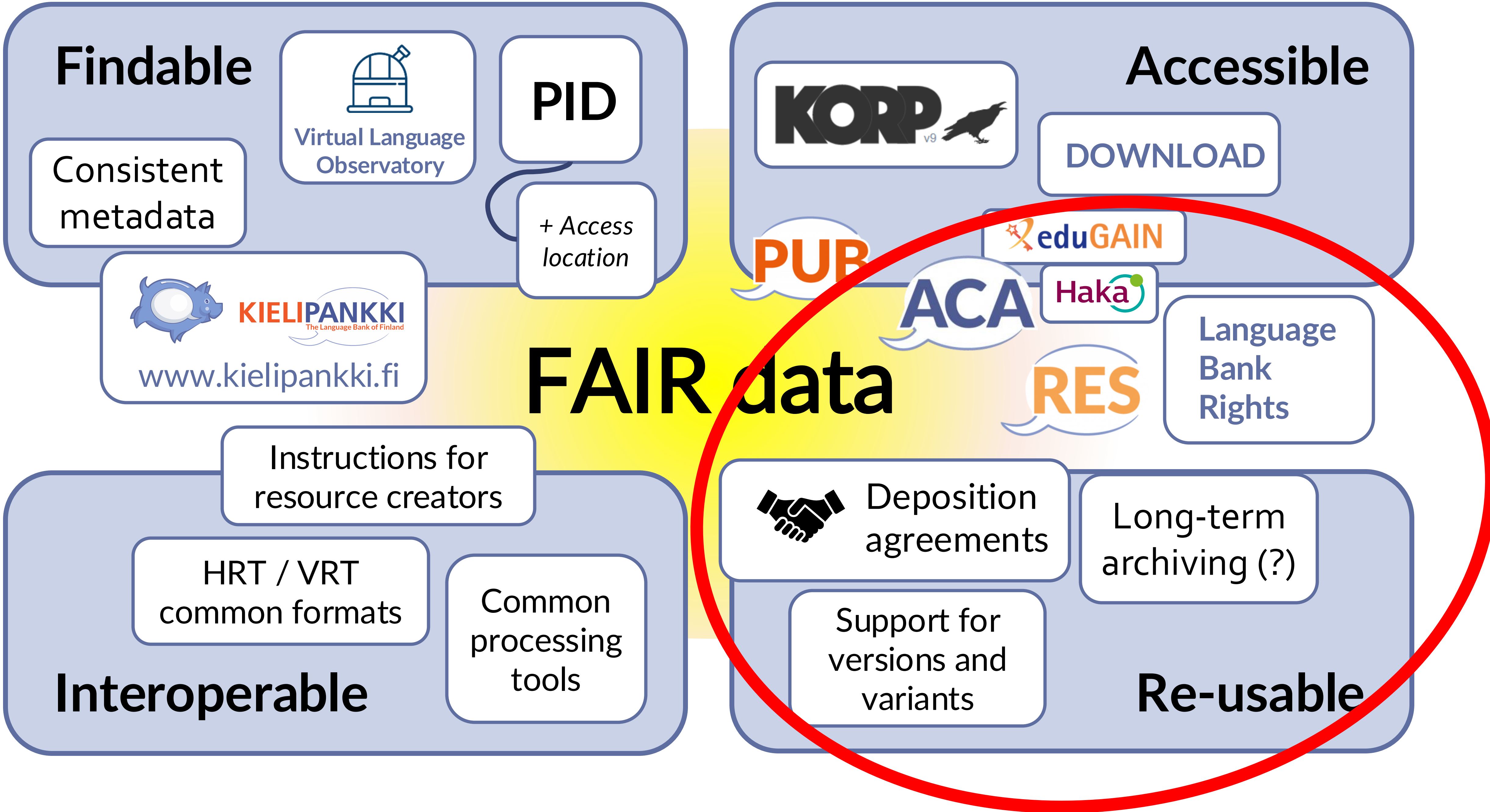
Common processing tools

Re-usable

Deposition agreements

Long-term archiving (?)

Support for versions and variants



FAIR data

Findable

Consistent metadata

Virtual Language Observatory

PID

+ Access location

KIELIPANKKI
The Language Bank of Finland
www.kielipankki.fi

Accessible

KORP v9

DOWNLOAD

PUB

eduGAIN

ACA

Haka

RES

Language Bank Rights

Interoperable

Instructions for resource creators

HRT / VRT common formats

Common processing tools

Re-usable

Deposition agreements

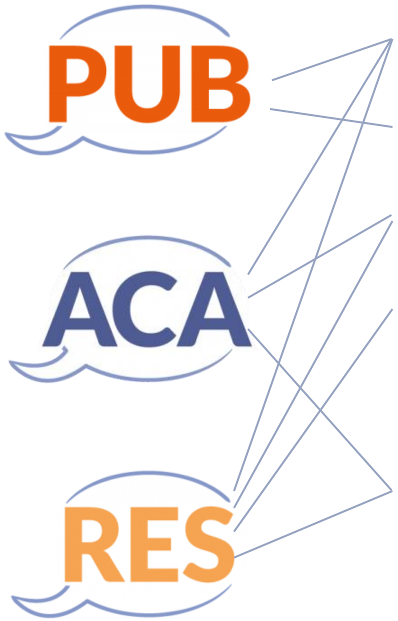
Long-term archiving (?)

Support for versions and variants

CLARIN license categories

PUB	Publicly available
ACA	Available for academic, logged in users
RES	Personal permission is required for access ***Language Bank Rights****, https://lbr.csc.fi*

More detailed license conditions



+BY	author must be cited
+NC	non-commercial use only
+ID	login is required
+PLAN	research plan is required
+PRIV	**contains personal data**
+NORED	redistribution is not allowed
+DEP	modified versions can be redistributed via CLARIN

and other resource-specific conditions, if required (e.g., data protection terms and conditions)

Processing data obtained via the Language Bank

- Read the license text and apply the appropriate protective measures when required.
- Follow the guidelines of your home organization.
- If processing personal data, submit the link to the public privacy notice of your research purpose.

https://www.kielipankki.fi/privacy-notices-from-users/

What kind of data can be deposited to the Language Bank of Finland?

- Text or speech in any natural language
- Make sure you have sufficient rights to distribute the data, at least for research purposes
  - Copyright and other Intellectual Property Rights
  - Personal data → information given to the data subjects! (Kielipankki brochure)
- Follow good practices in research ethics
  - In case ethical review is required, the process must be completed **prior to collecting the data**.

Intellectual Property Rights e.g., copyright and related rights

- Is your primary data **copyrighted**? (e.g., original texts, translations)
  - An explicit “copyright notice” is **not** required for the content to be protected by copyright!
  - Who might have rights to the content? Authors, translators, publishers… including, e.g., the companies who own social media platforms.
- Consider **related/neighboring rights**, e.g.:
  - Performers’ rights
  - Databases (*sui generis*)
  - Photographs

Ask for permission from the rightholder(s)?

- Make sure you are allowed to share the (copyrighted) material at least for research (& education) purposes, via Kielipankki.
- Social media platforms are legally challenging data sources
- If it is not possible to ask for permission, it *may* in *some cases* be possible to apply the “data mining exception” (Tekijänoikeuslaki 13 b §).

Personal data

- **Personal data** is any information that can be used for identifying an existing living natural person.
- **Processing personal data** is almost any conceivable action on the personal data, including the action of just keeping or storing the data.

Personal data in special categories

- Special categories:
  - Information indicating race or ethnic origin, political opinions, religious or a philosophical belief or union membership
  - Genetic information
  - Biometric data for unambiguous identification of a person
  - Health information
  - Information on sexual behavior and orientation
  - Information related to criminal convictions and violations or security measures

- Processing is only allowed for **scientific research**, to fulfill statutory obligations, or with the consent of the data subject:
  - Process only if absolutely necessary!
  - Usually **requires more stringent protective measures**.

Licensing a resource via the Language Bank

1. Deposition license agreement (DELA)

https://www.kielipankki.fi/support/dela/

- The depositor gives the Language Bank (formally, the University of Helsinki) the right to distribute the resource to end-users under specific terms and conditions.
- A DELA is usually required, unless the University of Helsinki is the rightholder or the resource is already available under a public license.
  - The DELA can also include a data processing agreement between the depositor's home university and the University of Helsinki.
  - Social media datasets are considered on a case by case basis.

2. End-User License Agreement (EULA)

- The End-User agrees to use the data under the general conditions of the Language Bank as well as under resource-specific conditions (according to the permission from the rightholder and/or the information given to the data subjects).
- For resources containing personal data, the end-user license also includes data protection terms and conditions. Example license: Donate Speech Corpus

How to prepare for depositing a resource with the Language Bank

If the resource contains personal data…

- **Follow the data protection guidelines of your home organization!**
- Have your documentation ready:
  - Ethical review statement, if required (see TENK)
  - Information provided to the research subjects (e.g., privacy notice or similar, Kielipankki brochure). See FSD: Data Management Guide
  - The specific wording of the consent, if applicable (i.e., to what the subjects have agreed).
- If required, review your preliminary evaluation of risks, considering the potential storage and distribution via the Language Bank.
- If necessary, perform a Data Protection Impact Assessment (DPIA).

Minimize the amount of personal data

- Data minimization principle: process no more than what you need for your purpose but **keep the data you need**!
- Pseudonymize or anonymize, if necessary and possible.
  - Read more in

    Finnish Social Science Data Archive (FSD):
    Data Management Guidelines
    https://www.fsd.tuni.fi/en/services/data-management-guidelines/

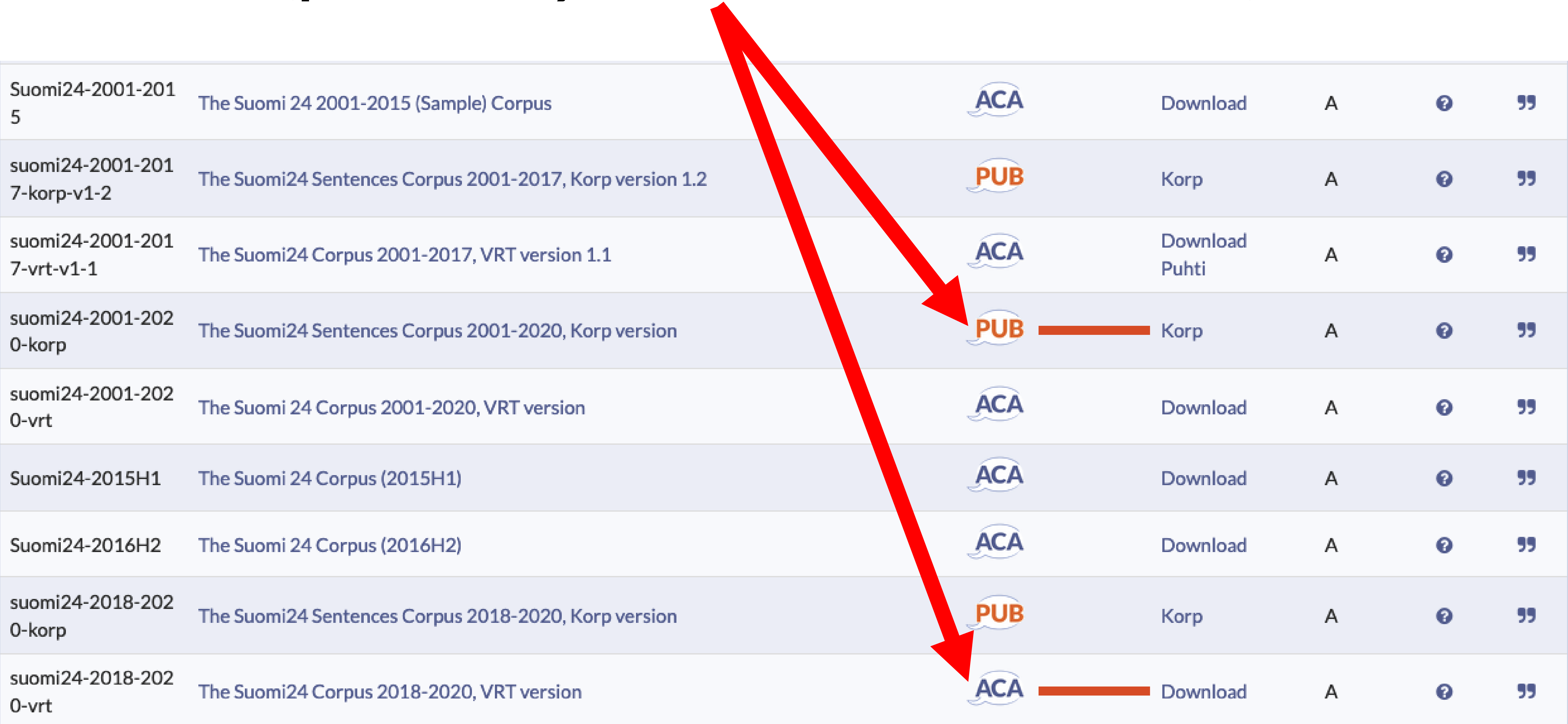
If the resource contains copyrighted data...

- You should have obtained the material legally.
- Keep a list or a description of the copyrighted works and their authors and other rightholders
- Can you contact the rightholders for their permission?
- Consider the possibility of pre-processing the data: is it still useful for research if the content is scrambled?
- **Follow the instructions given by your home organization!**

Are safety measures required for sharing?

- The Language Bank of Finland offers means of protection for restricted content:
  1. Access management, if needed: university login (ACA); access granted on an individual basis upon application (RES)
  2. Resource-specific data protection terms and conditions (must be accepted by the end-users)
  3. Data encryption for individuals (example resource: findarc)
  4. Sensitive Data (SD) services at CSC

Resources may have several versions or variants (potentially under different licenses!)



Suomi24-2001-2015	The Suomi 24 2001-2015 (Sample) Corpus	ACA	Download	A	?	❞
suomi24-2001-2017-korp-v1-2	The Suomi24 Sentences Corpus 2001-2017, Korp version 1.2	PUB	Korp	A	?	❞
suomi24-2001-2017-vrt-v1-1	The Suomi24 Corpus 2001-2017, VRT version 1.1	ACA	Download Puhti	A	?	❞
suomi24-2001-2020-korp	The Suomi24 Sentences Corpus 2001-2020, Korp version	PUB	Korp	A	?	❞
suomi24-2001-2020-vrt	The Suomi 24 Corpus 2001-2020, VRT version	ACA	Download	A	?	❞
Suomi24-2015H1	The Suomi 24 Corpus (2015H1)	ACA	Download	A	?	❞
Suomi24-2016H2	The Suomi 24 Corpus (2016H2)	ACA	Download	A	?	❞
suomi24-2018-2020-korp	The Suomi24 Sentences Corpus 2018-2020, Korp version	PUB	Korp	A	?	❞
suomi24-2018-2020-vrt	The Suomi24 Corpus 2018-2020, VRT version	ACA	Download	A	?	❞

Solutions

Example resource:
"CoLaGe" – a speech resource containing interviews recorded in Spain and Mexico

The resource contains personal data. The interviewees have been informed about the restricted distribution of their data for research purposes.

1. The Language Bank publishes the basic metadata provided by the PI of the project, mints a Persistent Identifier (PID) for the resource and offers citation instructions.
2. The researcher and the Language Bank (University of Helsinki) negotiate a deposition agreement. For data protection reasons, the license will be restricted to research purposes and requires an individual application.
3. The researcher uploads the data to one of the experts in the Language Bank.
4. The Language Bank repackages the files, adds the *readme.txt* and *LICENSE.txt* documents and publishes the data in the download service.

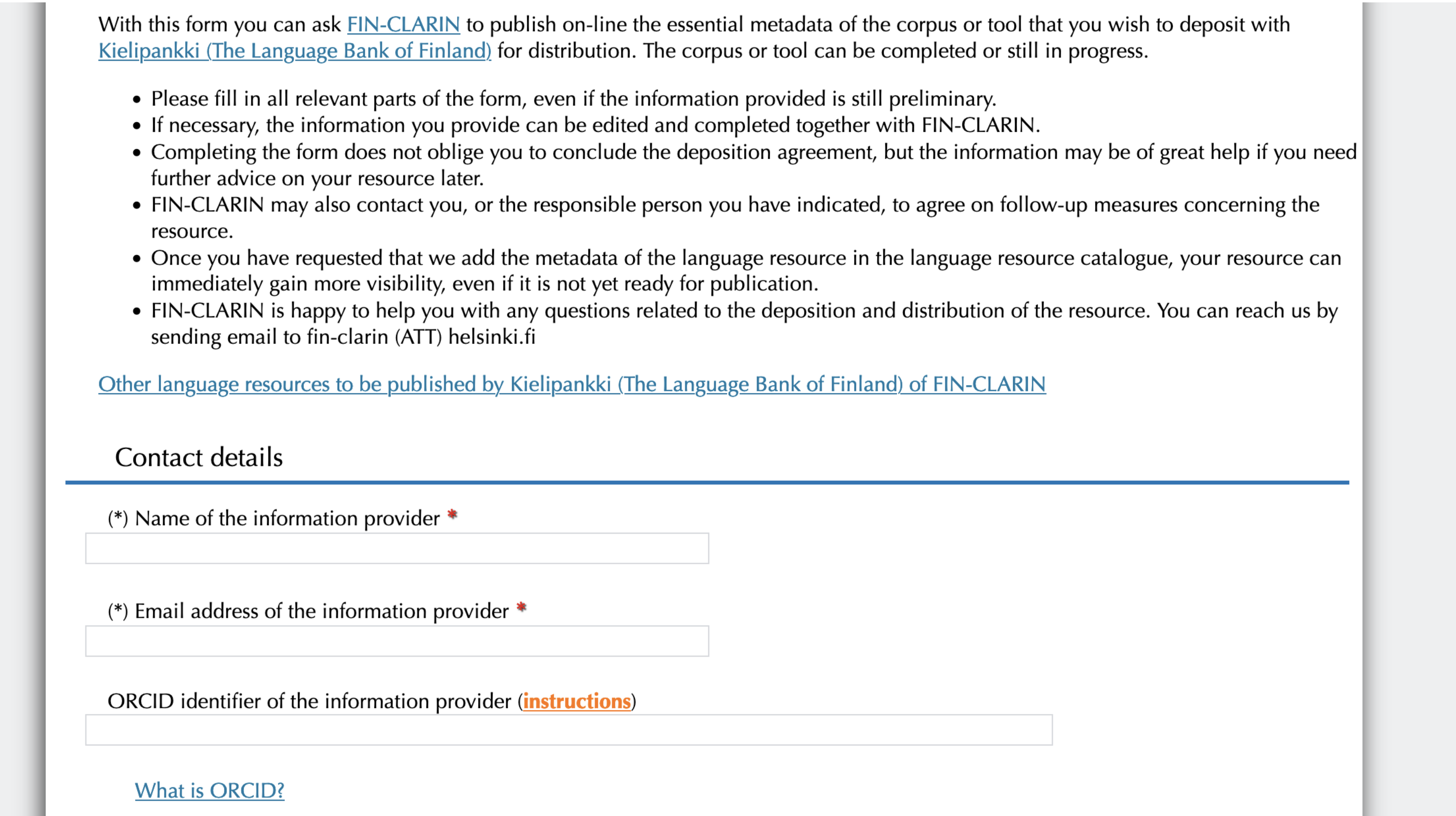
Example resource: “Somepressa24”

The corpus includes social media posts published in six different social media services (Twitter/X, Facebook, Instagram, TikTok, YouTube) and comments left under news items from Helsingin Sanomat and YLE.fi, related to the presidential elections in 2024.

1. The Language Bank publishes the basic metadata provided by the PI of the project, mints a Persistent Identifier (PID) for the resource and offers citation instructions.
2. The researchers and the Language Bank (University of Helsinki) negotiate on the license terms. Due to the unclear and risky situation with social media content, a great deal of legal expertise is required. Due to copyrighted content and for data protection, the license will be restricted to specific research purposes and an individual application will be required. Applications may need to be reviewed by a legal advisor at UHEL.
3. The researchers are preparing the dataset and will deliver it to the Language Bank.
4. The Language Bank will repackage the files, add the *readme.txt* and *LICENSE.txt* documents and publish the data for download and (hopefully) at a later point via the Korp system.

Suggest a dataset to the Language Bank of Finland!

http://urn.fi/urn:nbn:fi:lb-2021121422



With this form you can ask FIN-CLARIN to publish on-line the essential metadata of the corpus or tool that you wish to deposit with Kielipankki (The Language Bank of Finland) for distribution. The corpus or tool can be completed or still in progress.

- Please fill in all relevant parts of the form, even if the information provided is still preliminary.
- If necessary, the information you provide can be edited and completed together with FIN-CLARIN.
- Completing the form does not oblige you to conclude the deposition agreement, but the information may be of great help if you need further advice on your resource later.
- FIN-CLARIN may also contact you, or the responsible person you have indicated, to agree on follow-up measures concerning the resource.
- Once you have requested that we add the metadata of the language resource in the language resource catalogue, your resource can immediately gain more visibility, even if it is not yet ready for publication.
- FIN-CLARIN is happy to help you with any questions related to the deposition and distribution of the resource. You can reach us by sending email to fin-clarin (ATT) helsinki.fi

Other language resources to be published by Kielipankki (The Language Bank of Finland) of FIN-CLARIN

Contact details

(*) Name of the information provider *

(*) Email address of the information provider *

ORCID identifier of the information provider (instructions)

What is ORCID?

Sensitive Data services at CSC

Quick instructions by the Language Bank of Finland:

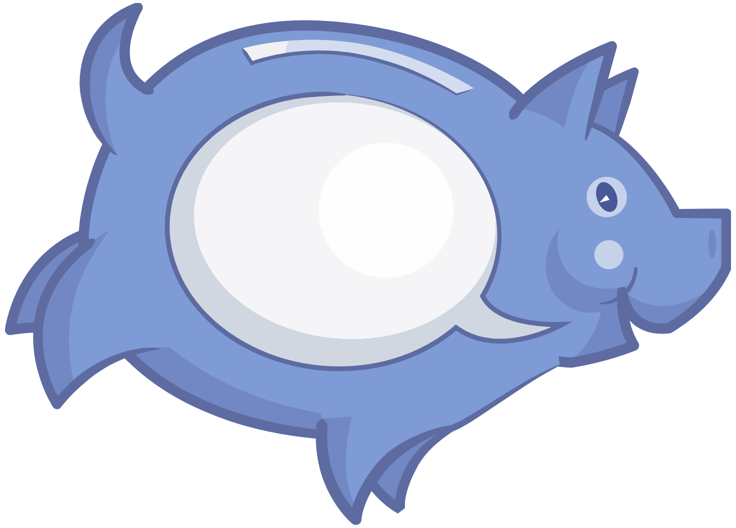
https://www.kielipankki.fi/support/sd-services/

https://www.kielipankki.fi/tuki/sd-services/

For up-to-date details, see documentation provided by CSC:

https://docs.csc.fi/data/sensitive-data/

KIELIPANKKI
The Language Bank of Finland



Kiitos! Tack! Thank you!

www.kielipankki.fi

General support
fin-clarin@helsinki.fi

Technical support
kielipankki@csc.fi

Time for questions and discussion!