OECD Case study report RUT

Magnus Eriksson
Business Architect

Maria Nilsson
Head of Unit for register-based research
**Overview of the initiative**

<table>
<thead>
<tr>
<th>Name of initiative</th>
<th>Infrastructures for Register-based Research – a government commission to the Swedish Research Council</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective</td>
<td>Increase the value of Swedish registers, research databases, sample collections and research datasets. To do this by providing an infrastructure to make knowledge regarding the content of the above mentioned data sources accessible through rich metadata and semantic descriptions, and by providing opportunities for reuse of this knowledge throughout the research process.</td>
</tr>
<tr>
<td>Type (strategy, policy, bill of law,…)</td>
<td>Government commission</td>
</tr>
<tr>
<td>Responsible policy making bodies</td>
<td>Ministry of Education and Research</td>
</tr>
<tr>
<td>Responsible implementing bodies</td>
<td>The Swedish Research Council</td>
</tr>
<tr>
<td>Target audience</td>
<td>Scientific Community, including higher education institutions and businesses conducting research based on register data</td>
</tr>
<tr>
<td>Total duration of initiative</td>
<td>Initially a 4 yr project but now a permanent commission</td>
</tr>
<tr>
<td>Total budget of initiative</td>
<td>5 M EURO (annual budget)</td>
</tr>
<tr>
<td>Sectoral focus</td>
<td>Register-based research spanning across several sectors and research areas</td>
</tr>
<tr>
<td>Type of data concerned</td>
<td>Public Sector Information, Research Data</td>
</tr>
<tr>
<td>Expected Results</td>
<td>Reduced time for data access to register data for researchers, High quality rich metadata and semantics that supports communication of knowledge and reuse – increased data FAIRness</td>
</tr>
</tbody>
</table>
Rationale, motives and key drivers

Background to the commission
In the 2012 Government Research and Innovation bill, the Swedish Research Council (SRC) received a commission to establish operations within the authority to improve access to register data for research purposes and to facilitate the use thereof.

The commission was named “Infrastructures for register-based research.” And had a background in a state public report regarding register data for research (SOU 2012:36).

The SRC ought to also provide researchers with information on registers and relevant legislation and establish a national entry point for register-based research. The task was initially allocated for the period of 2013-2016 and listed in the annual appropriations directions to SRC.

Motives for initiating the initiative
The Swedish authorities have a long tradition of collecting data in national registers for statistical purposes and the health care system use registers to improve quality in health care. The registers also includes the use of a unique national identifier for persons registered in Sweden that are allocated by the Swedish tax authority Skatteverket. This “personal number” acts as the common link between registers.

The data collected in registers provide a valuable source for transdisciplinary research in many different areas including socioeconomic and public health research. The registers have sometimes been described as goldmines for research but in order to utilise the resources that the register data comprises the process of finding, getting access to, integrating and reusing need to be more effective and less time consuming than it is today.

The concept of register data was set to have a wide scope and besides registers, data from sample collections in biobanks, research database and research datasets was included. That is there are several kinds of register holders, the organisations that keeps and may give access to register data for research.
The time it takes for a researcher to formulate a research question to get access to the desired data may extend over very long time (months, sometimes years). Since most of this data are sensitive the researchers cannot get access to the data before applying for ethical approval in Sweden. After the ethical approval the register holder also perform a harms test to evaluate the sensitivity of the data, the scope of the requested data and how the data will be protected after disclosure. It is important that the request for data is in line with the ethical approval, and that you have understood the meaning of the variables correctly.

One reason why the process is so long is that the register contents is not easily accessible online and there is very little and/or low quality information on variables, populations and collection procedures that supports the researchers request. The dialogue between the researcher and register holder is therefore not always straight forward due to this lack of access to information. An infrastructure that separates metadata from data make it possible for the researchers to evaluate the register content before getting access to the data and will allow a sharper and more correct request for data.

Objectives and expected results
The vision is to increase the value of Swedish registers, research databases, sample collections and research datasets by shortening the researchers time to desired data and enabling reuse.

In order to make this possible the researchers need to find descriptions of the data and its meaning and the register owners need to be able to communicate the knowledge of the content. We enable this by providing an infrastructure to make knowledge regarding the content accessible through rich metadata and semantic descriptions, and by providing...
opportunities for reuse of this knowledge throughout the research process. Another important part of the enabling process is to provide a method for metadata management together with metadata curation support and training.

International references/standards
When initiating the project an evaluation of several statistical metadata standards and frameworks took place. The decision was to use the Generic Statistical Information Model (GSIM), developed by The High-Level Group for the Modernisation of Official Statistics where OECD, Eurostat and UNECE is members.

When harvesting metadata into the GSIM metadata framework we also make sure that the GSIM principle of separating meaning from representation is fulfilled. The descriptions of the meaning of the data in the actual context is described through concepts holding definitions and concept system graphs providing meaning regarding the context.

Using this metadata we are able to provide findability through semantic search of variables using the concept systems and promote increased interoperability through the infrastructures graph knowledge representation.

The advantage of the GSIM is that it specifies what types of metadata should be included from a register in order to sufficiently describe its detailed contents. The GSIM does not specify what actual variables should be included in a register or in a domain standard but rather that the variables included are described with certain types of metadata. The GSIM also provides a common terminology and concept system describing the types of metadata that increase the semantic interoperability between domain standards and registers.

There is therefore no need for the register holders to agree on domain standards or harmonisation of variables and/or concepts. Instead they need only agree on how they should describe their contents, i.e., using what types of metadata. Since the register domains that RUT supports range from different kinds of administrative registers, health care quality registers to biobank sample collections and research databases, this generic approach is necessary in order to incorporate all domains with their different requirements.
**Governance**

As the commission was complex and initially time-limited, it was initially managed according to the SRC’s project model. The project was governed by an internal steering group with members representing high-level expertise in register-based research, IT and research infrastructures. The steering group is chaired by the Director General. In 2015, the Unit for Register-based Research was established and given the responsibility to coordinate activities related to the topic. As of 2018, this task has been included in the SRC’s instructions and is considered permanent.

In order to ensure that principally important decisions within the project receive the support of relevant authorities and organisations, the SRC board established the **Register Data Council** in 2014. This advisory body consists of researchers and senior executives representing the major register holders and other key players. The council is chaired by the Director General of the SRC. The Register Data Council’s mandate was initially time-limited but in 2017 it was extended, and is now permanent. The council meets regularly 2-3 times per year.

By assembling a reference group with researchers from different research domains and with different experience levels we got a forum for capturing the existing research based research process, the bottlenecks experienced and the major problems encountered during the different activities in the process.

The reference group also provided the project with requirements on what types of metadata and semantics they needed in different stages and could steer us in the right direction. This was used as a foundation for the evaluation of statistical metadata standards and frameworks and the cross domain contribution in the group made it clear that we needed a quite generic standard in order to manage metadata from the different domains.

The other stakeholders, that is the register owners, was as previously mentioned represented on the highest level in the Register Data Council. In order to set up the cooperation between the actual registers and the Swedish Research Council agreements on participation was made with each register owners organisation. Participation principles and activities was agreed upon and are functioning as a good base for the joint activities.
### Process

When the project started, it was well-known that register-based research is complex, expensive, and extremely time-consuming. A thorough mapping of the register-based research landscape was therefore conducted in order to understand the bottlenecks and hindrances that researchers encounter in their projects.

A general research process\(^1\) was identified based on interviews with researchers, register holders and other relevant actors including e.g., members of ethics review boards. Both the reference group of researchers and the register holders stressed the lack of a central entry point where information on register-based research and contents of Swedish registers at the variable level could be easily accessible. In order to decrease the lead-time, the most time-consuming steps of the research process were identified.

![Diagram of register-based research process](image)

**Fig 2 – Subset of identified register based research process. Early activities containing bottlenecks.**

From a register holder perspective, the researchers need to define which variables are needed to answer a specific research question since the register holder cannot give access to more data than is strictly needed to carry out the research project. The researcher sometimes applies for more data than necessary and has little or no access to information regarding the population and variable meaning in the register context, the representation used, and how this has changed over time. The dialogue between the researcher and the register holder could therefore become lengthy. From both perspectives, the lead-time would decrease if registers and variables could be easily identified and matched against the research question.

\(^1\) [https://www.registerforskning.se/en/](https://www.registerforskning.se/en/)
In order to be able to provide support to the researchers early in the process, and to contribute in resolving the bottlenecks identified early in the process, the project decided that the infrastructure would build on a principle separating the data from the non-sensitive metadata and semantics.

![Diagram of metadata and semantics separation](image)

**Fig 3 – Separation of metadata and semantics from data.**

Based on these findings and a requirements analysis, the SRC decided to collect the information relevant to register-based research on the website registerforskning.se and to develop an online metadata tool, Register Utiliser Tool (RUT short, in Swedish), in order to describe registers’ contents at the variable level. A vision was set of a national entry point for register-based research that would allow the researcher to have access to the tools needed in her research, which also includes the possibility to apply for data. The vision has received the support of the Register Data Council.

The question of harmonisation, to what extent harmonisation work should take place and who should own definitions of concepts used rose quite early. The reference groups input provided some guidance on the question since they explained that whether or not a source variable, the concept used to define it’s meaning and the representation used, was possible to harmonise into a study variable was totally dependent on the study in question. The researchers would then benefit mostly from a thorough description of a variable with rich metadata and semantics in order for them to make comparisons between variables and form a judgement on harmonisation possibilities.

This input led to the strategic decision to assign responsibility for metadata and semantics to the register owners organisation and that the infrastructure would not engage in harmonisation activities between the register owners. That said the project’s work on method and training together with the infrastructures application support aims to promote usage of standardised
terminologies, ontologies and classifications and to make it easy to chose from standards if appropriate when working with the metadata and semantics. The choice and decisions are however still made by the register owners organisations.

Fig 3 – Register owners organisation decide on concepts and metadata to be used to provide rich descriptions of the register contents.
Adoption and implementation of the initiative

Launch of a national entry point
The website registerforskning.se was launched early 2016 and strives to eventually serve as a national entry point for Swedish register-based research. Its purpose is to provide the research community with information on how register data can support research and to support the researchers during the planning of a project that includes register data.

On the website, the researcher receives information about existing registers, as well as support in each part of the process of identifying, requesting and using register data. The website provides information and insights into the legal requirements and the processes and activities associated with them and also function as an entry point to the Register Utiliser Tool.

The website have just undergone a graphic revision. An English version of the website, registerresearch.se, is launched summer 2018.

The metadata tool RUT
In order to meet the research community's need for qualified information on the contents of Swedish registers at the variable level, the SRC decided to develop a web-based metadata tool – the Register Utiliser Tool (RUT). The decision received the support of the reference group of researchers as well as the Register Data Council and the SRC’s legal advisers.

RUT’s objective is not only to allow the researcher to search for variables but also to support the researcher during the entire research process. This involves harvesting register metadata at the variable level as close to the source as possible (i.e., from the register the variables are located in), making it available throughout the research process and also allowing contributors to be credited in the later stages for their work by storing and displaying related provenance metadata. Provenance metadata describes which samples, source datasets, methods, etc. have been used in the research process as well as their origin.

This gives the opportunity to give credit to sources such as sample collections or registers and to easily find and display them. RUT will thus support increased findability, interoperability and re-use of the included registers and research datasets.

In addition, RUT will support the register holders in their management of researcher requests for data by providing support during the disclosure process. The search result and increased understanding of the detailed content could support the communication between the researcher and the register holder and provide a distinct list of variables. Such a shortened dialogue will decrease the lead-time until the data has been disclosed, a procedure that often takes months.

RUT as a tool to support data management according to the FAIR principles
Having included their metadata into RUT, some register holders have also expressed an interest in using the RUT metadata repository in their own data management ecosystem. RUT will thus facilitate the register owners’ daily work with use and management of the metadata curated into the repository. This is an unforeseen user case, and the SRC is therefore looking into which application programming interfaces (APIs) are needed to provide this service.
The RUT metadata repository supports and utilises the results of good data management practices in administrative registers, Clinical Quality Registers, biobank sample collections and major research databases. In the future RUT should also include research datasets produced in research utilising personal data with the help of SND and the networks of Data Access Units (DAUs) that are being built at Swedish universities and higher education institutes. In line with the SRC’s government commission to coordinate national work on open access to research data, RUT aims to support data management throughout the whole register-based research process.

The use of a standardised metadata framework based on the GSIM that the registers and researchers can use to create FAIR compliant registers and research datasets. Such findable and accessible metadata will have strong support for interoperability and will thus be the foundation of efficient reusability. To increase the metadata support throughout the research process and enhance reusability, the existing GSIM implementation is being complemented with an implementation of the W3C provenance\(^4\) standard. This is currently being implemented in an additional RUT component named “RUT Researcher Area” (RUT RA) that will enable the researchers to further re-use metadata harvested from the registers.

One objective of RUT RA is to enable re-used register metadata to be supplemented with the researcher’s own additional curated metadata from research datasets. This will be a starting point for describing the entities, activities and people involved in producing the study dataset that is used for publications. This not only increases the opportunities for and quality of re-used data but also provides a mechanism where used datasets, samples, methods and documents can be attributed to their sources. This is in line with the vision according to which RUT provides support during the whole research process. The requirements for the RUT RA are currently being collected and processed.
**International aspects**

The work with RUT has drawn some attention in the neighbouring Nordic countries and members of the project have taken part in NordForsk funded Nordic collaboration projects regarding provision of common access to metadata describing Nordic register data resources.

One outcome of these collaborations have been the adoption of or mapping to the GSIM standard framework in the Nordic statistical institutes (NSI) which increases the interoperability on a Nordic scale. Also, there is interest from some Nordic NSIs to visualise some register content in the RUT system. Another outcome is the contribution of experiences into a Nordic health data collaboration project.

One of the most important lessons learned and contributed to the Nordic community is the importance of semantic interoperability and the experience and added value of working closely with domain experts such as medical doctors, statistical subject area specialists and biobank specialists to create knowledge representations of the register content.

The project have also participated in UNECE hosted workshops on Implementing Standards for Statistical Modernisation and contributed with presentation of the projects practical experiences of GSIM and thoughts on development related to this.
**Monitoring and evaluation**

The commission to facilitate register-based research in Sweden is managed by the SRC department for research infrastructures, unit for register-based research. The work is reported annually to the ministry in the SRC annual report.

The development of the metadata tool RUT is run as a project according to the SRC project model. Different components of the tool is being developed and will be released during the coming years. In terms of contents, new registers are continuously being visualised in RUT. The aim is to have the registers that are most used in research in RUT over the coming years.

The planned activities are described in the operations planning and followed up during the year.

The work is anchored in the Register Data Council and monitored by a steering group with representatives from the SRC management. Depending on the development and interest from national and international register holders, the governance structure may need to be developed in the future.

**Lessons and Challenges ahead**

**Outlook**

The SRC will continue the development of infrastructures for register-based research including registerforskning.se and RUT, and work towards the vision of one national entry point. The RUT system will in the future be a tool for finding, evaluating, accessing, re-using and administering metadata that describes variables in population-based registers and research datasets. Since metadata in RUT is separate from the actual microdata, RUT can potentially act as a metadata hub in relation to many of the surrounding infrastructures in the national ecosystem when it comes to population-based data. One such surrounding infrastructure, SND, has expressed interest in making use of RUT both as a tool to support the Data Access Units (DAU’s) work within metadata curation for relevant datasets and for enabling re-use by harvesting dataset level metadata to SND’s own platform.

The SRC foresees an increased interest from Swedish register holders to include their registers in RUT due to the heightened focus on good data management. For example, an upcoming initiative regarding the National Quality Registers will aim to increase the quality of data management. In this context, the SRC’s main objective in the coming years is to include registers that are most frequently used in research. In order to increase the inclusion speed of new registers, focus is currently placed on the development of educational material alongside the RUT Management tool, so that the register holders could have a greater role in the modelling procedure.

The SRC’s vision of one national entry point for register-based research aims to facilitate each step in the process of data disclosure and data management in line with the FAIR principles. The SRC, Statistics Sweden, and the National Board of Health and Welfare are currently discussing the need for technical solutions for data disclosure that also enable linking register data with big data. There is great potential to expand this dialogue and to include other actors in the e-infrastructure ecosystem.
Lessons learned

**Competence** - The work providing metadata and semantics to describe the register content have proved to span over several competences that contribute to the creation and usage of the register resources. One lesson learned is the importance of involving representatives from all involved competences at the register owners organisation since they each provide an important contribution to the knowledge that need to be expressed with metadata and semantic descriptions in order to support understanding of the content.

**Understanding the domain** \(^2\) – Modelling concept systems that describe the semantics of the concepts used in the register context is fundamental to understanding the content. One lesson learned is that domain experts such as medical doctors have viewed concept system modelling as a easy to use tool to describe the meaning of the data they collect without needing to delve into technical details or complex variable names.

**GSIM Statistical framework provides good value** – Different registers and register types have different strengths and weaknesses when it comes to metadata management. The level of metadata differs, the level of semantic descriptions differs, and the level of history management differs, but GSIM acts as an effective common language of metadata and provides a framework for identifying what parts of the metadata and semantic descriptions need to be further curated and which parts can be harvested and used as is.

Furthermore the different domain specific standards that exists specifying what variables that should be included, what terminologies/ontologies should be used and so generally maps well to GSIM and the generic nature of the standard. It provides a common context while in the same time accepting the differences that exists between domains and registers when it comes to the content selected to be included by the register owners.

**Challenges ahead**

Challenges related to the RUT system is to deliver the planned components in due time and to populate the metadata tool with rich high quality metadata content. It is not until there are a critical mass of registers that the tool is of interest for a wider mass. Another critical point is that the register holders should be able to handle requests for data made via RUT in order to build the national entrance for register-based research.

**Training** -

The SRC business architects provides training on data management to the register owners. The register owners investment in developing their data management skills and curating their datasets make their dependency on specific individuals less critical since the knowledge about data content become a knowledge of the organisation instead of an individual. This is

\(^2\) “an area of interest or an area over which a person has control” (https://dictionary.cambridge.org/dictionary/english/domain)
important since building competence on the register content and meaning are sometimes a time consuming activity in the absence of rich metadata and semantic descriptions.