Pirkko Nuolijärvi / Toni Suutari

The landscape of the Finnish language research infrastructure

Abstract

Languages are a basic part of cultural heritage and the language collections and archives are an essential research infrastructure for humanities. Nowadays, there is an intensive co-operation between Finnish national institutes within the larger European context. This paper presents two nationwide projects in Finland working for the better use of language resources. These projects, FIN-CLARIN and National Digital Library belong, to some extent, also to the work of the Research Institute for the Languages of Finland, which is responsible for the nationally remarkable language collections.

Tiivistelmä

Kielet ovat kulttuuriperinnön perusta, ja kielenaineskokoelmat ja arkistot ovat humanistisille tieteenaloille tärkeä infrastruktuuri. Nykyisin Suomen kansalliset laitokset tekevät tiiviisti yhteistyötä keskenään ja myös laajemmin Euroopan mitassa. Artikkelissa esitellään lyhyesti Suomen kaksi laajaa kansallista hanketta, joissa työskennellään kieliresurssien paremman käytön edistämiseksi. Näissä hankkeissa, FIN-CLARINissa ja Kansallisessa digitaalisessa kirjastossa, on mukana myös Kotimaisten kielten tutkimuskeskus, joka vastaa kansallisesti merkittävistä kieliaineistoista.

Sammandrag

Språk är grunden för vårt kulturarv, och språksamlingar och arkiv utgör väsentlig infrastruktur för humanistisk forskning. Nuförtiden samarbetar de nationella institutionerna i Finland intensivt med varandra och med andra europeiska institutioner. I den här artikeln presenteras kort två omfattande finländska nationella projekt som båda har som mål att främja en bättre tillgång till språkresurser. De här två projekten, dvs. FIN-CLARIN och Det nationella digitala biblioteket, ingår också i arbetet på Forskningscentralen för de inhemska språken, som är den institution, som ansvarar för det nationellt viktiga språkmaterialet.

Languages are a basic part of cultural heritage and the language collections and archives are an essential research infrastructure for humanities. During the past decade Finnish society has become more aware of this matter. The state has provided funds for more work in this area than ever before and there is an organized and systematic cooperation of various institutes, building new ways to maintain and develop the culturally remarkable infrastructures.

At the moment, there are a number of projects working for the better use of language resources in Finland. This work means that there is intensive co-operation between national institutes within the larger European context. In this context, we will briefly describe two nationwide projects. The work in the Research Institute for the Languages of Finland is also, to some extent, involved with these projects, FIN-CLARIN and the National Digital Library.

First, we will briefly describe the mentioned projects and secondly, present the electronic databases of the Research Institute for the Languages of Finland.

1. FIN-CLARIN as a Finnish part of the European CLARIN

One of the priorities of European research policy is to develop research infrastructures. As mentioned in a number of chapters in this volume, the European Strategy Forum on Infrastructures (ESFRI) has drawn a roadmap for European research infrastructures. CLARIN (Common Language Resources and Technology Infrastructure) is one of the some 34 infrastructures chosen for the ESFRI roadmap. The goal of CLARIN is to provide access for all scholars to language materials and tools all over Europe.

In 2008, Finland's Ministry of Education and Culture provided funds for the Federation of Finnish Learned Societies for the mapping of research infrastructures at national level. This work concerned research infrastructures in all areas and, as a result, 20 projects were proposed for the roadmap of new infrastructures that are to be significantly developed. Thirteen of those projects are associated with European research infrastructures proposed by ESFRI. One of the associated projects is FIN-CLARIN (*Kansallisen tason infrastruktuurit: nykytila ja tiekartta* 2009), funded by the Ministry of Education and Culture as well as the Academy of Finland and the University of Helsinki.

FIN-CLARIN is the Finnish Language Resource Consortium and it is committed to building the Finnish language resource infrastructure and making it an integral part of the European CLARIN infrastructure. FIN-CLARIN as well as CLARIN will solve three problems which presently prevent the efficient use of existing language materials and tools: First, even if digital material exists, it is difficult to find out where the material is located. Hence, common metadata for various materials are needed. Secondly, even if the user finds the material, it is difficult to know how to get permission to use it. Hence, standardized licensing types and a common system for authorization and authentication are needed. Thirdly, even if the user gets the permission to use the material, the parts of it are in different formats and not compatible with each other or with the tools available. Hence, standardizing formats and interfaces are needed. (For more about FIN-CLARIN see http://www.ling.helsinki.fi/finclarin/ and https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN.)

Topics of FIN-CLARIN are: relevant standards and guidelines for resources; accessing and acquiring resources from other sources; collecting and using text, speech, lexical and other resources; developing methods and tools to better utilize and use the resources; intellectual property rights (IPR) of the resources and training and education of related topics.

The Finnish Language Resource Consortium will build its activities around the CSC (the IT Center for Science) Language Bank, which will be the main depository of resources, tools and knowledge. CSC is funded by the Finnish state. There are a number of participants in FIN-CLARIN, universities and the Research Institute for the Languages of Finland.

The preparatory phase that has been performed by CLARIN ended in 2010. FIN-CLARIN will continue its work by implementing the recommendations produced by CLARIN in close collaboration with the European META-NET project and its northern part META-NORD.

In addition, several other research projects operate in close collaboration with FIN-CLARIN, e.g., the Finnish Treebank project, the Finnish WordNet and the HFST project (*Helsinki Finite-State Transducer Technology*) at the Department of Modern Languages, University of Helsinki (www.ling.helsinki.fi/finclarin/intro.html).

To sum up, the main task of FIN-CLARIN is to make the use of materials easier for both researchers and laymen. The whole project is made from the users' point of view.

2. National Digital Library

The information society has radically changed the environment of the various types of collections: archives, libraries and museums. During the last decade, significant investments have been made in digitizing traditional collections and distributing the materials online. In addition, the organizations must accumulate existing digital materials. The main archives, libraries and museums have an obligation to preserve materials in digital format for a long period of time.

The National Digital Library (NDL) is implemented in the project launched by the Ministry of Education and Culture with 35 Finnish organizations, e.g. scientific and public libraries, museums, archives and other organizations and key interest groups. It is one of the key electric research and culture infrastructures currently under construction in Finland.

The aim of the project is to improve accessibility and long-term preservation (LTP) of the electronic materials of libraries, archives and museums. It offers new possibilities for information seeking and ensures that the information remains in active use in the future as well. The project also contributes to the European Union's objectives concerning the digitisation of cultural materials and scientific information as well as their digital availability and long-term preservation (www.kdk.fi/en).

There are four support services that should be provided within the scope of the NDL:

– permanent actionable identifiers of digital objects (such as URN identifiers);
– an authority database, i.e. a system that interconnects the names of persons and organizations in different languages and forms;
– maintenance of the standard portfolio;
– services related to competence development.
(*The National Digital Library – collaborating and interoperating* 2011, 14)

There are different types of material: publications, government publications, museum materials, archival documents and manuscripts, radio and TV programs, audio and video recordings of various types. Naturally, the rules and practices for describing these materials vary. Also the technical capabilities of the organizations and their present solutions for managing, using, distributing and preserving materials vary significantly.

The public interface gives access to the electronic information resources and services of libraries, archives and museums. The web service makes it easy to gain access to materials on any given subject matter, such as pictures, documents, newspapers, research, video and audio recordings. Aim is to launch the service in 2012.

Through the public interface, the materials of libraries, archives and museums come to form a whole. The public interface is intended to enable users to find the information they need through one interface, irrespective of which organization has produced the information. Hence, the information seeker no longer needs to know who owns or manages the materials. It is enough to want to gain knowledge or experiences. The information resources from various organizations can be found in one service and with one search. It helps information seekers not only to find the information they need but also other pertinent information. The user can also receive the electronic services connected with the materials from the same address.

Organizations will continue to be responsible for the production, cataloguing and management of their own digital resources. The public interface will facilitate access to the diverse resources of libraries, museums and archives for research, teaching and other information acquisition. Organizations will be able to customize the public interface for their own unique requirements and will also be able to create default views for different groups of users. (See www.kdk.fi/en/public-interface.)

The role of the long-term preservation sub-project (see e.g. Merenmies 2010) of the National Digital Library project is to coordinate the development of the long-term preservation of cultural heritage content data objects. The users of the centralized long-term preservation solution are primarily those responsible for the preservation of published and other kinds of material cultural heritage operating within the sector of the Ministry of Education and Culture, e.g. the National Archives, the National Library, and the Institute for the Languages of Finland.

The basis for designing the centralized long-term preservation solution is to offer archives, libraries and museums a system that is reliable, versatile and provides the required preservation services such as the controlled migration of the preserved content.

An expert group of long-term preservation in NDL was established in spring 2011. This includes representatives from archives, libraries and museums. CSC, the IT Center for Science, will be the organization responsible for the first phase of the long-term preservation system implementation project. This phase will be completed by the end of 2014. In the second phase of the project, which according to plans will start in 2015, will include a transition from the project phase into permanent system administration and full-scale use. The LTP system should be operational by 2016, at the earliest.

One of the main benefits of the NDL project is the unification of work processes, data structures and systems in archives, libraries and museums including the production and distribution of administrative metadata required for long-term preservation. Organizations will need to have a collection policy and clear operating principles guiding the compiling, management and preservation of digital collections within the framework of legislative and other obligations. The objective is that the national data resources will be widely available to the use of entire society. (See www.kdk2011.fi/images/stories/KDK_PAS_jarjestelma_metatiedot_v0.9.pdf.)

National Digital Library – Enterprise Architecture by the Finnish National Digital Library project steering group describes how the various elements – organizational units,

people, processes, information and information systems – relate to each other and function as a whole. Enterprise architecture is subdivided into four areas as follows:

– Business architecture: the project's services, stakeholders and processes;
– Data architecture: the key glossaries being used, the central information resources and the relationship between information categories and systems;
– Application architecture: the content of the information system portfolio;
– Technical architecture: the technology portfolio, reference architectures and interfaces.

A centralised long-term preservation solution for the digital materials will secure transitions between generations of systems, software and equipment, keeping digital information coherent and understandable for future users. Even in the long-term preservation system, the ownership of materials will remain with the organization who stored them. The system will be designed to allow the preservation of electronic data resources for research in the future.

The National Digital Library is the most extensive cooperation project between libraries, archives and museums so far in Finland. During the project, cooperation both between and within the library, archive and museum sectors has increased and intensified. (*Putting data into use* 2011; see also www.minedu.fi/OPM/Julkaisut/2011/Kansallinen_digitaalinen_kirjasto.html?lang=fi&extra_locale=en.)

4. Electronic archives and collections held by the Research Institute for the Languages of Finland

The Archives and Collections of Linguistic Corpora and Collections of Electronic Linguistic Corpora of the Research Institute for the Languages of Finland belong to the national infrastructures (*Kansallisen tason tutkimusinfrastruktuurit. Nykytila ja tiekartta* 2009, 26). As mentioned above, the institute is one participant of both projects, FIN-CLARIN and National Digital Library.

The extensive archives and collections (www.kotus.fi/collections) held by the Research Institute for the Languages of Finland have been assembled over more than a century. Besides the material held in paper form, there are also audio and video recordings and an ever increasing volume of electronic data. An on-line data service named ‘Kaino’ was launched in December 2006. This includes Finnish texts dating back as far as the 1500s as well as a separate Atlas of Place Names and etymological data on the Saami languages (Álgu – Origins of Saami Words). The Finland-Swedish data is mostly available in the electronic and manual archives of Svenska Litteratursällskapet i Finland (The Society of Swedish Literature in Finland) or in the data bank of the University of Gothenburg in Sweden.

At present, there is a number of electronic data in the Research Institute for the Languages in Finland (http://kaino.kotus.fi/korpus). A large part of the data is available for everybody. The electronic freely accessible on-line data service is 2011 as follows:

Finnish:

- Corpus of Old Literary Finnish 1543-1809;
- Corpus of Early Modern Finnish 1809-1899;
- Modern Finnish Lexicon by Research Institute for the Languages of Finland;
- Corpus of Finnish Literary Classics 1880s-1930s;
- Corpus of Proverbs and Other Colloquial Expressions;
- New Year Speeches of the President of the Republic of Finland 1935-2007;
- Etymological Reference Database;
- Atlas of Place Names;
- Toponymic Database, 162,774 place names;
- Collection Database of Audio Recordings Archive.

Languages related to Finnish:

- Álgu – Origins of Saami Words;
- Etymological Reference Database;
- Dictionary of Karelian;
- Vepsian Word List.

The freely accessible on-line data service includes other materials, and it increases continuously.

Another part of the electronic data requires user authorization. It includes materials as follows:

Finnish:

- Corpus of the Finnish Language = Finnish Text Collection (access via CSC, Language Bank), contains written Finnish from 1990s;
- Corpus of Magazines and Periodicals 20th century;
- Syntax Archive Data: The data is owned by the Research Institute for the Languages in Finland and the School of Languages and Translation Studies (Finnish Language) at the University of Turku. The Syntax Archive Data contains dialects from 132 Finnish parishes (one hour from each parish) and literary Finnish (40 units);
- Lexical Data from the Archive of Modern Finnish;
- Headwords in the Dictionary of Modern Finnish (= Nykysuomen sanakirja 1-6, 1951-1961);
- Oulu Corpus, a representative sample of the Finnish language in the 1960s media (access via CSC, Language Bank, www.csc.fi/english/research/software/oulu);
- Texts from the Samples of Finnish Dialects Collection, text and audio (access via CSC);
- Corpus of Entries from the Dictionary of Finnish Dialects.

Finland Swedish:

– Finland Swedish Text Corpus = Finnish-Swedish Text Collection 1997-2000 (access via CSC, Language Bank),
– Swedish-Finnish Parallel Text Corpus 21th century (CSC, Language Bank).

The Research Institute for the Languages of Finland has also other types of digital data, which requires user authorization: audio and video recordings, manuscripts and photos.

The list of the databases of the Research Institute for the Languages of Finland will increase and develop in the future. This work is funded in the frame of the budget of the institute, but, however, it is supported by the experts in the national projects, FIN-CLARIN and National Digital Library.

5. References

Kansallisen tason infrastruktuurit: nykytila ja tiekartta. (= Opetusministeriön julkaisuja 2009:1). Helsinki: Opetusministeriö. www.minedu.fi/OPM/Julkaisut/2009/Kansallisen_tason_tutkimusinfrastruktuurit._Nykytila_ja_tiekartta.html?lang=fi&extra_locale=en.

Merenmies, M. (2010): *The National Digital Library Initiative Long-term Preservation Project. Final Report*. *8th European Conference on Digital Archiving, Geneva, 28-30 April 2010*. www.kdk.fi/en.

Putting data into use. A roadmap for the utilization of electronic data in research. (= Reports of the Ministry of Education and Culture, Finland 2011:4). Helsinki: Opetus- ja kulttuuriministeriö.

The National Digital Library – collaborating and interoperating. (= Publications of the Ministry of Education and Culture 2011:26). Helsinki: Opetus- ja kulttuuriministeriö. www.minedu.fi/OPM/Julkaisut/2011/Kansallinen_digitaalinen_kirjasto.html?lang=fi&extra_ locale=en.