Making the impact on research and society

a case study: open repository and crowdsourcing solutions developed for the Finno-Ugric Digitization Pilot Project at the National Library of Finland

Jussi-Pekka Hakkarainen
Project Manager
Research Library
National Library of Finland

S094. History of science and the ecology of knowledge: the limitations, expectations, and needs of four knowledge communities
26 July 2013, ICSHTM, Manchester

DIGITIZATION PROJECT OF KINDRED LANGUAGES

Kone Foundation Language Programme

The National Library of Finland (NLF) is executing a pilot project (7/2012-10/2013) that aims for digitizing and publishing of Finno-Ugric material for the benefit of the linquistic research within the Kone Foundation Language Programme (2012-2016). Decision on the follow-up project for 2014-2016 is pending.

The objective of the **Kone Foundation Language Programme** is to advance the documentation and status of small Finno-Ugrian languages, the Finnish language, and minority languages in Finland. Both, the scientific community and all language users, will profit from the results of this research and documentation.

DIGITIZATION PROJECT OF KINDRED LANGUAGES

Objectives of the Pilot Project

Another key objective of the project is to foster a culture of openness and interaction in linguistic research. This entails, for example, **the unlimited availability, accessibility and usability of source material** and research results via a virtual library, as well as **the participation of the language community** in various stages of documentation and application of research results.

It is essential not only ensure the availability of language materials, but to make it easy for different users to approach and use them. Free, open access to the material ensures that it can be used by both the **academic community** and the **speakers of the kindred languages** of Finnish.

DIGITIZATION PROJECT OF KINDRED LANGUAGES

Materials for Digitization

The focus is on the Finno-Ugric languages, which suddenly became socially important, at the beginning of the Soviet era in the 1920s and 1930s. No contemporary literature will be digitised.

The researchers made the selection of materials for the digitization plan in autumn 2011. The selection consists of 17 000 pages of monographs in **Veps, Ingrian, Mari (Meadow and Hill Mari) and Mordvinic (Erzya and Moksha)** languages and around 20 000 pages of newspapers in Mari and Mordvinic languages. Monographs are mainly school and text books and in many cases they are translations from Russia to the local languages.

DIGITIZATION PROJECT OF KINDRED LANGUAGES

Co-operation with Russian Partners

During the pilot project the NLF has produced the research infrastructure (repository and OCR editor) and takes care of co-operation with the Finnish (**Helsinki University Library**) and Russian partners (**National Library of Russia**, NLR and **National Library Resources**).

The material was digitised from the collections of the NLR. This is the first time that material published in the former Soviet Union has been made freely available for public use in the NLF (or any foreign?) data systems.

DIGITIZATION PROJECT OF KINDRED LANGUAGES

Erzya Language as a typical example

Mordvinic language, **Erzya,** was converted into a medium of popular education, enlightenment and dissemination of information pertinent to the developing political agenda of the Soviet state. The "deluge" of popular Erzya literature, 1920s-1930s, suddenly challenged **the lexical orthographic norms** of the limited ecclesiastical publications from the 1880s.

Newspapers were written in orthographies and in word forms that the locals would understand. Schoolbooks were written to address the separate needs of both the adults and children. **New concepts were introduced in the native language**. It was the beginning of a renaissance and period of enlightenment.

DIGITIZATION PROJECT OF KINDRED LANGUAGES

Open Access to the Material

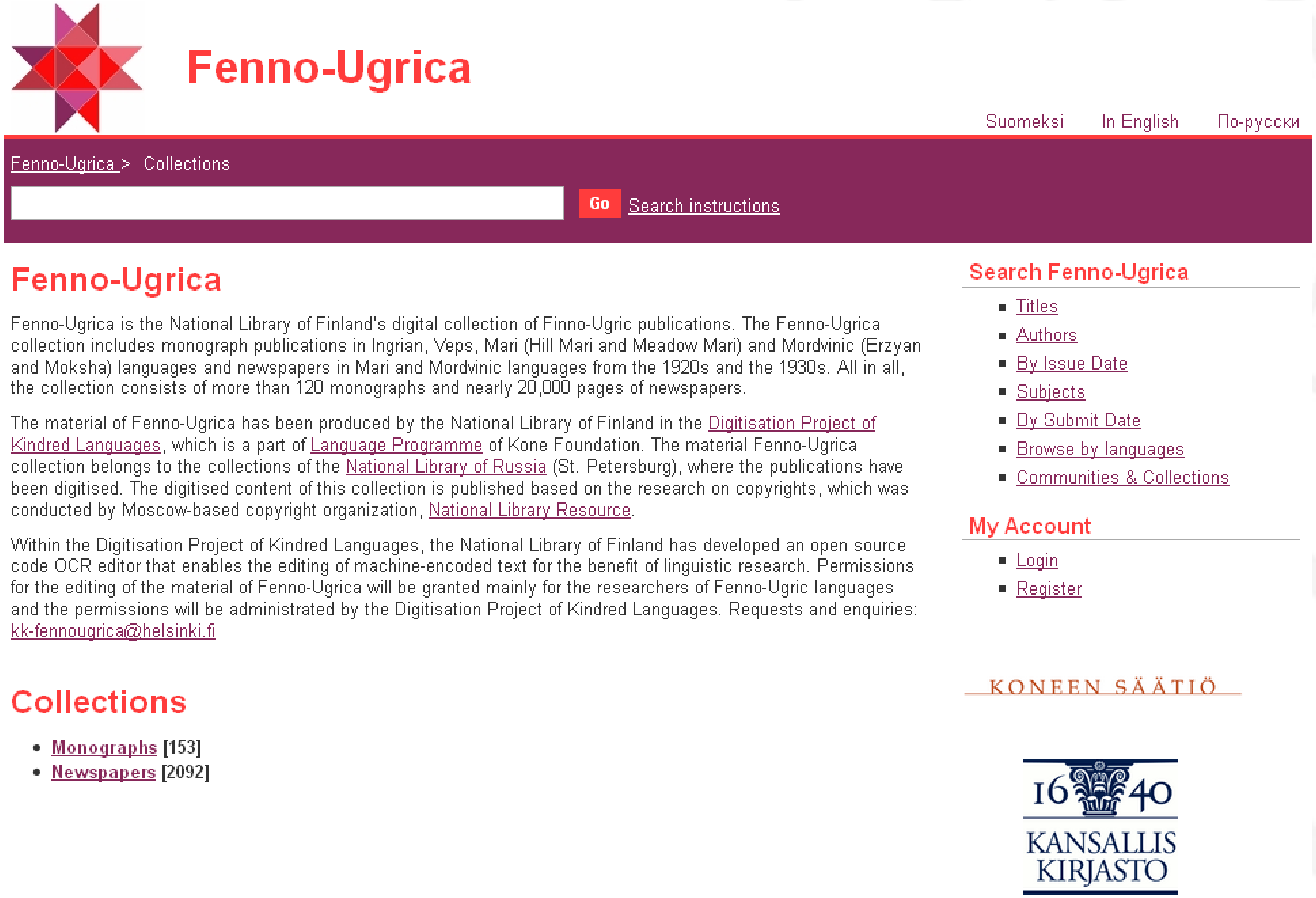
Since the Kone Foundation Language Programme has an objective of free access to the materials, NLF will publish the material as its own repository and provides an open access **without geographical or IP restrictions**.

In order to publish the material as public domain, the copyrights regarding material was needed to be cleared. The research on copyrights was conducted by Moscow-based **National Library Resource** during winter 2013.

Public domain allows NLF to donate the language-resources after editing to the **FIN-Clarin** for the benefit of other research (linguistic) communities.

DIGITIZATION PROJECT OF KINDRED LANGUAGES

Open Repository



Fenno-Ugrica

Suomeksi In English По-русски

Fenno-Ugrica > Collections

Search instructions

Fenno-Ugrica

Fenno-Ugrica is the National Library of Finland's digital collection of Finno-Ugric publications. The Fenno-Ugrica collection includes monograph publications in Ingrian, Veps, Mari (Hill Mari and Meadow Mari) and Mordvinic (Erzyan and Moksha) languages and newspapers in Mari and Mordvinic languages from the 1920s and the 1930s. All in all, the collection consists of more than 120 monographs and nearly 20,000 pages of newspapers.

The material of Fenno-Ugrica has been produced by the National Library of Finland in the Digitisation Project of Kindred Languages, which is a part of Language Programme of Kone Foundation. The material Fenno-Ugrica collection belongs to the collections of the National Library of Russia (St. Petersburg), where the publications have been digitised. The digitised content of this collection is published based on the research on copyrights, which was conducted by Moscow-based copyright organization, National Library Resource.

Within the Digitisation Project of Kindred Languages, the National Library of Finland has developed an open source code OCR editor that enables the editing of machine-encoded text for the benefit of linguistic research. Permissions for the editing of the material of Fenno-Ugrica will be granted mainly for the researchers of Fenno-Ugric languages and the permissions will be administrated by the Digitisation Project of Kindred Languages. Requests and enquiries: kk-fennougrica@helsinki.fi

Collections

- Monographs [153]
- Newspapers [2092]

Search Fenno-Ugrica

- Titles
- Authors
- By Issue Date
- Subjects
- By Submit Date
- Browse by languages
- Communities & Collections

My Account

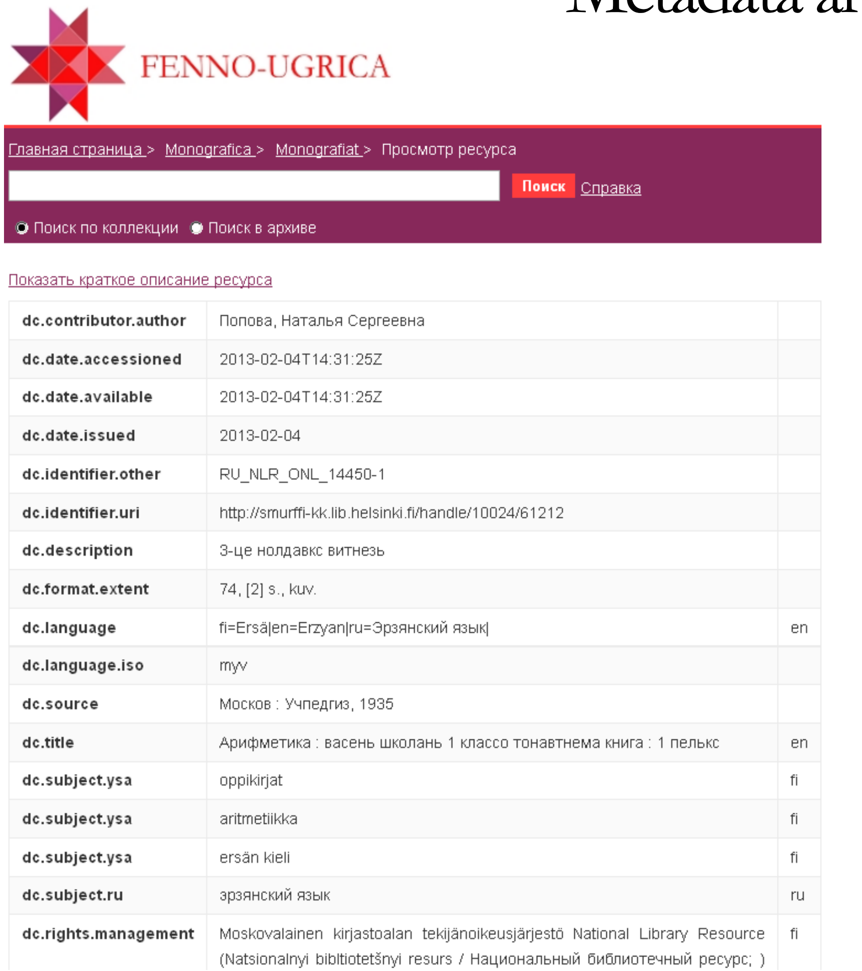
- Login
- Register

KONEEN SÄÄTIÖ

16 40 KANSALLIS KIRJASTO

DIGITIZATION PROJECT OF KINDRED LANGUAGES

Metadata and Availability


FENNO-UGRICA

Главная страница > Monografica > Monografiat > Просмотр ресурса

Поиск Справка

● Поиск по коллекции ○ Поиск в архиве

Показать краткое описание ресурса

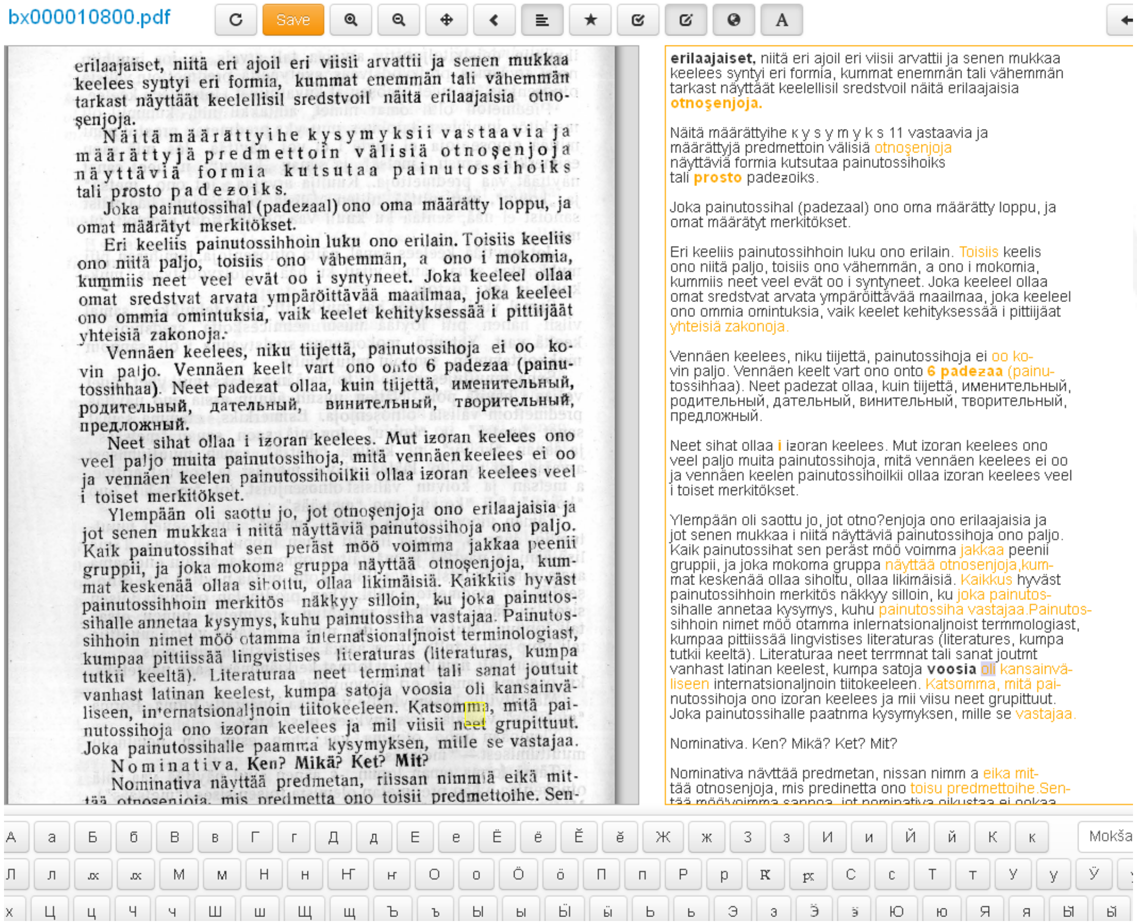
dc.contributor.author	Попова, Наталья Сергеевна	
dc.date.accessioned	2013-02-04T14:31:25Z	
dc.date.available	2013-02-04T14:31:25Z	
dc.date.issued	2013-02-04	
dc.identifier.other	RU_NLR_ONL_14450-1	
dc.identifier.uri	http://smurffi-kk.lib.helsinki.fi/handle/10024/61212	
dc.description	3-це нолдавкс витнезь	
dc.format.extent	74, [2] s., kuv.	
dc.language	fi=Ersä|en=Erzyan|ru=Эрзянский язык|	en
dc.language.iso	myv	
dc.source	Москов : Учпедгиз, 1935	
dc.title	Арифметика : васень школань 1 классо тонавтнема книга : 1 пелькс	en
dc.subject.ysa	oppikirjat	fi
dc.subject.ysa	aritmetiikka	fi
dc.subject.ysa	ersän kieli	fi
dc.subject.ru	эрзянский язык	ru
dc.rights.management	Moskovalainen kirjastoalan tekijänoikeusjärjestö National Library Resource (Natsionalnyi bibltiotetšnyi resurs / Национальный библиотечный ресурс;)	fi

The material is catalogued directly to the repository in **Dublin Core** format, but the metadata will be linked to the local library cataloques too.

In order to ease the access to the material, the material will be linked to **Europeana** and the **National Digital Library** of Finland and it can be browsed through its interface, **Finna**.

DIGITIZATION PROJECT OF KINDRED LANGUAGES

OCR Editor for Enriching the Text



erilaajaiset, niitä eri ajoil eri viisii arvattii ja senen mukkaa keelees syntyi eri formia, kummat enemmän tali vähemmän tarkast näyttäät keelellisil sredstvoil näitä erilaajaisia otnoşenjoja.

Näitä määrättyihe kysymyksii vastaavia ja määrättyjä predmettoin välisiä otnoşenjoja näyttäviä formia kutsutaa painutossihoiks tali prosto padezoiks.

Joka painutossihal (padezaal) ono oma määrätty loppu, ja omat määrätyt merkitökset.

Eri keeliis painutossihhoin luku ono erilain. Toisiis keeliis ono niitä paljo, toisiis ono vähemmän, a ono i mokomia, kummiis neet veel evät oo i syntyneet. Joka keeleel ollaa omat sredstvat arvata ympäröittävää maailmaa, joka keeleel ono ommia omintuksia, vaik keelet kehityksessää i pittiijäät yhteisiä zakonoja.

Vennäen keelees, niku tiijettä, painutossihoja ei oo kovin paljo. Vennäen keelt vart ono onto 6 padezaa (painutossihhaa). Neet padezat ollaa, kuin tiijettä, именительный, родительный, дательный, винительный, творительный, предложный.

Neet sihat ollaa i izoran keelees. Mut izoran keelees ono veel paljo muita painutossihoja, mitä vennäen keelees ei oo ja vennäen keelen painutossihoilkii ollaa izoran keelees veel i toiset merkitökset.

Ylempään oli saottu jo, jot otnoşenjoja ono erilaajaisia ja jot senen mukkaa i niitä näyttäviä painutossihoja ono paljo. Kaik painutossihat sen peräst möö voimma jakkaa peenii gruppii, ja joka mokoma gruppa näyttää otnoşenjoja, kummat keskenää ollaa sihottu, ollaa likimäisiä. Kaikkiis hyväst painutossihhoin merkitös näkkyy silloin, ku joka painutossihalle annetaa kysymys, kuhu painutossiha vastajaa. Painutossihhoin nimet möö otamma internatsionaljnoist terminologiast, kumpaa pittiissää lingvistises literaturas (literaturas, kumpa tutkii keeltä). Literaturaa neet terminat tali sanat joutuit vanhast latinan keelest, kumpa satoja voosia oli kansainväliseen, internatsionaljnoin tiitokeeleen. Katsomma, mitä painutossihoja ono izoran keelees ja mil viisii neet grupittuut. Joka painutossihalle paamma kysymyksen, mille se vastajaa.

Nominativa. **Ken? Mikä? Ket? Mit?**

Nominativa näyttää predmetan, riissan nimmiä eikä mittää otnosenjoja, mis predmetta ono toisii predmettoihe. Sen-

erilaajaiset, niitä eri ajoil eri viisii arvattii ja senen mukkaa keelees syntyi eri formia, kummat enemmän tali vähemmän tarkast näyttäät keelellisil sredstvoil näitä erilaajaisia **otnoşenjoja.**

Näitä määrättyihe к y s y m y k s 11 vastaavia ja määrättyjä predmettoin välisiä otnoşenjoja näyttäviä formia kutsutaa painutossihoiks tali **prosto** padezoiks.

Joka painutossihal (padezaal) ono oma määrätty loppu, ja omat määrätyt merkitökset.

Eri keeliis painutossihhoin luku ono erilain. Toisiis keelis ono niitä paljo, toisiis ono vähemmän, a ono i mokomia, kummiis neet veel evät oo i syntyneet. Joka keeleel ollaa omat sredstvat arvata ympäröittävää maailmaa, joka keeleel ono ommia omintuksia, vaik keelet kehityksessää i pittiijäat yhteisiä zakonoja.

Vennäen keelees, niku tiijettä, painutossihoja ei oo kovin paljo. Vennäen keelt vart ono onto **6 padezaa** (painutossihhaa). Neet padezat ollaa, kuin tiijettä, именительный, родительный, дательный, винительный, творительный, предложный.

Neet sihat ollaa **i** izoran keelees. Mut izoran keelees ono veel paljo muita painutossihoja, mitä vennäen keelees ei oo ja vennäen keelen painutossihoilkii ollaa izoran keelees veel i toiset merkitökset.

Ylempään oli saottu jo, jot otno?enjoja ono erilaajaisia ja jot senen mukkaa i niitä näyttäviä painutossihoja ono paljo. Kaik painutossihat sen peräst möö voimma jakkaa peenii gruppii, ja joka mokoma gruppa näyttää otnosenjoja,kummat keskenää ollaa siholtu, ollaa likimäisiä. Kaikkus hyväst painutossihhoin merkitös näkkyy silloin, ku joka painutossihalle annetaa kysymys, kuhu painutossiha vastajaa.Painutossihhoin nimet möö otamma inlernatsionaljnoist termmologiast, kumpaa pittiissää lingvistises literaturas (literatures, kumpa tutkii keeltä). Literaturaa neet terrmnat tali sanat joutmt vanhast latinan keelest, kumpa satoja **voosia** oli kansainväliseen internatsionaljnoin tiitokeeleen. Katsomma, mitä painutossihoja ono izoran keelees ja mii viisu neet grupittuut. Joka painutossihalle paatnma kysymyksen, mille se vastajaa.

Nominativa. Ken? Mikä? Ket? Mit?

Nominativa nävttää predmetan, nissan nimm a eika mittää otnosenjoja, mis predinetta ono toisu predmettoihe.Sen-

А а Б б В в Г г Д д Е е Ё ё Ĕ ĕ Ж ж З з И и Й й К к
Л л Ӆ ӆ М м Н н Ӈ ӈ О о Ӧ ӧ П п Р р Ꝛ ꝛ С с Т т У у Ӱ
Х Ц ц Ч ч Ш ш Щ щ Ъ ъ Ы ы Ӹ ӹ Ь ь Э э Ӭ ӭ Ю ю Я я Ӹ ӹ

The NLF has developed an **OCR editor** to support the research use of the material. The editor allows text that has undergone a process of machine identification to be edited for the purposes of linguistic research.

Editor ia capable **for correcting the alphabets that cannot be recognized** upon digitization or will be misread by the OCR programme.

Once the text will be corrected, the edited material **will be re-uploaded.**

DIGITIZATION PROJECT OF KINDRED LANGUAGES

Impact on research and society

Are the **scientific community** and **language users**, profiting from the results of this research and documentation?

As for the research of Finno-Ugric languages, the publication of open-access and searchable written materials from the 1920s and 1930s is a **“gold mine”**. The linguistically oriented population can also find writings to their delight:

1) **lexical items specific to a given publication**
2) **orthographically documented specifics of phonetics.**

Also the historians, social scientists and laymen with interests in specific local publications can now find text materials pertinent to their studies.

DIGITIZATION PROJECT OF KINDRED LANGUAGES

Impact on research and society

When it comes to the **societies of the minority languages**, the impact cannot be specifically measured. However, one can notice slight changes in attitudes:

1) Community participation and interaction were also supported and through the interactive research with the **citizen scientists,** who carried out proofing work and thus contributed to research directly. New ”scientists” are willing to join the project and there are a plenty of local initiatives in about to take off.
2) The published material was unlike to be digitized by the Russian libraries, so the project **made the preservation of the material** possible. The Russian libraries have already started the speak more about open access etc.

DIGITIZATION PROJECT OF KINDRED LANGUAGES

Resources and Services Online

Project Web Site

www.nationallibrary.fi/services/digitaalisetkokoelmat/finnougric_en_ru.html

Fenno-Ugrica Collection

fennougrica.kansalliskirjasto.fi/

Fennio-Ugrica Blog

blogs.helsinki.fi/fennougrica/

National Library of Finland

www.nationallibrary.fi/

DIGITIZATION PROJECT OF KINDRED LANGUAGES
Contact Details and Further Information

Jussi-Pekka Hakkarainen
Project Manager
National Library of Finland
Research Library

The National Library of Finland
P.O. Box 26 (Teollisuuskatu 23)
00014 University of Helsinki
kk-fennougrica@helsinki.fi