

Summary

The objective of the MIDRAS project for remote access has been to investigate and clarify how register-based research can be supported and the availability of register-based data sets can be enhanced. The project, funded by the Ministry of Education and Culture, was executed by CSC-IT Scientific Computing Center and the Finnish Information Centre for Register Research (ReTki). The proposed operational model will facilitate and diversify the means of register-based research, provide a new technical solution to support research, and improve access and reuse of information collected with public funding. In particular, the operational model will improve the availability and linking of confidential sensitive data sets from several authorities. In this report the abbreviation MIDRAS (Micro Data Remote Access System) is used for the remote access system.

This report shall, in accordance with the objectives of the project, focus on the promoting of research using confidential unit level data (microdata). However, the MIDRAS system can be used as well for disseminating other types of data sets, and more generally as a research and analysis platform for e.g. reports compiled by the authorities.

The MIDRAS system in response to present problems facing register-based research

In this report, register-based data means micro data that is owned by authorities, has been collected for administrative and statistical purposes using public funds, and the use of which are subject to authorization.

Register-based data sets are important data sources for research, but there are currently many problems and challenges in their use for research. From the researcher's perspective the problems are, in particular:

- A lot of time is taken by the authorities for granting permits and for compiling the research project specific data sets.
- There are many public authorities that keep registers, and the practices and laws pertaining to the delivery of the data vary.
- From the researcher's point of view, there are shortcomings in the register descriptions and the metadata of the data sets.
- It is not always possible to give out the best research data for the planned research because of the requirement for anonymisation that is given by the Statistics Act.
- The price of some data sets is very high.

- If the data passes through many authorities before being delivered to research, the data may be already outdated.

From the perspective of the authorities, the key challenges for the promotion of the use of register data are:

- Research service is not the primary task of the authorities, so the resources for this are scarce.
- The requirement for anonymization (set by the Statistics Act) adds to the work load for compiling research data sets.
- The current delivery methods of research data sets (CD or Flash drive) induce security risks and it is not possible to monitor the use and disposal of the data sets.
- Research has not been taken into account in data set documentation, production and storage.
- The content and storage format of the metadata is not standardized.
- Researchers have poor knowledge of the data sets, so they hand in incomplete research plans and permit applications.

The MIDRAS operational model offers clear improvements compared to the current situation (figures 1 and 2): it will enhance and streamline the processes of using register-based data for research and thus it will save both researchers' and authorities' time and effort. The MIDRAS operational model supports a broader and more diverse use of data, offers a technically novel solution for supporting research, and substantially improves data security and data protection of data that has not been anonymized.

Figure 1 Current situation, where the researcher applies for permits and data sets from different authorities

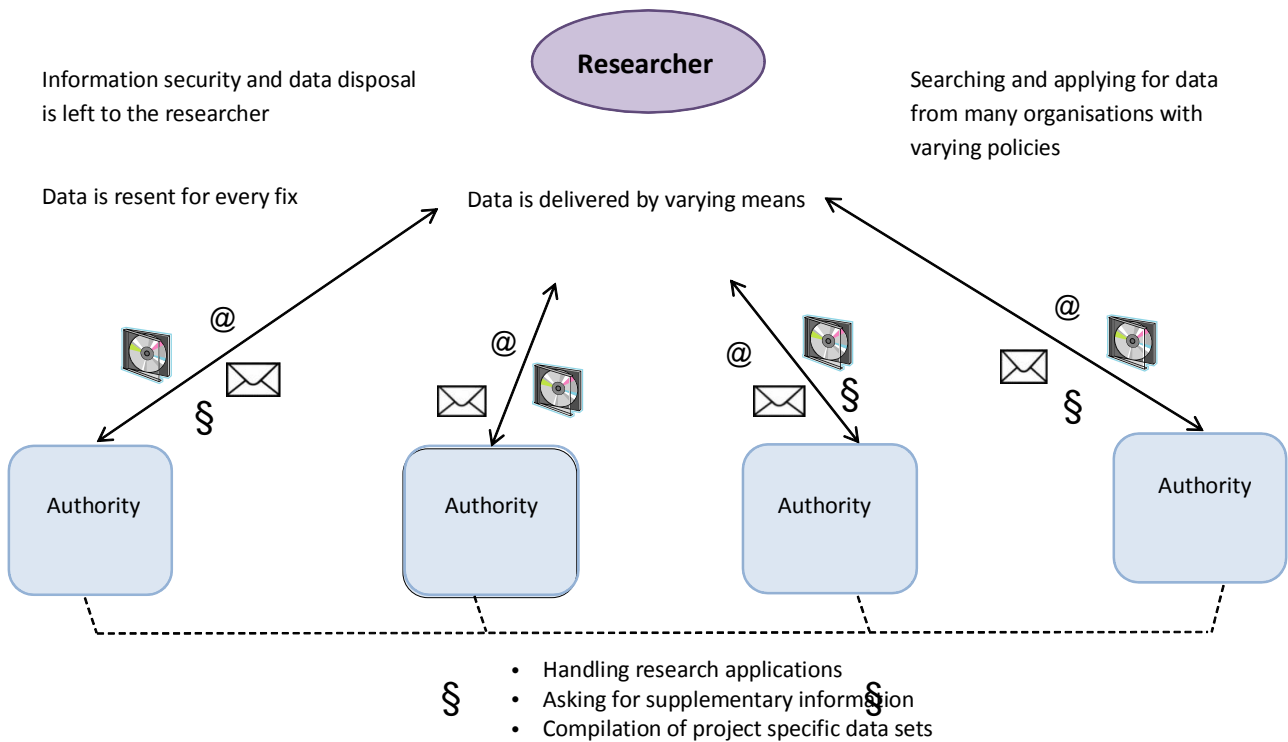
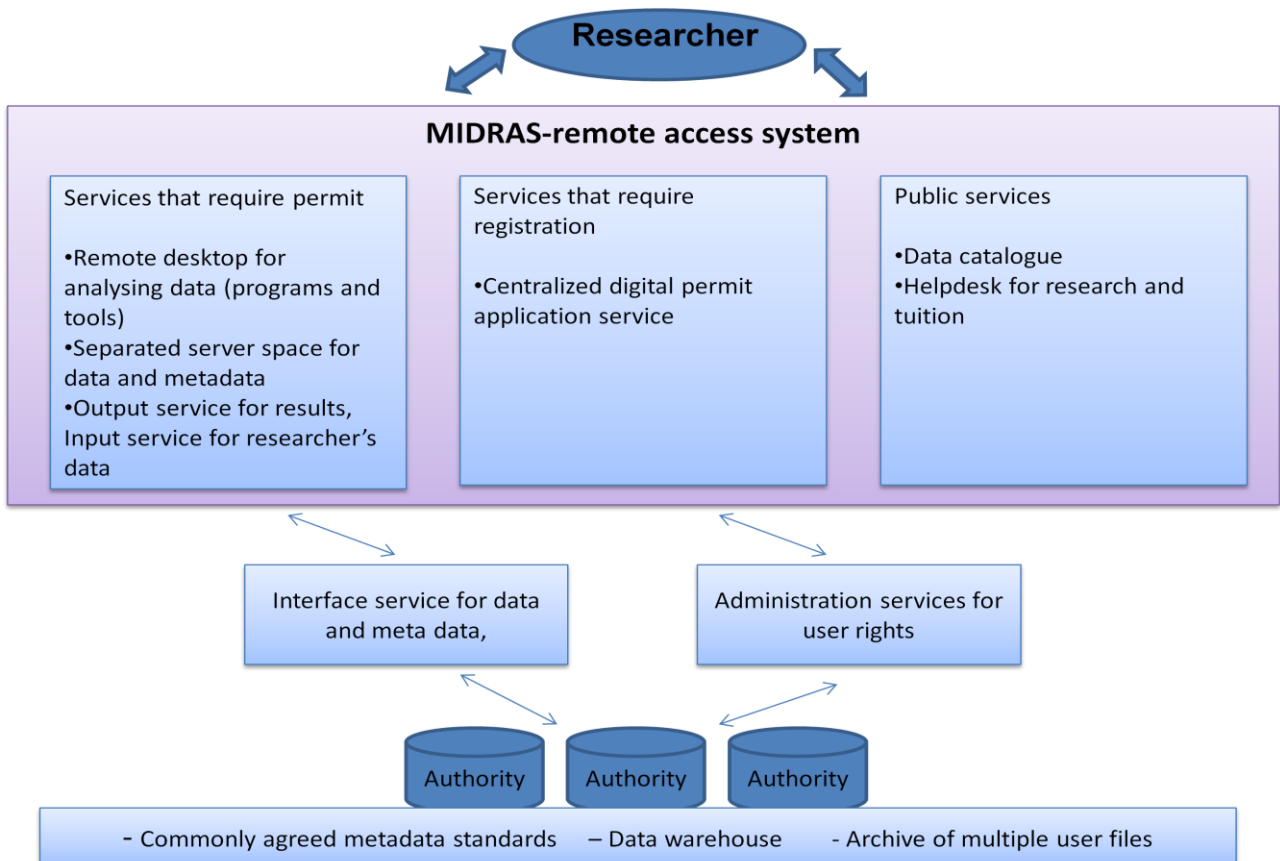


Figure 2 Depiction of the MIDRAS-operation model and services according to the vision



The MIDRAS system works as a channel that can be used for disseminating licensed register-based research data, especially sensitive unit level data, compiled from official registers for research purposes.

From the view point of researchers, the core services of the proposed MIDRAS system are:

1. A restricted remote desktop environment, including services and applications for processing and analyzing research data sets. The researcher cannot download the research data, but the data will be made available on a remote desktop, which the researcher can log on to from his own desktop.
2. A catalogue over available data sources that makes it easier to apply for research permits and to plan the research, and a centralized research application service based on the catalogue. The data catalogue is a web service where metadata on available datasets is collected in a common form.

These are the core services for the authorities who produce the data:

1. Data dissemination interface between the MIDRAS maintenance organization and the authorities. The system's data sets, metadata and other data are delivered through the interface.
2. A tool for managing research permits and access authorization. The access management tool allows administrators to grant permissions, to manage and to view which research projects and researchers have access to which data.

Vision

MIDRAS will provide a national remote access system through which the authorities' micro data of restricted use can be used for research in a comprehensive, secure, cost-effective and easy way

The MIDRAS operational model creates common practices and recommendations that harmonize and streamline the research process. As a result, researchers will invent new ways to exploit old and produce new knowledge. The MIDRAS system provides a good and versatile working environment, a catalogue that makes it easier to find data, and a comprehensive collection of data sets in one system. The solutions of the MIDRAS system facilitate and speed up the authorities' work in analyzing, planning and reporting. The new tools make it possible to leave the analyzing and processing of the data sets for researchers, and only the straightforward tasks, which can be automated, will remain the responsibility of the owners of the data. The work of both the researcher and the data producers is reduced by automation, the unified description of the data sets and the research application service based on these descriptions.

Proposals for the construction and implementation of the remote access system

The experience in the project shows that all parties feel that there is a need for a fully usable remote access system for register data in Finland. We propose that:

1. An agreement will be reached in 2011 on the management model and financing of MIDRAS.
2. Of the services provided by the system, the data catalogue and electronic research application form will be implemented first.
3. The pseudonymization model will be decided together with interest groups during year 2011. The pseudonymization model may not, however, be used for data sets affected by the current Statistical Act
4. Data sets that are to be used through the system will be introduced in phases starting from year 2012. From the researcher's point of view, some appropriate data sets are the care register (hospital discharge register), the birth register, the cause of death register, the drug reimbursement register, the data set on regional employment and the longitudinal census file.