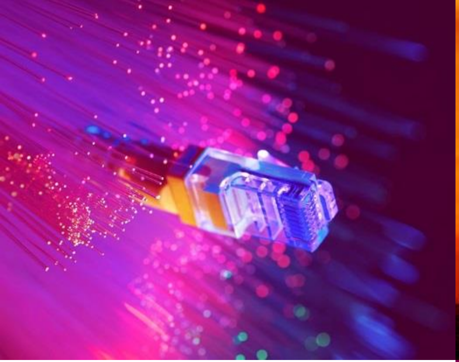
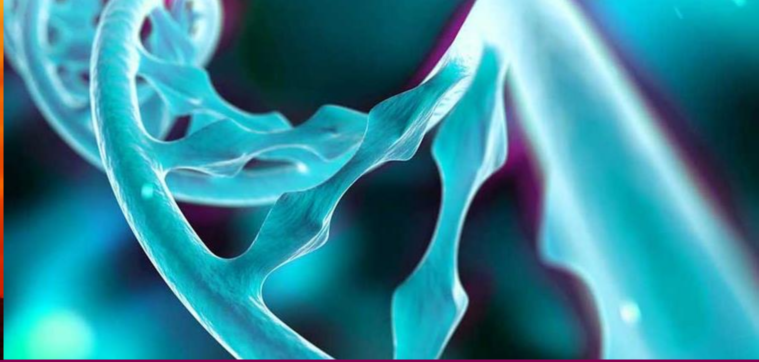




Access Management at Kielipankki – The Language Bank of Finland

Martin Matthiesen

DELAD Workshop Cork, 16.11.2017

CSC – Suomalainen tutkimuksen, koulutuksen, kulttuurin ja julkishallinnon ICT-osaamiskeskus

Contents

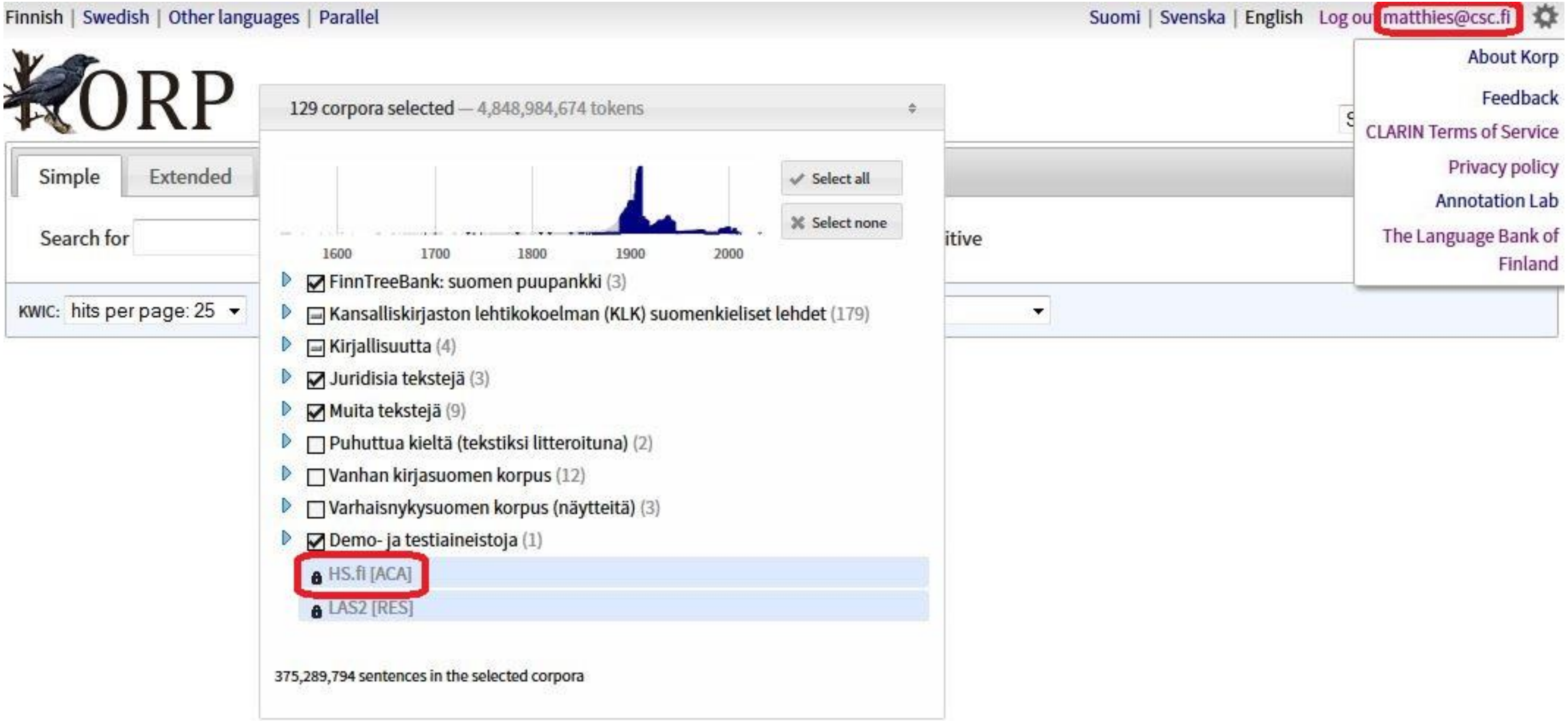
- CSC?
- A use case example: HS.fi
- Registration of new resources
- Application and Approval Process
- Time for questions

CSC – IT Center for Science

- Non-profit, (mostly) publicly funded company
- Owned by the Finnish ministry of education
- Supporting research & education institutions
  - “We don’t do research, we support research.”
  - Network (FUNET)
  - High-performance super computing
  - Data services
- Home to E-Infrastructures like Elixir, Kielipankki

The Use Case HS.fi

- **Helsingin Sanomat newspaper corpus**
  9/2011-9/2012
- 94 000 articles
- 600,000 sentences
- 8 Million words
- Licensed with the **CLARIN ACA +NC license**
- Available at https://korp.csc.fi/



Finnish | Swedish | Other languages | Parallel

Suomi | Svenska | English Log ou matthies@csc.fi

About Korp

Feedback

CLARIN Terms of Service

Privacy policy

Annotation Lab

The Language Bank of Finland

KORP

Simple Extended

Search for

KWIC: hits per page: 25

129 corpora selected — 4,848,984,674 tokens

- ☑ FinnTreeBank: suomen puupankki (3)
- ⊟ Kansalliskirjaston lehtikokoelman (KLK) suomenkieliset lehdet (179)
- ⊟ Kirjallisuutta (4)
- ☑ Juridisia tekstejä (3)
- ☑ Muita tekstejä (9)
- ☐ Puhuttua kieltä (tekstiksi litteroituna) (2)
- ☐ Vanhan kirjasuomen korpus (12)
- ☐ Varhaisnykysuomen korpus (näytteitä) (3)
- ☑ Demo- ja testiaineistoja (1)
- 🔒 HS.fi [ACA]
- 🔒 LAS2 [RES]

375,289,794 sentences in the selected corpora

Ingredients

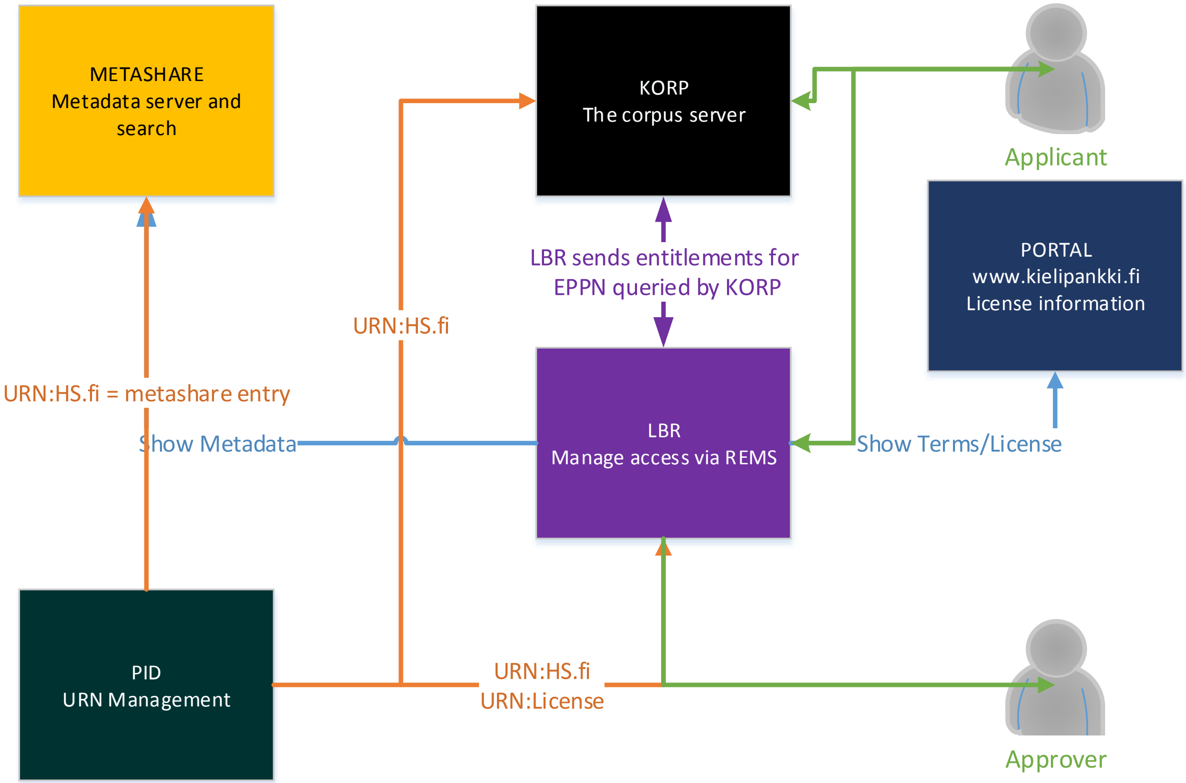
We need: The data, a contract and

lbr.csc.fi (LBR)	**“Language Bank Rights” Access Management**
korp.csc.fi (KORP)	The corpus viewer
metashare.csc.fi (METASHARE)	Metadata management
pid.csc.fi (PID)	PID management (URN/Handle)
www.kielipankki.fi (PORTAL)	License information

Prepare registration of corpus

- Create metadata in METASHARE
- Assign Persistent Identifier, like a URN/Handle (here: URN:HS.fi) in PID
- License text available in PORTAL
- Register corpus in LBR

Dependencies



METASHARE
Metadata server and search

KORP
The corpus server

Applicant

PORTAL
www.kielipankki.fi
License information

LBR sends entitlements for EPPN queried by KORP

URN:HS.fi

URN:HS.fi = metashare entry

Show Metadata

LBR
Manage access via REMS

Show Terms/License

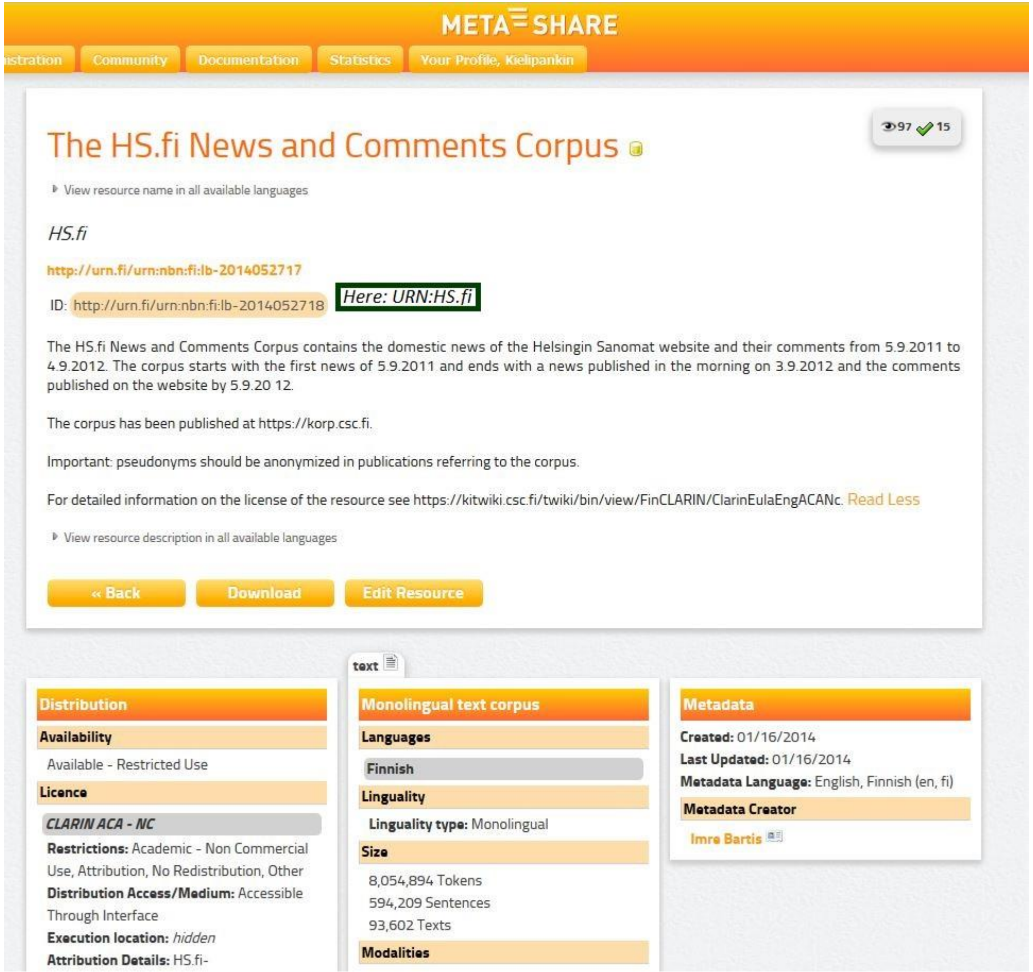
PID
URN Management

URN:HS.fi
URN:License

Approver

Create metadata in METASHARE

Register a URN in PID ("URN:HS.fi") to point to this page



The HS.fi News and Comments Corpus

▸ View resource name in all available languages

HS.fi

http://urn.fi/urn:nbn:fi:lb-2014052717

ID: http://urn.fi/urn:nbn:fi:lb-2014052718 *Here: URN:HS.fi*

The HS.fi News and Comments Corpus contains the domestic news of the Helsingin Sanomat website and their comments from 5.9.2011 to 4.9.2012. The corpus starts with the first news of 5.9.2011 and ends with a news published in the morning on 3.9.2012 and the comments published on the website by 5.9.20 12.

The corpus has been published at https://korp.csc.fi.

Important: pseudonyms should be anonymized in publications referring to the corpus.

For detailed information on the license of the resource see https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/ClarinEulaEngACANc. Read Less

▸ View resource description in all available languages

Distribution

Availability

Available - Restricted Use

Licence

CLARIN ACA - NC

Restrictions: Academic - Non Commercial Use, Attribution, No Redistribution, Other

Distribution Access/Medium: Accessible Through Interface

Execution location: *hidden*

Attribution Details: HS.fi-

Monolingual text corpus

Languages

Finnish

Linguality

Linguality type: Monolingual

Size

8,054,894 Tokens

594,209 Sentences

93,602 Texts

Modalities

Metadata

Created: 01/16/2014

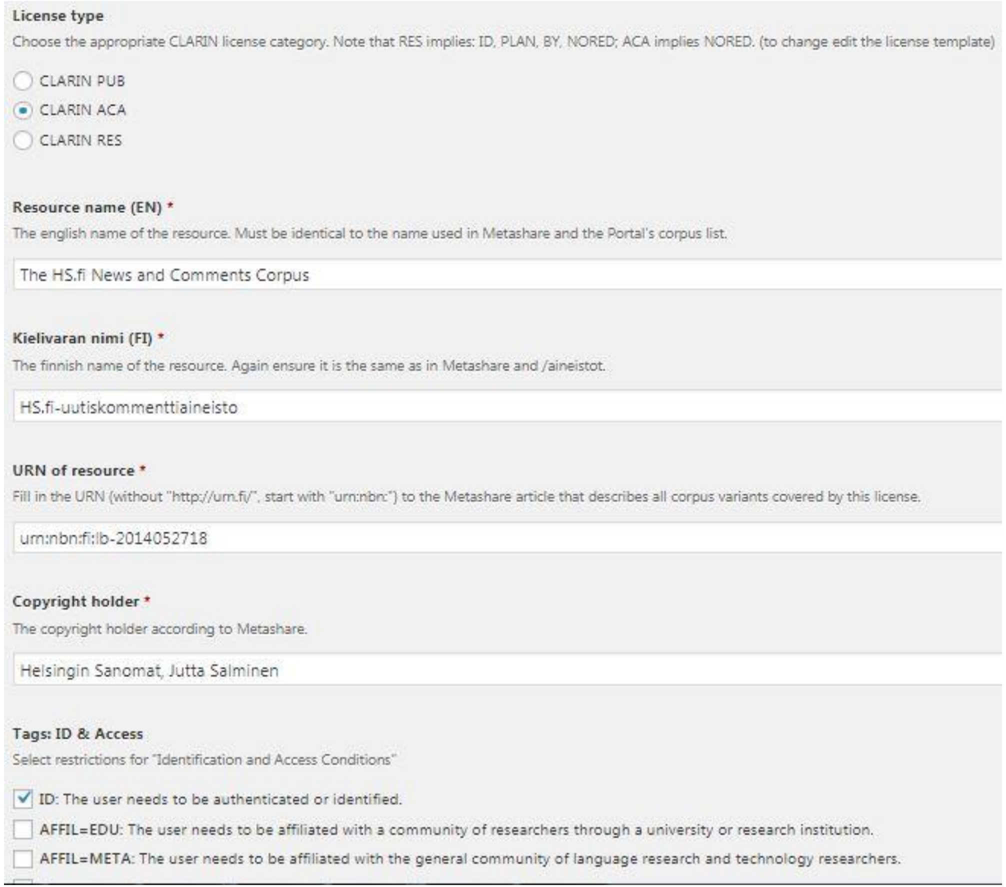
Last Updated: 01/16/2014

Metadata Language: English, Finnish (en, fi)

Metadata Creator

Imre Bartis

Make the license text available in PORTAL using templates



License type

Choose the appropriate CLARIN license category. Note that RES implies: ID, PLAN, BY, NORED; ACA implies NORED. (to change edit the license template)

○ CLARIN PUB
● CLARIN ACA
○ CLARIN RES

Resource name (EN) *

The english name of the resource. Must be identical to the name used in Metashare and the Portal's corpus list.

The HS.fi News and Comments Corpus

Kielivaran nimi (FI) *

The finnish name of the resource. Again ensure it is the same as in Metashare and /aineistot.

HS.fi-uutiskommenttiaineisto

URN of resource *

Fill in the URN (without "http://urn.fi/", start with "urn:nbn:") to the Metashare article that describes all corpus variants covered by this license.

urn:nbn:fi:lb-2014052718

Copyright holder *

The copyright holder according to Metashare.

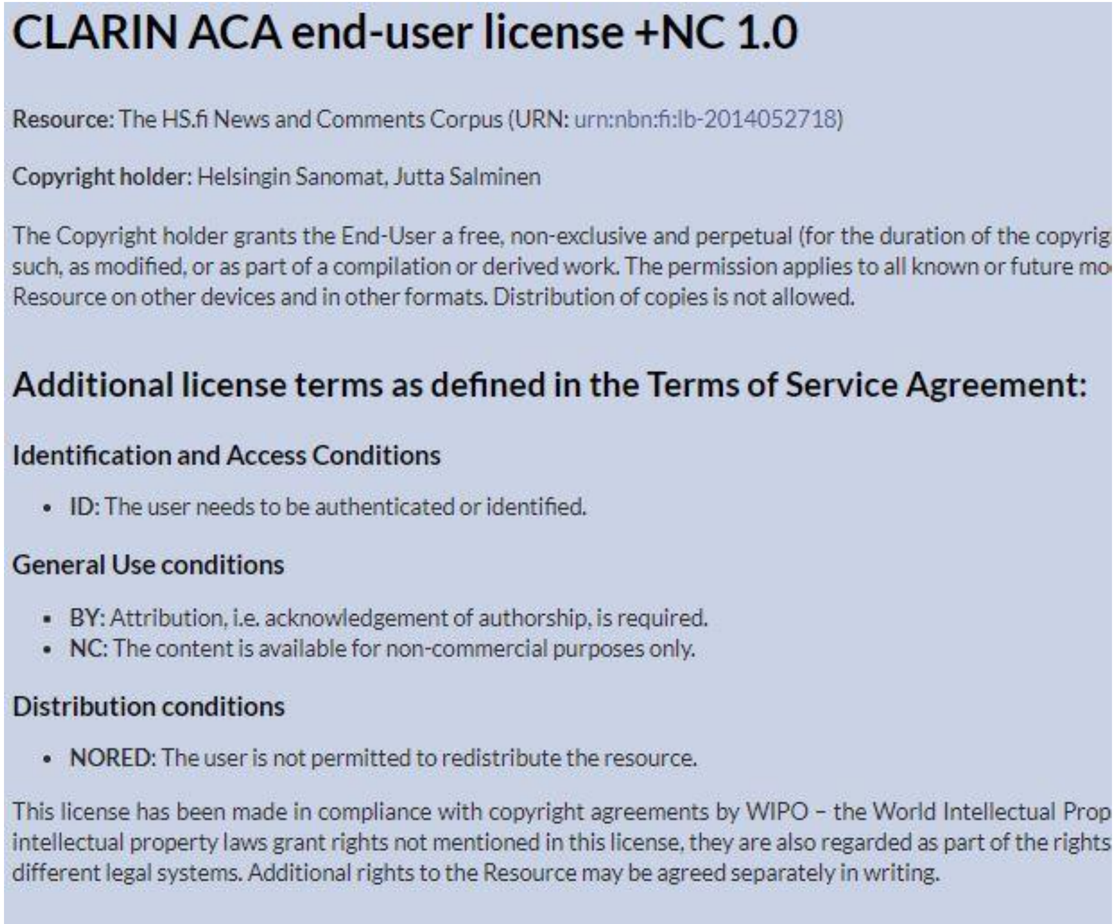
Helsingin Sanomat, Jutta Salminen

Tags: ID & Access

Select restrictions for "Identification and Access Conditions"

☑ ID: The user needs to be authenticated or identified.
□ AFFIL=EDU: The user needs to be affiliated with a community of researchers through a university or research institution.
□ AFFIL=META: The user needs to be affiliated with the general community of language research and technology researchers.

The license in PORTAL



CLARIN ACA end-user license +NC 1.0

Resource: The HS.fi News and Comments Corpus (URN: urn:nbn:fi:lb-2014052718)

Copyright holder: Helsingin Sanomat, Jutta Salminen

The Copyright holder grants the End-User a free, non-exclusive and perpetual (for the duration of the copyrig such, as modified, or as part of a compilation or derived work. The permission applies to all known or future mo Resource on other devices and in other formats. Distribution of copies is not allowed.

Additional license terms as defined in the Terms of Service Agreement:

Identification and Access Conditions

- **ID:** The user needs to be authenticated or identified.

General Use conditions

- **BY:** Attribution, i.e. acknowledgement of authorship, is required.
- **NC:** The content is available for non-commercial purposes only.

Distribution conditions

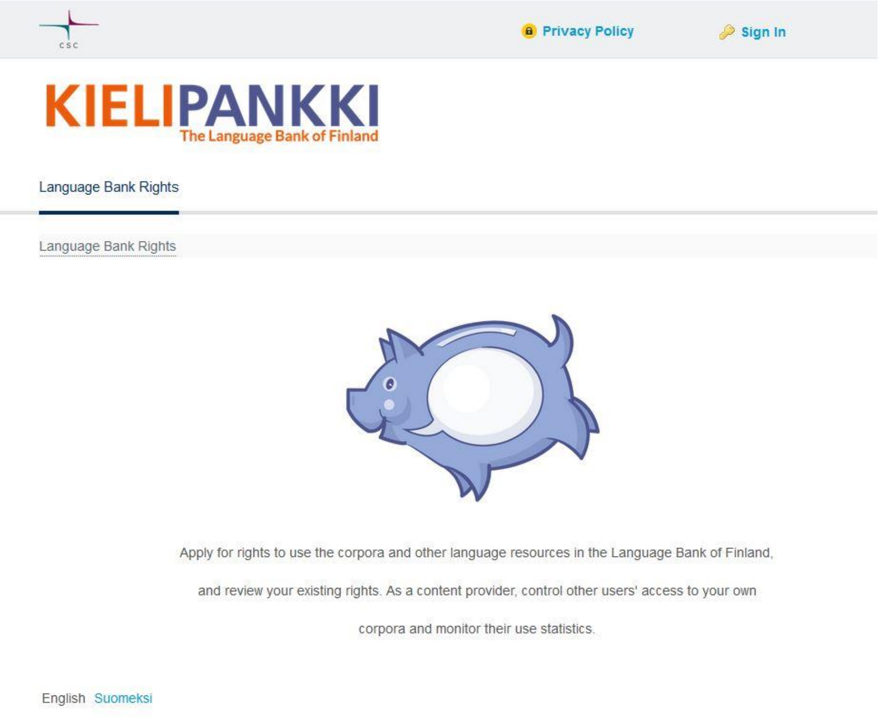
- **NORED:** The user is not permitted to redistribute the resource.

This license has been made in compliance with copyright agreements by WIPO – the World Intellectual Prop intellectual property laws grant rights not mentioned in this license, they are also regarded as part of the rights different legal systems. Additional rights to the Resource may be agreed separately in writing.

Language Bank Rights at Kielipankki

- Manages applications, access rights
- Instance of CSC’s Resource Entitlement Management System REMS
- REMS was originally developed for Bioscience (Elixir)
- Used at Kielipankki since April 2015 as Language Bank Rights (LBR)
- Controls 47 resources (10/2015: 35 resources)

Register the corpus in LBR



Apply for rights to use the corpora and other language resources in the Language Bank of Finland, and review your existing rights. As a content provider, control other users' access to your own corpora and monitor their use statistics.

- Register the license(s).
- Register HS.fi as a “resource” (using “URN:HS.fi”).
- Create forms to be filled in by applicants.
- Create a workflow that is followed during an application.
- Connect all of the above to a ”Catalog item”.

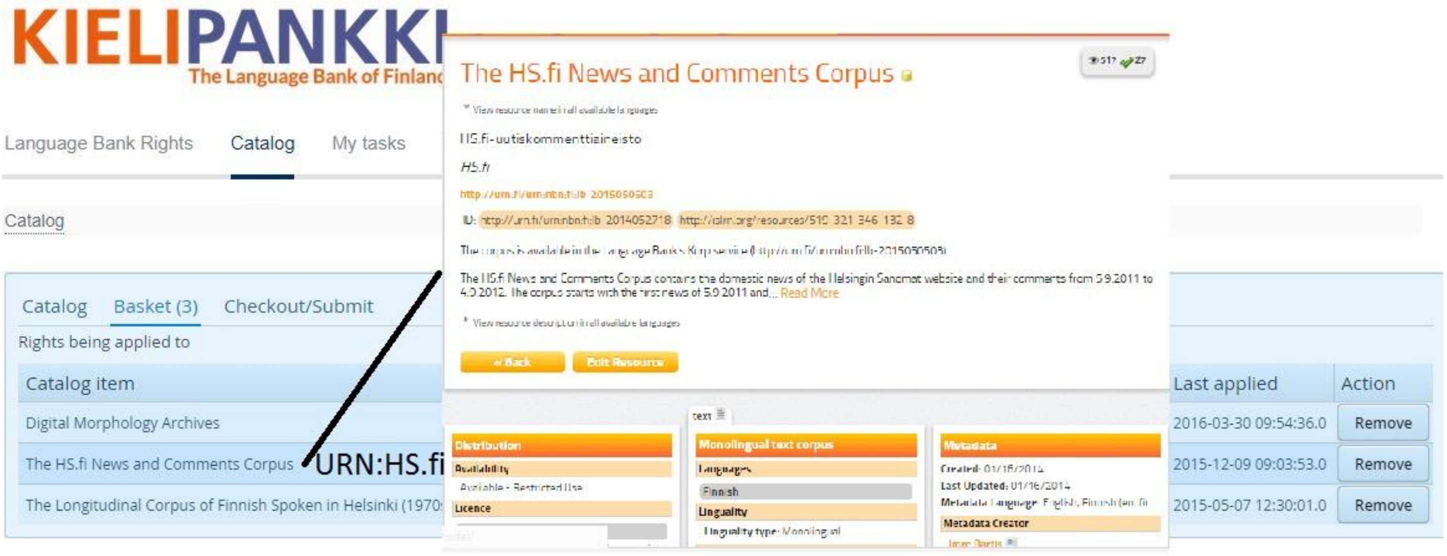
Registration of corpus completed

- Created metadata in METASHARE
- Assigned URN/Handle (here: URN:HS.fi) in PID
- License text now available in PORTAL
- Registered corpus in LBR

Language Bank Rights (LBR) Application and Approval

Applicant’s view

- Name is link to URN (to metadata)
- Applicant can add items to basket (webshop metaphor)



Rights being applied to

Catalog item	Last applied	Action
Digital Morphology Archives	2016-03-30 09:54:36.0	Remove
The HS.fi News and Comments Corpus	2015-12-09 09:03:53.0	Remove
The Longitudinal Corpus of Finnish Spoken in Helsinki (1970s	2015-05-07 12:30:01.0	Remove

The HS.fi News and Comments Corpus

517 27

▾ View resource name in all available languages

HS.fi-uutiskommenttiaineisto

HS.fi

http://urn.fi/urn:nbn:fi:lb-2015050503

ID: http://urn.fi/urn:nbn:fi:lb-2014052718 http://islrn.org/resources/519-321-346-132-8

The corpus is available in the Language Bank's Korp service (http://urn.fi/urn:nbn:fi:lb-2015050503)

The HS.fi News and Comments Corpus contains the domestic news of the Helsingin Sanomat website and their comments from 5.9.2011 to 4.9.2012. The corpus starts with the first news of 5.9.2011 and... Read More

▸ View resource description in all available languages

Distribution

Availability

Available - Restricted Use

Licence

Monolingual text corpus

Languages

Finnish

Linguality

Linguality type: Monolingual

Metadata

Created: 01/16/2014

Last Updated: 01/16/2014

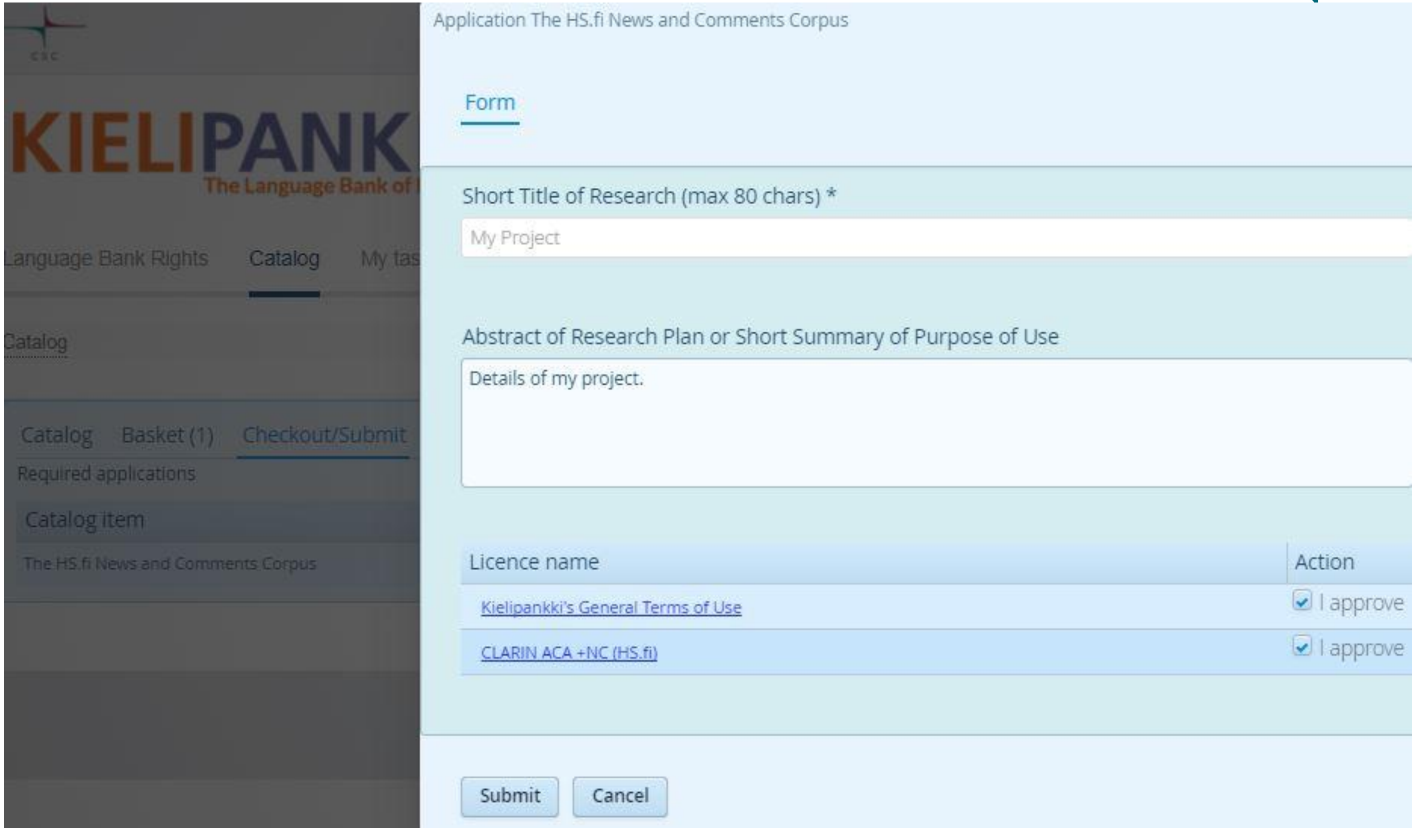
Metadata Language: English, Finnish (en, fi)

Metadata Creator

Imre Bartis

The application

- Fill out form
- Accept licenses
- ”Submit” sends application to approvers defined in workflow



Application The HS.fi News and Comments Corpus

Short Title of Research (max 80 chars) *

My Project

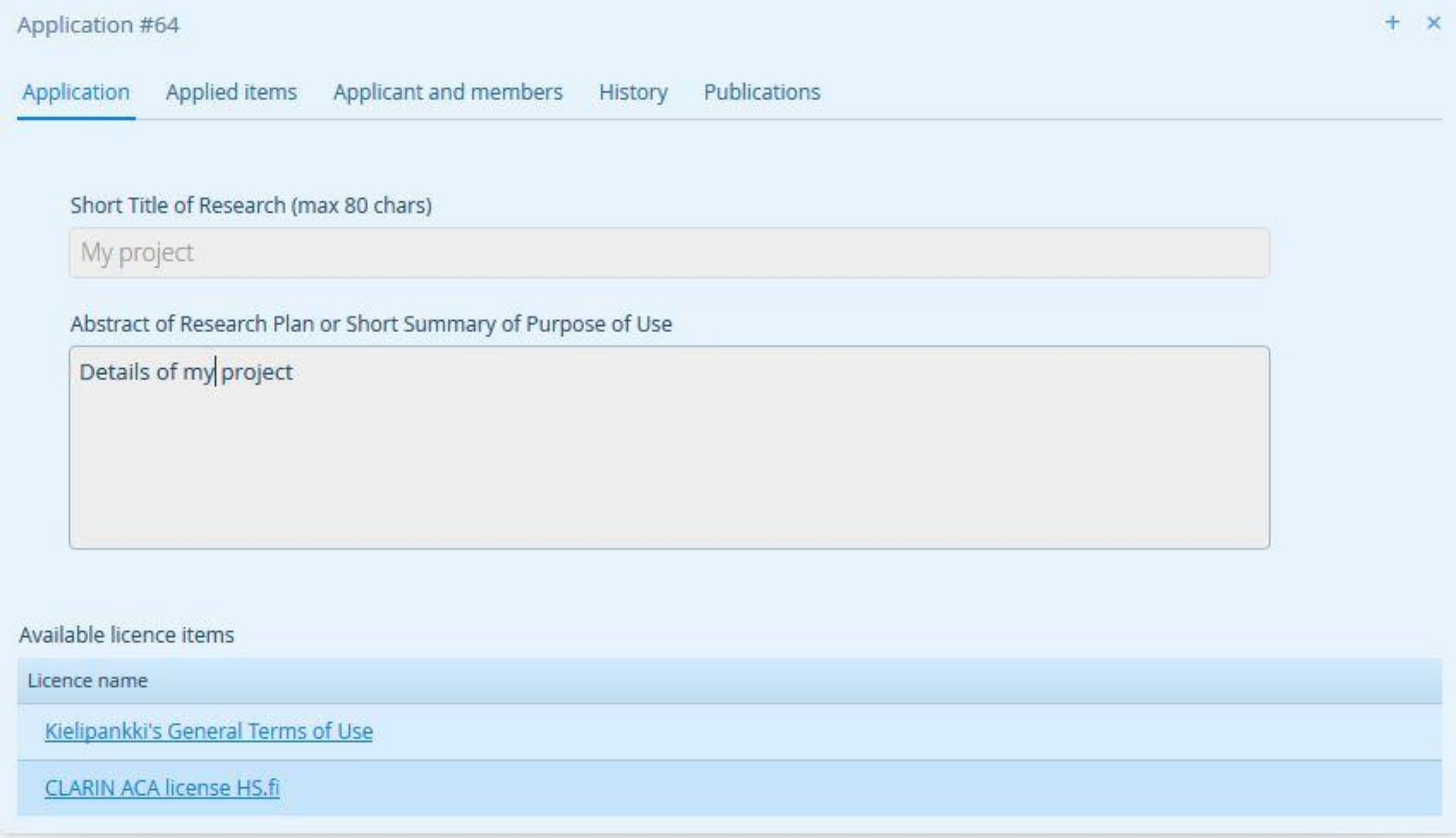
Abstract of Research Plan or Short Summary of Purpose of Use

Details of my project.

Licence name	Action
Kielipankki's General Terms of Use	☑ I approve
CLARIN ACA +NC (HS.fi)	☑ I approve

The approver view (1/4)

- The application itself



Application #64

Short Title of Research (max 80 chars)

My project

Abstract of Research Plan or Short Summary of Purpose of Use

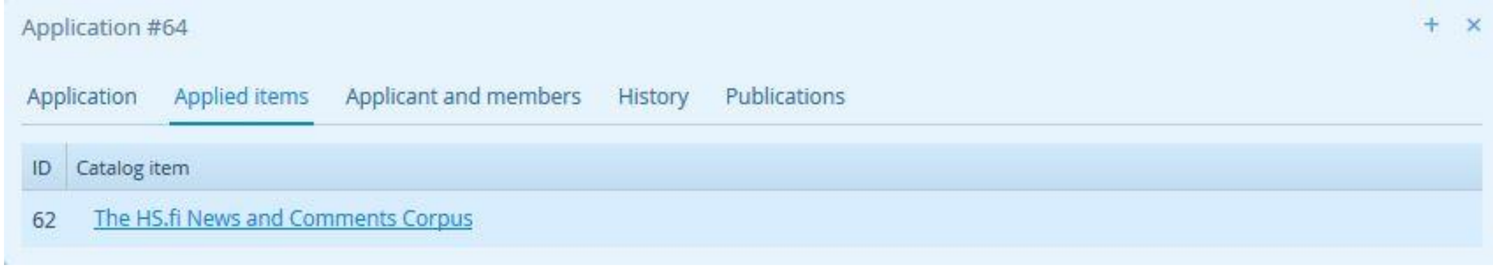
Details of my project

Available licence items

Licence name
Kielipankki's General Terms of Use
CLARIN ACA license HS.fi

The approver view (2/4)

- Applied items (could be more)

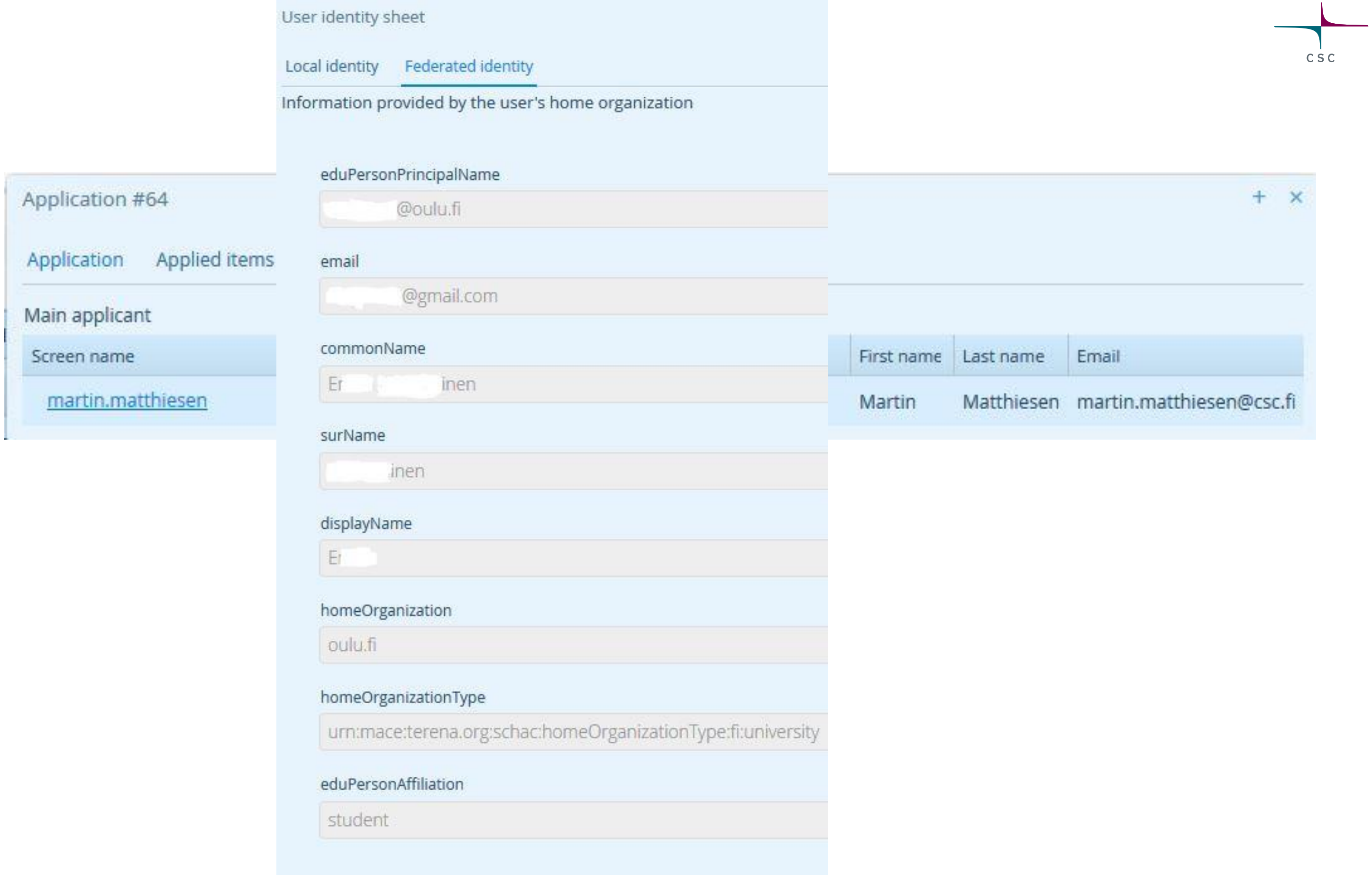


Application #64

ID	Catalog item
62	The HS.fi News and Comments Corpus

The approver view (3/4)

- Applicant
- More info available behind name.



Application #64

Main applicant

Screen name	First name	Last name	Email
martin.matthiesen	Martin	Matthiesen	martin.matthiesen@csc.fi

User identity sheet

Information provided by the user's home organization

eduPersonPrincipalName: [redacted]@oulu.fi

email: [redacted]@gmail.com

commonName: Er[redacted]inen

surName: [redacted]inen

displayName: Er[redacted]

homeOrganization: oulu.fi

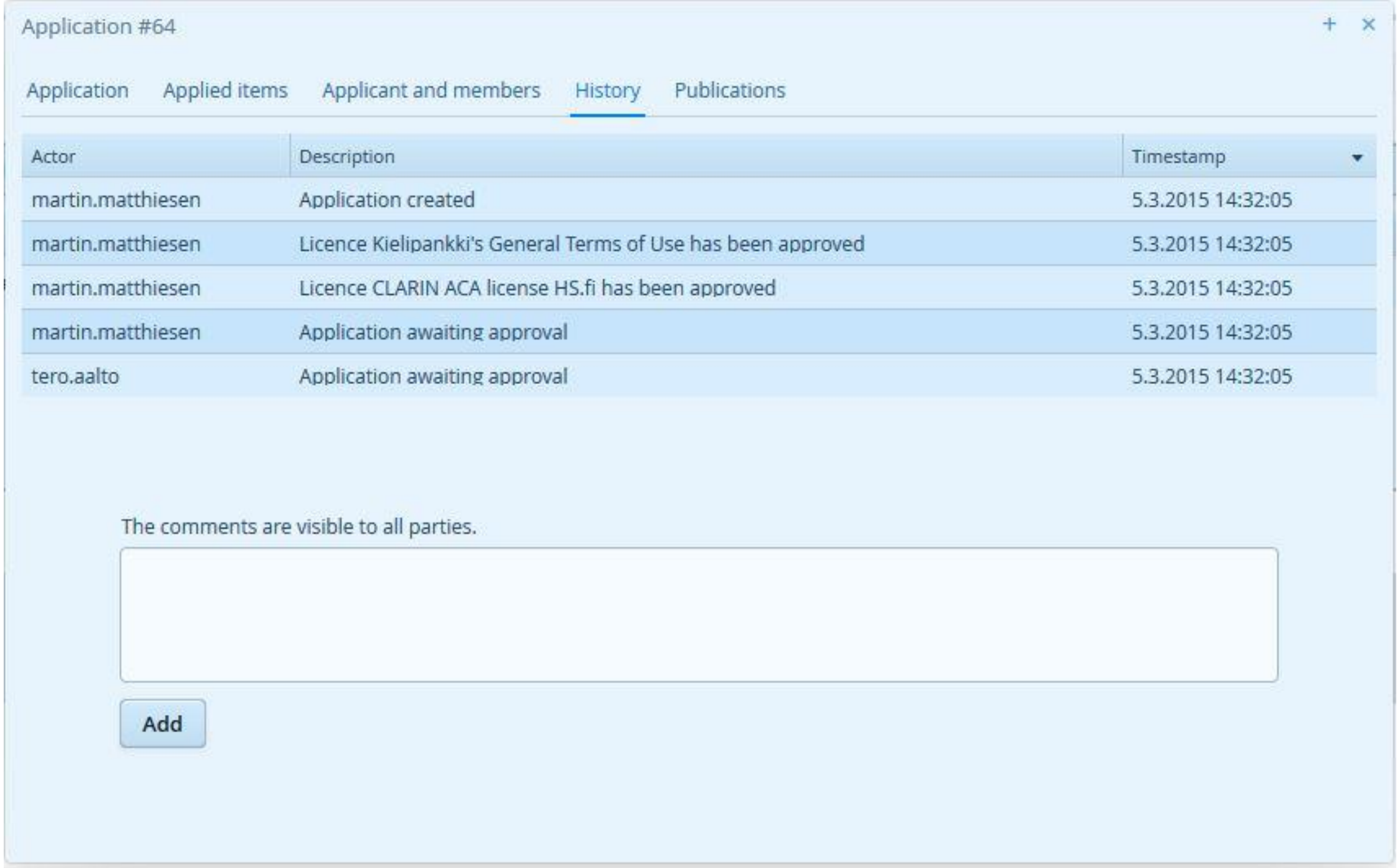
homeOrganizationType: urn:mace:terena.org:schac:homeOrganizationType:fi:university

eduPersonAffiliation: student

The approver view (4/4)

- History of application
- Who what when
- Comments can be added

(Publications: Applicant can augment application later with relevant publications, skipped here)

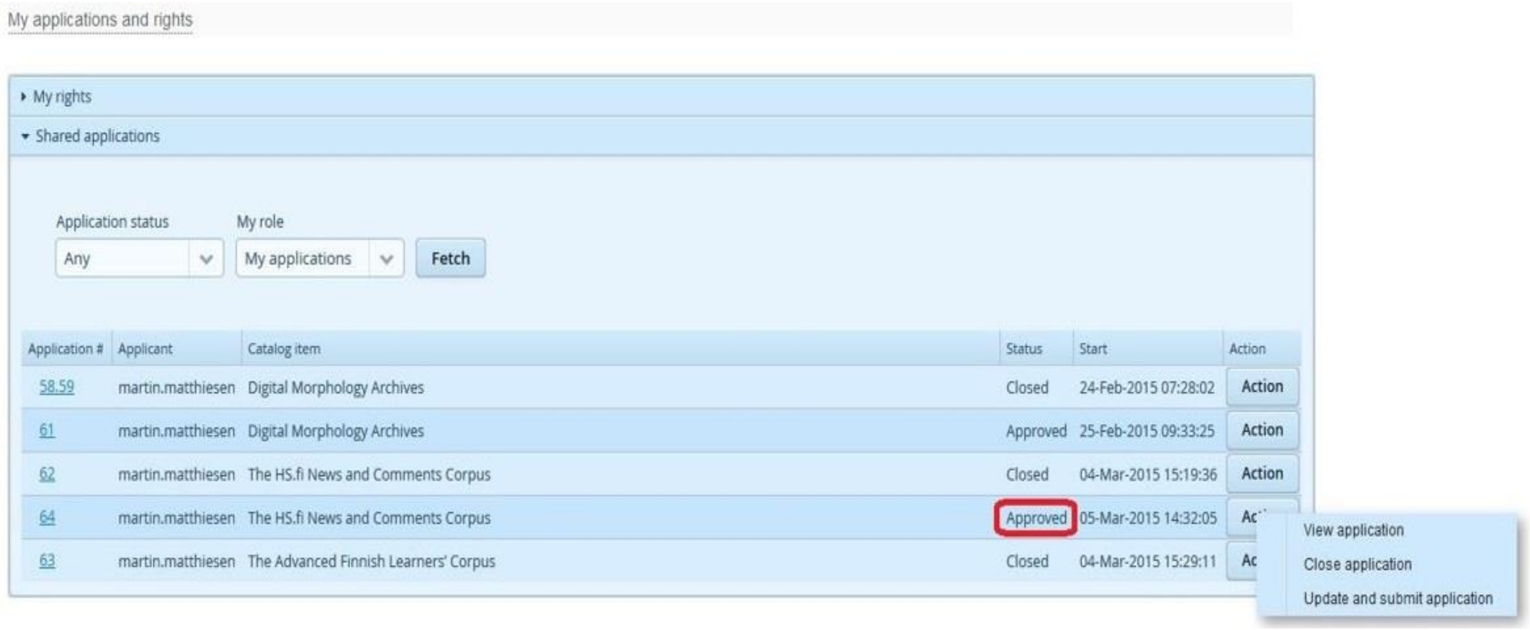


Actor	Description	Timestamp
martin.matthiesen	Application created	5.3.2015 14:32:05
martin.matthiesen	Licence Kielipankki's General Terms of Use has been approved	5.3.2015 14:32:05
martin.matthiesen	Licence CLARIN ACA license HS.fi has been approved	5.3.2015 14:32:05
martin.matthiesen	Application awaiting approval	5.3.2015 14:32:05
tero.aalto	Application awaiting approval	5.3.2015 14:32:05

The comments are visible to all parties.

The applicant's view

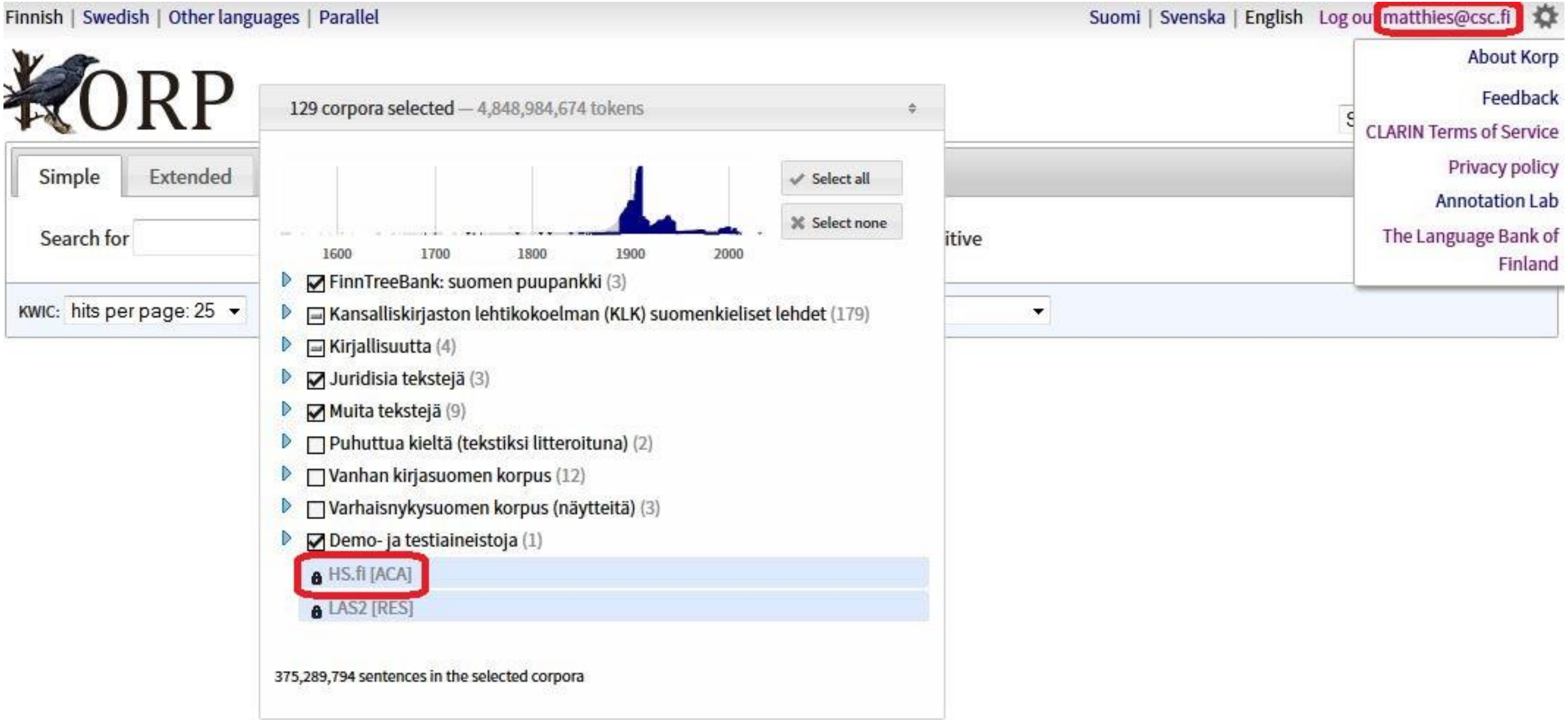
- Status of application(s)
- Applicant can Close or Update



Application status: Any

My role: My applications

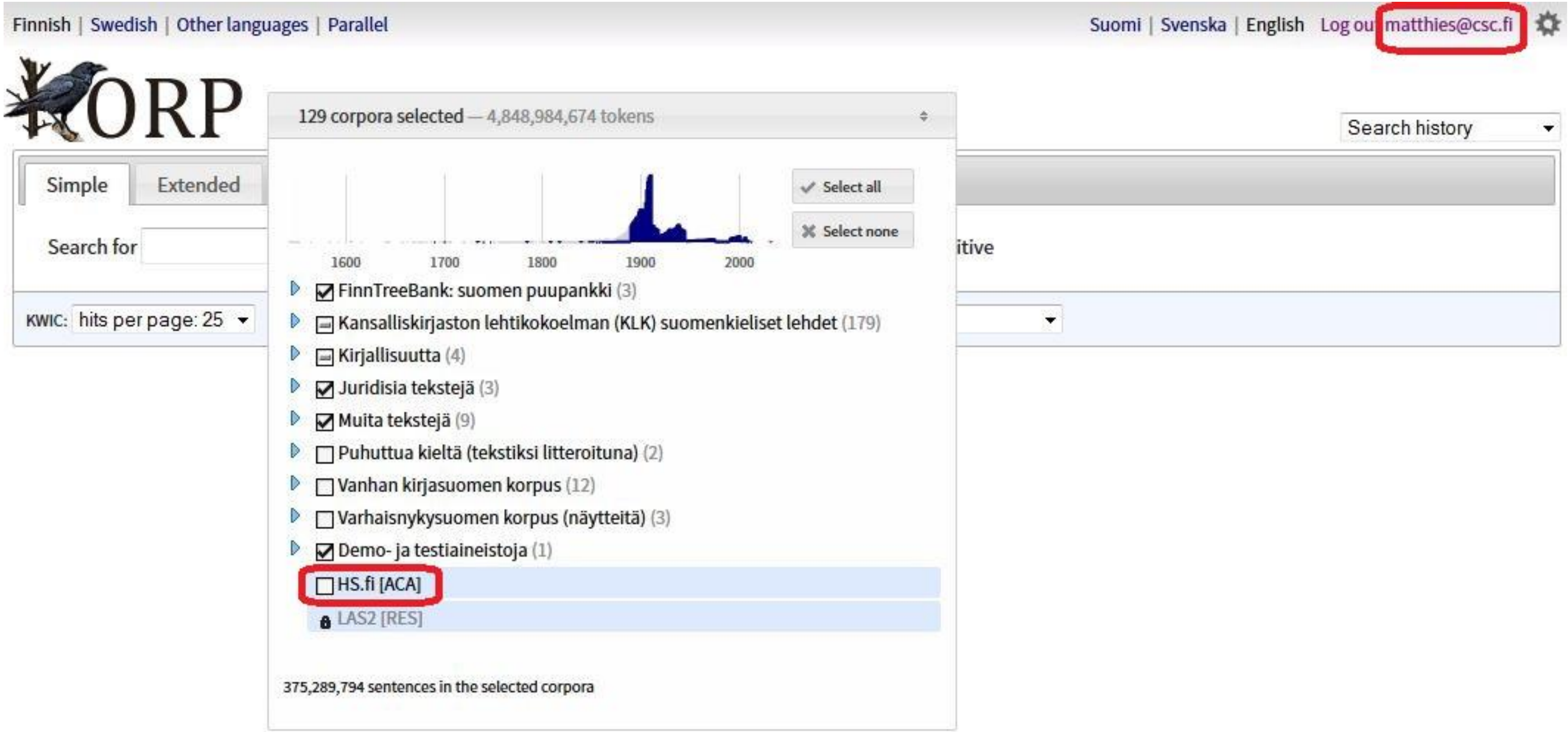
Application #	Applicant	Catalog item	Status	Start	Action
58.59	martin.matthiesen	Digital Morphology Archives	Closed	24-Feb-2015 07:28:02	Action
61	martin.matthiesen	Digital Morphology Archives	Approved	25-Feb-2015 09:33:25	Action
62	martin.matthiesen	The HS.fi News and Comments Corpus	Closed	04-Mar-2015 15:19:36	Action
64	martin.matthiesen	The HS.fi News and Comments Corpus	Approved	05-Mar-2015 14:32:05	Ac
63	martin.matthiesen	The Advanced Finnish Learners' Corpus	Closed	04-Mar-2015 15:29:11	Ac



129 corpora selected — 4,848,984,674 tokens

- ☑ FinnTreeBank: suomen puupankki (3)
- ⊟ Kansalliskirjaston lehtikokoelman (KLK) suomenkieliset lehdet (179)
- ⊟ Kirjallisuutta (4)
- ☑ Juridisia tekstejä (3)
- ☑ Muita tekstejä (9)
- □ Puhuttua kieltä (tekstiksi litteroituna) (2)
- □ Vanhan kirjasuomen korpus (12)
- □ Varhaisnykysuomen korpus (näytteitä) (3)
- ☑ Demo- ja testiaineistoja (1)
- 🔒 HS.fi [ACA]
- 🔒 LAS2 [RES]

375,289,794 sentences in the selected corpora



129 corpora selected — 4,848,984,674 tokens

- ☑ FinnTreeBank: suomen puupankki (3)
- ⊟ Kansalliskirjaston lehtikokoelman (KLK) suomenkieliset lehdet (179)
- ⊟ Kirjallisuutta (4)
- ☑ Juridisia tekstejä (3)
- ☑ Muita tekstejä (9)
- □ Puhuttua kieltä (tekstiksi litteroituna) (2)
- □ Vanhan kirjasuomen korpus (12)
- □ Varhaisnykysuomen korpus (näytteitä) (3)
- ☑ Demo- ja testiaineistoja (1)
- □ HS.fi [ACA]
- LAS2 [RES]

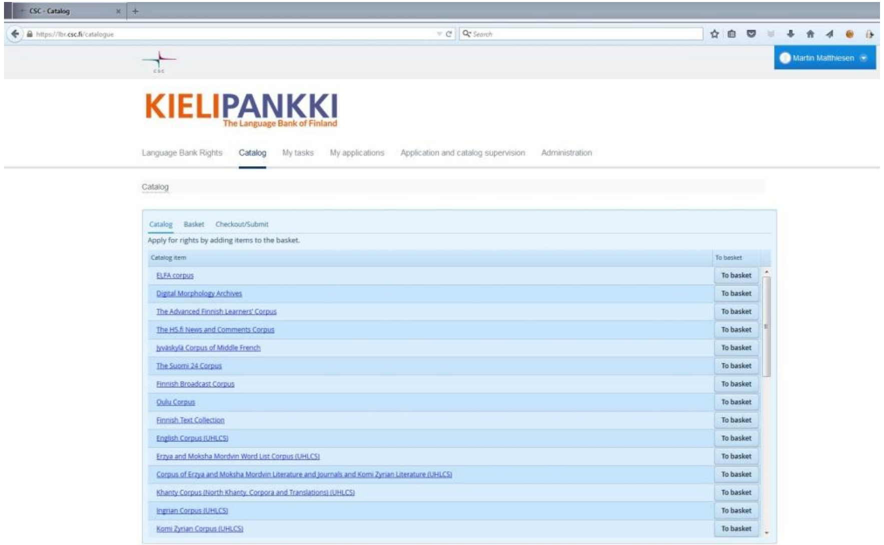
375,289,794 sentences in the selected corpora

Recap: Language Bank Rights at Kielipankki

- Based on Resource Entitlement Management System (REMS)
- Connects
  - resource users
  - approvers / licensors
  - licenses
  - resources
- Keeps electronic **paper trail**
- Heavily relies on **SAML2 Single Sign On** (eduGAIN, local federations, CLARIN IdP, Finnish Eduuni)

Thank you!

- The Language Bank:
- https://www.kielipankki.fi/language-bank/
- Language Bank Rights: https://lbr.csc.fi/
- Demo: https://remsdemo.csc.fi/
- Time for comments, questions.



Apply for rights by adding items to the basket.

Catalog item	To basket
ELFA corpus	To basket
Digital Morphology Archives	To basket
The Advanced Finnish Learners' Corpus	To basket
The HS.fi News and Comments Corpus	To basket
Jyväskylä Corpus of Middle French	To basket
The Suomi 24 Corpus	To basket
Finnish Broadcast Corpus	To basket
Oulu Corpus	To basket
Finnish Text Collection	To basket
English Corpus (UHLCS)	To basket
Erzya and Moksha Mordvin Word List Corpus (UHLCS)	To basket
Corpus of Erzya and Moksha Mordvin Literature and Journals and Komi Zyrian Literature (UHLCS)	To basket
Khanty Corpus (North Khanty, Corpora and Translations) (UHLCS)	To basket
Ingrian Corpus (UHLCS)	To basket
Komi Zyrian Corpus (UHLCS)	To basket