**Working Paper**
ENGLISH ONLY

**UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION**

**STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)**

**Joint UNECE/Eurostat work session on statistical data confidentiality**
(Helsinki, Finland, 5 to 7 October 2015)

Topic (v): Practicum: Case Studies and Software

# The RAIRD Project: Remote Access Insfrastructure for Register Data.

Johan Heldal[*], Elin Monstad[**], Terje Risberg[*] and Ørnulf Risnes[**]

[*] Statistics Norway (SN)

[**] Norwegian Social Science Data Services (NSD)

**Abstract:** Norway has a large number of registers of individuals established for administrative and statistical purposes, covering the entire population or significant subpopulations. The registers can be merged by personal identification numbers to form event history data bases. The merged registers are used for production of statistics in Statistics Norway and represent a valuable source of data for research. Trusted researchers in approved research institutions may apply for access to the data at their own site. The approval procedure is sometimes long and tedious. There is a desire to simplify the procedure and at the same time make it more safe though remote access and other measures.

In 2012 the Norwegian Research Council funded a four million € project (RAIRD) aiming at creating an analysis server for easier and safe remote access to register data. The project is a joint venture for NSD and SN.

Among the conditions for the project were

1. Remote access (RA)
2. Micro data are invisible, only statistical output should be visible (Anonymous User Interface, AUI)..
3. Users should be allowed to combine data from different sources.
4. The statistical output should be safe.

This paper describes the confidentiality risks that we consider for such a system and how we want to deal with it. The funding runs until 2017 at which time RAIRD should be up going.

There will be demonstration of a prototype.

# 1. Background

For many years, register data and sometimes register data combined with survey data has formed a backbone of empirical research in social sciences in Norway. The registers have been established either for administrative or statistical purposes and can be merged with a unique person identification number which is used in all public and some private administration in Norway. As examples of such registers we will mention

- The Central Population Register (CPR, administrative)
- National Database for Educations (NUDB, statistical)
- Taxation Registers (TR, complete income tax returns, administrative)
- Pensions and retirements
- Employer-employees (for administration of sick-leaves)
- Unemployment benefits
- Social security

The registers are continuously maintained and events are recorded. By merging the registers event histories within the fields covered by the registers can be established for every individual in the population.

As the number of registers has increased and the data have become richer, the demand for such register data has been increasing. Research institutions satisfying certain criteria can apply for approval and be granted access to data for specified research projects at a need to know basis. Trusted researchers in these institutions can get access to data at their own premises. This is a weakness of the present system.

To get access to specific data can be a long and tedious process where permissions are needed from the owners of all registers involved and from the Norwegian Data Protection Authority. Data preparations (by Statistics Norway) can also be resource intensive. The process can take a year and having to wait for the needed permissions and preparations means waste of researchers' resources. Research based on register data has shown a high pay-off. We wish to meet the demand and extend their use to a wider range of researchers than those who have access today.

In 2012 a group of persons from NSD and SN filed an application to the Norwegian Research Council, convincing them that they could build a "cheap to run" and safe analysis server for dissemination of register data to researchers. They were granted 35 mill. NOK (≈ 4 mill. €) to be shared between SN and NSD. Among the conditions for the project were

1. Remote Access (RA).
2. Micro data are invisible, only statistical output should be visible (Anonymous User Interface, AUI).
3. Users should be allowed to combine data from different sources.
4. The statistical output should be safe.

The project aiming at establishing such a system is called RAIRD, Remote Access Infrastructure for Register Data.

RAIRD will differ from most analysis servers discussed in the literature for at least two reasons,

- It intends to handle register data that comprises the entire population

- It contains event history data which can change values at specified points in time

The event histories go back to 1992, for some variables back to 1970.

Australian Bureau of Statistics (ABS) offers Table Builder that enables users to order custom-made census tables on the web from ABS and have them confidentialised on the fly and Data Analyser that offers remote exploratory analysis, data manipulation and regression analysis with invisible data. But these products are limited to cross sectional, i.e. static data. Statistics Denmark offers register data RA to researcher's desks through their Research Services. The Danish services are more extensive than our RAIRD system in terms of data access. It encompasses business data as well as personal data and although it works with de-identified data, it does not have Anonymous User Interface. It goes beyond what present Norwegian legislation would allow.

Like the Danish solution we want researchers to have access to RAIRD from their ordinary desk computers, not through Safe Centers. We want to avoid manual (and perhaps sometimes arbitrary) output checking. Disclosure control should be automated, which means that it should either be built into the available data or take place on the fly through automatic disclosure control. In order to maintain maximum utility of the research data we will prefer the latter if we can.

Section 2 of this paper describes the information structure of RAIRD as it stands at the time of this writing.

Section 3 describes the system for access to RAIRD.

Section 4 describes the disclosure risks that we see as relevant.

Section 5 describes the confidentiality measures we plan to implement for RAIRD.

Section 6 concludes the discussion.

## 2. The RAIRD information model

Since microdata are invisible, the interface to RAIRD will be completely through metadata which must have all the qualities needed to meet researchers' needs. Figure 1 sketches the present information structure model in RAIRD.

Essentially, the users extract variables from the Event History Data Store to build an Analysis Data Set with the variables of interest. For instance, to fetch marital status at first of July 2010, write the command

import STATUS '2010-07-01' *marit_stat_10_07_01*

where *marit_stat_10_07_01* is the variable name to be used in the Analysis Data Set. The interface of the present prototype is Stata-like. In this process the study population is also delimited. The Virtual Statistical Machine processes the queries from the user and a provisional output is produced. The output is then checked in the Disclosure Control System and eventually modified before the final output is either presented or censored.

The prototype running today is based on modules from the Python programming environment like SciPy (www.scipy.org/), NumPy (www.numpy.org/), Pandas (http://pandas.pydata.org/ ) and Statsmodels (http://statsmodels.sourceforge.net/devel/)

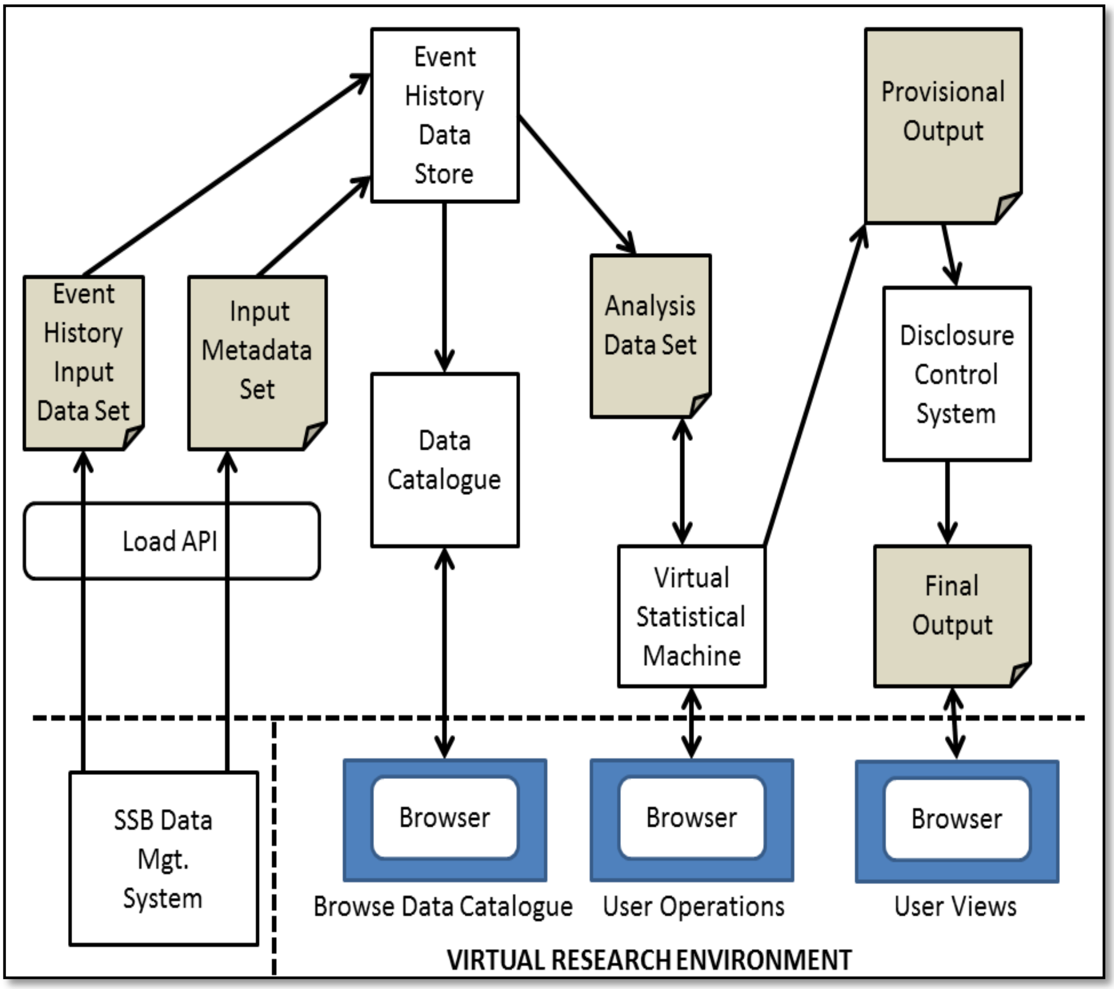

**Figure 1.** The RAIRD Information Model (RIM)

Users cannot invoke these modules directly; instead all operations go through the Virtual Statistical Machine (VSM) as illustrated in Figure 1 above. The VSM is a crucial design aspect of RAIRD as allows for both controlled exposure of methods as well as a managed execution environment where user commands are interpreted and executed separately.

This means that users at any given time will be able to execute only an enabled and safe subset of the functions available in the underlying modules and platforms. The linear algebra routines available in NumPy will e.g. be out of reach for users even if they are installed as part of the NumPy library.

Furthermore, the VSM prevents users from direct access to files (data files, output files, temporary files, cross product matrices, etc). This limits the statistical disclosure interface to analytical output and other output generated by the VSM-component.

## 3. Access to RAIRD

With RA and AUI it will be possible to get around many of the legal restrictions that exist on distribution of register data to researchers' own premises. It will also be possible to extend access to institutions and researchers that are not qualified today. But we may have to introduce different access levels where users at different levels will have access to different detail and different variables. At the lowest level one may find master students at universities and colleges and at the top level the most trusted and experienced researchers who already today have an extensive access. The details in this hierarchy have not been worked out yet.

All researchers that want to use RAIRD, have to be accepted as a RAIRD-user, and identify themselves when accessing RAIRD. The researcher can only see the metadata and the controlled outputs (through AUI) and the micro data as such will never leave Statistics Norway's servers. All communication with the data will go through encrypted communication channels and only by using commands that has been approved by the RAIRD-project in advance.

In RAIRD all user activity will be logged and administrators will be able to replay all work done into the smallest detail. We do believe that such monitoring will make aggressive attacks less likely.

## 4. Disclosure Risks

A considerable literature on attack scenarios and countermeasures has grown during the last years in the wake of the development of analysis-servers (e.g. Sparks et al. (2008), Gomatam et al. (2005), O'Keefe and Chipperfield (2013), Chipperfield and O'Keefe (2014)), notably the Australian Data Analyser (Thompson et al. (2013)).

When microdata are invisible (AUI) the disclosure risks change from *identity disclosure*, the risks of identification of records in a dataset to be a question of *attribute disclosure*, what can be disclosed through statistical output, and so called *inferential disclosure*, the possibility of accurately predicting unknown variable values based on estimated models and known external data.

When a National Statistical Office releases statistics, the output is fixed. The population for each statistic and design of tables and other output is in the hands of the NSO and can be protected before release. However, a person with access to the microdata can manipulate analysis populations and output design in almost indefinitely many ways in order to disclose sensitive information even when microdata are invisible, at least if there are no restrictions or limitations on output.

The fact that RAIRD aims at covering the entire population makes it more challenging. Using a sample, say 20 per cent, would reduce the risks. When samples are taken from registers there is no way a user can have "selection knowledge", not even the selected persons themselves. However, unless explicitly expressed, the discussion here will be based on a 100 per cent assumption.

In the context of RAIRD the disclosure risks considered in the literature have to be seen in relation to Norwegian legislation and confidentiality rules and in relation to which attack scenarios we consider as realistic possibilities. Researchers that have the highest trust today will not need serious restrictions while master students at the lowest level may meet a more restrictions on data and output. Some variables may be released with different detail depending on access level.

A risk assessment may include a description of which personal data will be processed and the security requirements for these. Risk is the product of the consequence of an incident and the (subjective) probability that the incident will occur. The fact that RAIRD is an event history data base also poses extra challenges. Most of the literature on disclosure risk and disclosure control is for cross-sectional data. Some papers have been written on panel surveys, but we don't know literature on panel surveys or event history data bases in analysis servers like RAIRD. We don't feel we have a good overview of possible attack scenarios for such data in a RAIRD setting, but we do mean that the scenarios that are relevant for cross-sectional data also apply to event history data and that protection relevant to cross sectional data is also relevant for event history data.

### 4.1. Tabular output

Without restrictions a RAIRD-user would be able to limit a table population to some few people with known characteristics and known identities and use the table to disclose the values of unknown variables. This can be avoided by having a minimum threshold for the size for any study population selected in RAIRD, but this restriction can be bypassed by *differencing* two tables based on nearly the same population.

Another kind of disclosure is *group disclosure* which occurs if for a combination of one set of variables *A* there is only one combination occurring for another set of variables *B*. Then the (common) values of all the variables in *B* can be disclosed for people with the actual *A*-combination.

*Small counts*, in particular 1 and 2 in frequency tables, are often seen as hazardous. In a RAIRD context we see two kinds of risks associated with small frequencies,

1. If a person who is alone in a table cell can be identified, the RAIRD-user can produce magnitude tables for the same table population and disclose the values of any magnitude variable associated with that person. This calls for counts of 1 or 2 never to be released. What to do with such counts is discussed in section 5.1.
2. If the RAIRD-user has access to external datasets, for instance sample survey data, the user will be able to use the tables to verify that a sample unique in the survey is also population unique on the same set of variables.

Disclosure risks with *magnitude tables* are mostly associated with business statistics. If there are only one or two contributors release will almost immediately lead to disclosure. If there are more than two contributors there is then an implicit assumption that the identities of the largest contributors to the cell totals and their order of magnitudes are known. In particular the two largest contributors are assumed to know which of them is the largest and which is the second largest. Such knowledge is rare when it comes to people, but may also occur in some contexts since individuals with extreme values on particular variables may stand out and be visible in the population under study. The treatment of such cases should be considered in connection with the treatment of extreme values, max and min. The issue is then to ensure a sufficient uncertainty with respect to their actual contribution. The classical methods of cell-suppression used by statistical agencies do not apply in a context where the table design can be manipulated and equations set up to solve for the unknowns.

Raising values to powers can be used to make the most extreme magnitudes more extreme with relatively larger contributions to sums. Actually, if $x_1 > x_2 \geq \cdots \geq x_n \geq 0$, then

$$\lim_{p\to\infty}\left(\sum_{i=1}^{n} x_i^p\right)^{1/p} = x_1 \qquad (1)$$

So just by choosing $p$ large enough, the largest contribution to a cell total can be revealed as closely as one likes. This raises the question on which transformations of the data should be allowed.

### 4.2. Graphical output and extreme values

Several kinds of graphical output represent disclosure risks, most prominently scatter plots where the value of the 'second' variable can be read too accurately if the value of the 'first' is unique and known or the regression fit can be calculated for an given individual in the data set. Also box-plots showing extreme values are risky, in particular for a data set consisting of an entire population where the 'inclusion probability' is one. An extreme value based on a 100 per cent sample, i.e. who has the highest income in a certain subpopulation, are primarily 'sensitive' if the value itself is not publicly known but it can be known to whom it belongs. In a sample, of say 20 per cent, a known extreme value may reveal that a specific person has been selected to the sample, disrupting some of the protection sampling would give compared to taking all.

Exact median value in a box plot from a 100 per cent sample can hardly be said to be disclosive. If a known value for an identified person shows up as a median the only thing learned is that this person is at the median. But this is not an attribute information and hardly very 'sensitive'. On the other hand, if such a value shows up as the median in a 20 per cent sample it may make it very likely that the person to whom it belongs is in the sample. In a sample such knowledge is undesirable.

### 4.3. Analysis output

Most attack scenarios in the literature associated with analysis output deals with manipulation of regression variables and regression output. But the issues associated with cell totals in magnitude tables apply to any sum of magnitude variables whether the sum occurs in a tabular context or not. Most kinds of analysis are based on sufficient statistics that are sums of transformations of variables, such as a cross-product matrix $\mathbf{X}'\mathbf{X}$, where $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)'$ is the data matrix. In standard statistical software these statistics can be output and saved to a file. This will not be allowed in RAIRD, although it could be computationally efficient when running a large number of regression models based on a large data set (entire population). Cross-product matrices contain sums of squares of magnitude variables and will therefore be more disclosive with respect to the largest contributions than ordinary sums of magnitudes. However, the covariance matrix of the OLS regression parameters is $\widehat{Var}(\hat{\beta}) = \hat{\sigma}^2$ $(\mathbf{X}'\mathbf{X})^{-1}$, the scaled inverse of the cross product matrix where $\hat{\sigma}^2$ is the estimated residual variance. The cross product matrix can be reproduced from this covariance matrix. It will therefore be undesirable to make the true estimated covariance matrix visible to the user of the system. But the variances of the regression parameters are absolutely required output. It has been suggested (Gomatam et al.) to suppress the variances, only providing $p$-values for the estimated parameters, but for most researchers this will not be satisfactory.

Regression parameters are most often not considered to be risky in themselves, but as pointed out (… Gomatam et al. (2005) and O'Keefe and Chipperfield (2013)) by running a large number of regressions it is possible to set up equations between the estimated parameters and the elements of the cross-product matrices and calculate them. Sparks (2008) and O'Keefe and Chipperfield (2013)) demonstrate how single observations may be disclosed by giving them a high leverage through transformations and other kinds of manipulation.

### 4.4. Inferential disclosure

The concept of inferential disclosure goes back to Duncan and Lambert (1986) and is about how sensitive values can be inferred indirectly through analyses, for instance very accurate predictions. In the traditional system for disseminating micro data to researchers in Norway it has not been possible to give that issue much attention. It is uncertain to which extent it will be given attention in the future. Predictions that are 'too accurate' are rare in social sciences. To the extent very precise model fits occur in real research, such results can be of high scientific value even though they may stigmatize groups of people. But this is not a closed issue.

## 5. Disclosure risk avoidance

All activity on RAIRD will be monitored by logged. Obviously suspicious activity will cause alarm. We are aware that it is difficult to distinguish legitimate and illegitimate activity on such systems, but we do believe that such monitoring will make aggressive attacks less likely.

No disclosure protection can be perfect. No single method can avoid all kinds of risks. Information loss associated with disclosure risk avoidance must be balanced against the utility the data have for research. We want to do as little as possible with the underlying micro data although users with different access levels may have some variables with different levels of detail. Generally we say that it should not be possible with ‘reasonable means’ to trace a variable value to some specific person. The concept ‘reasonable means’ is not quite precise and may change content with changing availability of externally known information, but an attempt to disclose a value should take more effort and have a higher cost than it is likely to be worth.

The safety of a system like RAIRD can be put under scrutiny in the public debate. Such a debate can be harmful for research carried out with RAIRD even though a register based system like RAIRD will not depend on people’s willingness to ‘respond’. The disclosure protection in RAIRD should be convincing enough to stand up against such scrutiny.

In our search for disclosure protection tools for RAIRD we have been fascinated by the ideas behind the Australian Table Builder and Data Analyser where a random number, a so called Record Key, is associated with each individual and aggregated along with counts and magnitudes to a 'Cell Key' as a foundation for perturbations of statistical output. Such methods will most likely be implemented in RAIRD, but they are not sufficient. We consider some general measures and some measures associated with specific kinds of statistics. General measures considered are for instance

- The study population for every analysis must exceed a given threshold, e.g. 100 individuals, to be carried out.
- For every analysis the number of observations per parameter must exceed a minimum value, e.g. 20 individuals. This also applies to tabular output where each cell total is one parameter.
- Noise addition on all aggregations, counts or magnitudes that will be output. Noise addition will follow the same principles as described in Thompson et al (2013),

  - Constant noise. Repeated runs of the same aggregations based on the same data -should result in the same perturbed totals
  - Approximately constant variation. Noise should not increase with increasing number of observations
- Removal of a random subset of the observations, say 5 per cent, as suggested by Sparks et. al (2008), perhaps so that the size of the final dataset is divisible by 3 or 5 (Lucero et al (2011)). Such a random subsetting must be the same for every repeated run of the same query on the same population, but may differ across queries. This can be achieved using Record Keys and gives some protection against differencing, or simpler; by letting the RAIRD database consist of only 95 per cent of the population. The main purpose of this is to make differencing more difficult.

### 5.1. Tabular protection

Some special rules will be applied to tabular output.

- The proportion of empty cells in a table cannot exceed a threshold, say 20 per cent.
- The proportion of cells with small counts (1 and 2) should (before perturbation) not exceed a threshold, say 10 percent.
- The small counts will either be set to zero or rounded (unbiasedly) to 0 or 3, eventually 5.
- Perturbation of other cells counts will be a function of the cell key as well as the cell count.
- Magnitude sums for the cells will first be adjusted according to the perturbations of the underlying counts so that averages are preserved. If this does not give enough protection of the largest contribution, some additional noise may be given as well.
- Permissible transformations of variables may be restricted, in particular raising magnitude variables to high powers as in formula (1).

### 5.2. Graphical output and extreme values

Scatterplots must be presented in such a way that the value of the ‘second’ variable cannot be read too accurately from the ‘first’. The criterion for what is ‘too accurate’ should be the same as when protecting magnitude tables. Thompson et al (2013) suggests using hex-bin plots as a protective measure. Hex-bin plots are readily available in R and Phyton. Sparks et al (2008) suggests replacing scatterplots with modified parallel boxplots. Such modified boxplots can replace ordinary boxplots in general, but care should be taken that boxplot representations of distributions of continuous variables are consistent with tabular and other summaries of such variables.

### 5.3.Analysis output

As mentioned in section 4.3, estimates of regression parameters and most other model parameters are rarely very disclosive in themselves, but the cross-product matrix can be. Particularly the squared sums on the diagonals of these matrices may be disclosive for the largest contributor. One way of attacking this problem is to perturb the aggregations in the cross product matrix in the same way as tabular sums. Another approach is to add noise directly to the estimates of the regression parameters to make back-calculation of $\mathbf{X}'\mathbf{X}$ uncertain. Chipperfield and O’Keefe (2014) suggest doing this adding noise to the estimating equations, e.g.

$$Sc(\boldsymbol{\beta}; \mathbf{X}) = \mathbf{X'}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{E}^* \qquad (2)$$

Here $Sc()$ is the score function, $\boldsymbol{\mu} = E(\boldsymbol{y})$, and $\mathbf{E}^* = (E_1^*, \cdots, E_K^*)$ is a vector of independent perturbations that replaces $\mathbf{0}$. $E_k^*$ may be of the form $E_k^* = e_k u_k^*$ where $e_k = \max_i\{x_{ik}(y_i - \mu_i)\}$ , the largest contribution a record in the micro data makes to the $k$-th coefficient of the estimating equations. The $u_k^*{}'s$ are independently generated from the uniform distribution on (-δ, δ) where the parameter δ >0 controls the amount of noise. An advantage of this method is that it is general and applies to large classes of estimation problems. The challenge is to set δ to give the right level of noise.

Several authors have pointed out the possibillty of using transformations to increase leverage of specific records or records with specific values on the explanatory variables in order to create regressions that disclose the response values of associated with these records. Gomatam et al (2005) and O'Keefe and Chipperfield (2013) suggest countermeasures. We are considering the feasibility of implementing these measures in RAIRD.

## 6. Conclusion

This paper has aimed at describing RAIRD and the confidentiality challenges that we see as most important for such a system and our thinking about how these challenges can be met. At the time this paper is written many details in the disclosure control design have not been clarified, but they are crystallizing. We are open for comments on which ideas are good or not so good, strengths and weaknesses. By the time this paper will be presented in Helsinki more details will be made clear. So far only the most common analysis tools, such as different forms of regression analysis, have been studied thoroughly in a remote server disclosure control context. We do believe these studies apply to regressions for event history data as well, such as for instance Cox-regression. RAIRD will have to start with these methods. Other statistical methods will be included as we learn which disclosure risks they represent and how to overcome them.

It is important that the system will be accepted by the researchers for whom it is meant. Some aspects of the system, and in particular the Anonymous User Interface, have been met with skepticism by researchers who have been used to be able to view the microdata they are working on, but we have had a good dialogue with some of them. Some researchers have been allowed to test our prototype. This has increased their understanding of our ideas and given us much more positive feedback. We believe that in the future, RAIRD will set new standards for micro data dissemination in Norway.

## References

Chipperfield, J.O. (2014). *Disclosure-Protected Inference with Linked Mocrodata Using a Remota Analysis Server.* Journal of Official Statistics, Vol. 30, No. 1, pp123-146.

Chipperfield, J.O. and O'Keefe, C.M. (2014). *Disclosure-protected Inference Using Generalised Linear Models.* International Statistical Review, 82,3 pp 371-391.

Duncan, G.T. and Lambert , D. (1986). *Disclosure-limited data dissemination (wit discussion).* J. Amer. Statist. Assoc. 81, pp 10-28.

Gomatam, S., Karr, A.F., Reiter, J.P. and Sanil, A.P. (2005). *Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Analysis Servers.* Statistical Science, Vol. 20 No. 2 pp 163-177.

Krenzke, T., Gentleman, J.F., Li, J. and Moriarity, C. (2013). *Addressing Disclosure Concerns and Analytic Demands in a Real Time Online Analytic System.* Journal of Official Statistics, Vol. 29, No. 1, pp 99-124.

O'Keefe, C.M. and Chipperfield, J.O. (2013). *A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems*. International Statistical Review, 81, 3 pp 426-455.

Sparks, R., Carter, C., Donnelly, J.B., O'Keefe, C.M., Duncan, J., Keighley, T. and McAullay, D. (2008). *Remote access method for exploratory data analysis and statistical modelling:Privacy-Preseving Analytics.* Computer Methods and Programs in Biomedicine, 91, pp208-222.

Thompson, G., Broadfoot, S. and Elazar, D. (2013). *Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics.* Joint UNECE/Eurostat work session on statistical data confidentiality.