Instant Access to Microdata – microdata.no

Johan Heldal (johan.heldal@ssb.no)[1], Svein Johansen (svein.johansen@ssb.no)[2], Ørnulf Risnes (ornulf.risnes@nsd.uib.no)[3]

Keywords: micro-data access, architecture, metadata, data protection, disclosure control

1. INTRODUCTION

Norway has a large number of registers on individuals that have been established for administrative and statistical purposes, covering the entire population or significant subpopulations. The merged registers are used for production of statistics and represent a valuable source of data for research. Trusted researchers in approved research institutions have been able to apply for access to the data at their own site. The approval procedure is complicated as well as time- and resource demanding. There has been is a desire to simplify the procedure and at the same time make it safer through remote access and other measures. The entering into force of GDPR this May, makes the process even more comprehensive.

In 2012 the Norwegian Research Council funded a (then approx. four million €) project Remote Access Infrastructure for Register Data (RAIRD) aiming at creating an analysis server for easier and safe remote access to register data. The project is a joint venture where the grant was divided equally between NSD – Norwegian Centre for Research Data and Statistics Norway. Among the conditions for the project were

1. Online Remote Access (RA)
2. Micro data are invisible, only statistical output will show.
3. Users should be allowed to combine data from different sources.
4. All statistical results should be confidentially safe.

In March this year the RAIRD technology was made operative in the research data service *microdata.no*.

Section 2 will sketch the data structure, the metadata structure, and user interface. Section 3 will deal with security and confidentiality and section 4 will present experiences and plans for the future.

2. ARCHITECTURE, DATA AND METADATA MODELS

2.1. System architecture

The RAIRD system architecture was developed to provide researchers with rich, intuitive, efficient and flexible tools, in a managed executional context that prevents unwanted data disclosure. The ambition is to create an environment for researchers to work safely with personal data in an informed, autonomous and ergonomic way.

All end-user transactions (e.g. analytical or data transformation command) are subject to automated introspection, monitoring and disclosure control before the result of the transaction is returned. It is the infrastructure layout, autonomous and specialized

[1] Statistics Norway, Akersveien 26, N-0177 Oslo, Norway

[2] Statistics Norway, Oterveien 23, N-2211 Kongsvinger, Norway

[3] NSD - Norwegian Centre for Research Data, Harald Hårfagres gate 29, N-5007 Bergen, Norway

components, data- and metadata models that together enable the desired characteristics of microdata.no.

2.2. Data

When microdata.no was launched, the system contained 124 variables and 10 million units. Most variables are either event-data or have other temporal components.

Data is organized on a variable-by-variable basis, where each of the 124 variables is contextualized with encrypted unit identifiers and temporal information. As such, each contextualized variable is self-contained, and has sufficient information to be merged with other variables that have equivalent identifier domains.

Because of the temporal information associated with every record in every variable, this structure also supports data extraction based upon temporal criteria.

The initial ideas behind the data organizing are described in RAIRD Information Model [1].

2.3. Metadata

Metadata plays several roles in RAIRD/microdata.no;

- It informs users about definitions, data types, temporal nature and codes
- It drives most technical components; data may only be accessed *through* metadata
- It is used to assist users interactively when formulating scripts in the user interface

Each of the 124 contextualized variables (i.e. long, narrow data sets) is associated with detailed metadata records according to a metadata model developed during the RAIRD project. The model is informed and inspired by conceptual models from GSIM [2] and DDI [3], both of whom have terminologies and constructions that support both informative and technical metadata.

Furthermore, both GSIM and DDI have developed specialized models that address the relationship between a variable's measure component (substantive content), its identifier component(s) and its attribute(s) (e.g. temporal attributes).

The models were extended, e.g. to enable support for code lists that evolve over time. Users can inspect code evolutions in various parts of the user interface and use this information to inform their data transformations and analytical interpretation.

Content-wise, metadata in microdata.no builds upon several of SN's metadata management systems, including the relatively new classification system KLASS [4].

2.4. User interface

Microdata.no has two main end-user-interfaces;

- a variable catalogue available for all users (including non-registered users)
- a tool for working with and analysing data available for accredited users only

Both interfaces show the complete metadata for all variables, but in order to work with data (in command-line or script-mode), users need to be registered by their accredited home institution. The analytical environment includes support for data extraction from the sources into the user's virtual workspace, data transformations and merging - as well as a wide range

of analytical commands. In version 1.0 the user interface is in Norwegian only. An English version is planned in coming launches.

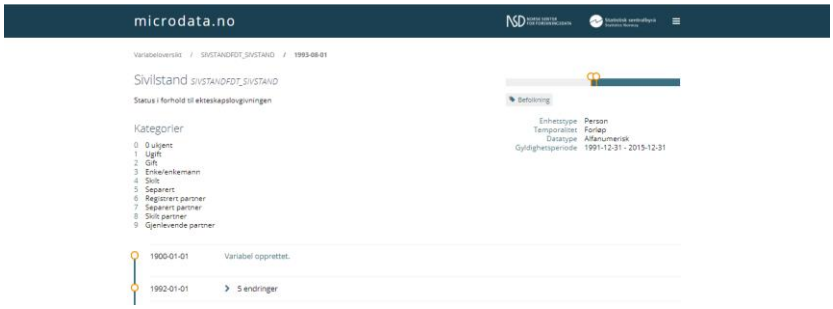

Figure 1. Screenshot from the variable catalogue

3. SECURITY AND CONFIDENTIALITY

3.1. Data Access

microdata.no is a service provided to master-/PhD-students and researchers at universities and research institutions. Each institution enters into an agreement of use with NSD and SN, giving them the right to sign in students and research fellows. Both the contract and the user administration are digital and self-administered by the institution.

Accredited users access through Norway's national digital log in solution to public services; *ID-porten* [5]. The solution requires a permanent or temporary Norwegian national identity number. Later versions will include international log in. At their first log in, the users accept the terms of use, e.g. not to attempt any form of data download, or use the service for other purposes than research.

During a two years' start-up phase, the use of microdata.no is free of charge. After that the institutions will be charged a fixed subscription fee and variable costs related to their actual use of the service.

3.2. Disclosure Control

Users of microdata.no will be able to edit and manipulate microdata in different ways. Therefore, even though individual records are invisible for the users of microdata.no, there would be ways for a user to disclose individual information if nothing had been done to prevent it. Primary methods for disclosure control are

- Minimum size of populations to be analysed (at least 1000 individuals)
- Noise addition on counts (max ±5) and on numerical aggregates. The noise is constant. It will always be the same when the same total is requested in the same population setting.
- Winsorization of all numerical variables with 1 per cent at each end of the distribution at every subsetting of the population.

The constant noise technology is inspired from the Australian Table Builder and Data Analyser [6]. The kind of noise used is not the kind suggested by Differential Privacy theory [7]. DP-noise would have caused too much information loss and would have harmed the utility of the results too much. Instead, winsorization was used to avoid the influence of extreme values.

All activity on microdata.no is logged. This enables NSD/SN to reconstruct all user activity. If any activity that is inconsistent with the terms of the user contract is discovered, the user can be excluded from further access to microdata.no.

4. CONCLUSIONS

4.1. Experiences so far

At the outset of the project, the research community was sceptical to metadata driven data access with no access to individual records. During the process and at the final launch the scepticism has turned to optimism and positive curiosity. The main reasons are the no-need for formal applications, instant access to data and the very favourable economic conditions. By end of May 13 institutions have signed up, among them the largest universities and the major research institutes.

4.2. Future challenges

The main purpose of the RAIRD project was the development of the technology. Therefore, functionality and data content are limited in V1.0. The usability of the system depends on the further extensions of both. A main challenge is the inclusion of third party data from outside SN, both register data and sample data. NSD and SN will apply to the Norwegian Research Council for further grants to develop the infrastructure in these directions.

Furthermore, the data protection features built into the technology opens for wider use of the infrastructure beyond research, e.g. public administration, business and even public access.

REFERENCES

[1] Linnerud, Jenny SSB, Ørnulf Risnes NSD, Arofan Gregory MTNA (2014) RAIRD Information Model RIM v1_0 https://statswiki.unece.org/display/gsim/RAIRD+Information+Model+RIM+v1_0

[2] Hamilton, Alistair, Choi, InKyung (2018) Generic Statistical Information Model https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model

[3] DDI - Data Documentation Initiative. http://www.ddialliance.org/

[4] Statistics Norway (2018) Statistical Classifications and Codelists https://www.ssb.no/en/klass/

[5] Agency for Public Management and eGovernment (Difi) (2018) ID-porten http://eid.difi.no/en/id-porten

[6] Thompson, G. et al (2013) Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics, *UNECE Work Session on Statistical Confidentiality, Ottawa, Canada, 20-23 October 2013*

[7] Dwork, C. (2006) Differential Privacy. In *Proceedings of the 33rd International Symposium on Automata, Languages and Programming (ICALP)*, 2:1-12