Information Matching Bulletin
News from the Office of the Privacy Commissioner – December 2013

In this edition

First Approved Information Sharing Agreement

The first Approved Information Sharing Agreement came into force on 3 October 2013 (Privacy Act schedule 2A amended by clause 14 of the Privacy (Information Sharing Agreement between Inland Revenue and Internal Affairs) Order 2013 (SR 2013/374). The agreement is between the Department of Internal Affairs (DIA) and Inland Revenue.

The agreement allows DIA to share information from adult passport applications with Inland Revenue to help it contact overseas-based student loan borrowers who are in default and child support liable parents who are in arrears or have out-dated contact details. The agreement can be accessed from the Inland Revenue website at http://www.ird.govt.nz/aboutir/agreements/.

The Commissioner may prepare and publish a report on an approved information sharing agreement under section 96P of the Privacy Act. Our report on this first agreement can be accessed from our website at http://privacy.org.nz/information-sharing/information-sharing-introduction/approved-information-sharing-agreements/.

New IR match enables dual application process

From July 2013, Inland Revenue has offered people registering a birth with DIA the option to also apply for an IR number for their newborn. The new programme, authorised by section 78A of the Births Deaths and Marriages Registration Act 1995 involves DIA disclosing information to Inland Revenue to enable it to allocate an IR number to a newborn.

By re-using information collected by DIA (including verified identity information), Inland Revenue saves customers time and money by removing the need to separately apply for an IRD number. In turn, Inland Revenue reduces its cost of processing new applications and the potential for identity fraud. The initial results show a strong uptake of the service by the public. The first formal reporting on the programme will be in our next annual report.

Review of MSD Integrity Intervention Centre programmes

A report on the review of nine programmes operated by MSD at its Integrity Intervention Centre was tabled in Parliament on 8 August 2013. The report covered the following programmes:

- ACC/MSD Benefit Eligibility
- BDM/MSD Deaths
- BDM/MSD Identity Verification
- BDM/MSD Marriages

OPC/1795 /A335286

- Corrections/MSD Prisoners
- Customs/MSD Arrivals and Departures
- HNZ/MSD Benefit Eligibility
- IR/MSD Commencement Cessation Benefits
- IR/MSD Commencement Cessation Students

In the report we recommended that MSD either cease operating the HNZ/MSD Benefit Eligibility programme and divert resources into other programmes, or modify the operation of the programme to improve its return on investment. For all the other programmes we were satisfied that they should continue. The report is available at http://privacy.org.nz/information-sharing/information-matching-reports-and-reviews/

Review of MSD Senior Services programmes

A report on the review of seven programmes operated by MSD Senior Services was tabled in Parliament on 8 August 2013. The report covered the following programmes:

- Centrelink/MSD Change in Circumstances
- Centrelink/MSD Periods of Residence
- Customs/MSD (Centrelink) Periods of Residence
- Netherlands (SVB)/MSD Change in Circumstances
- Netherlands (SVB)/MSD General Adjustment
- Netherlands (SVB)/MSD Debt Recovery
- IRD/MSD (SVB) Tax Information

In the report we recommended that the Centrelink/MSD Periods of Residence programme be discontinued because of the continuing low level of positive results. For all the other programmes we were satisfied that they should continue. The report is available at http://privacy.org.nz/information-sharing/information-matching-reports-and-reviews/.

2012/13 annual report

Details about each active programme are contained in the Annual report, to be available on our website within the next couple of weeks.

Four new programmes went live during the year:

- BDM/MSD Overseas Born Name Change Match
- BDM/IR Child Support Processing Match
- DIA Identity Verification Service Programme (IVS)
- Customs/IR Student Loans Alerts.

Parliament passed three information matching authorisations during the year that have translated into active programmes:

- The Electronic Identity Verification Act 2012 authorising the DIA Identity Verification Service programme

- The Student Loan Scheme Amendment Act 2013 authorising the Customs/IR Student Loans Alerts programme

- The Social Security (Benefit Categories and Work Focus) Amendment Act 2013 authorising the Justice/MSD Warrants to Arrest Programme.

Changes to how we assess programme compliance

During 2012/13, we made changes to how we assess compliance in order to:

- streamline reporting requirements on agencies, and
- obtain evidence on compliance with each of the information matching rules.

In order to streamline reporting requirements, we reviewed the requirement for an annual audit to see whether it was still the most efficient way for us to get the information we need to manage privacy risks.

We considered that requiring agencies to conduct an annual audit where there were no substantive issues identified in the previous year's audit report, and no substantive changes to the system used represented an unnecessary demand on agency resources. As a result we have changed our requirements so that an audit is required only at intervals of up to three years in these circumstances.

In lieu of an annual audit, we instead required a letter from the agency giving formal assurances about any changes made to the operation of the programme, measures taken to ensure people are aware of the match, measures taken to destroy information in accordance with the technical standards report, and the contents of adverse action letters sent as a result of the match.

This change only affects the requirement for an audit and does not affect any requirements for statistics or other information that we normally receive about any of these programmes.

Through the review, we also identified that where programmes were not subject to regular audits, we did not have sufficient information to be fully confident about agencies' compliance with destruction requirements set out in Part 10 and Schedule 4 of the Privacy Act.

This year, we have asked agencies to provided details about how they manage match information in order to comply with these requirements. This additional information has revealed that a number of matching programmes do not have sufficiently robust processes for destroying data in accordance with the Act. We will be working with agencies to resolve these issues during the coming year.

This year we have also changed the way we describe programmes' compliance. There are three levels:

Compliant: where the evidence we have been provided indicates that the programme complies with the information matching rules

Not compliant – minor technical issues: where reporting has identified practices that are not compliant with the information matching rules, but genuine efforts have been made to implement a compliant programme, and the risks to individual privacy are low.

Not compliant – substantive issues: where reporting has identified practices that are not compliant with the information matching rules or other provisions of the Privacy Act that cannot be considered minor technical issues.

The effect of data cleaning on record linkage quality
Colin Trotter

Randall et al.: The effect of data cleaning on record linkage quality. *BMC Medical Infomatics and Decision Making* 2013 13:64 found at http://www.biomedcentral.com/1472-6947/13/64 is an informative article with interesting insights about data cleaning and record linkage (the latter process sounding very similar to data matching!). I've attempted to summarise the key ideas in the article and present them below with some references to information matching in New Zealand.

Testing the theory

The authors noted that to their knowledge there had been no systematic assessment of the extent that data cleaning improves linkage quality, or which techniques are most effective. To answer that question, two datasets were used for the study, a 'synthetic' made up dataset and a 'real world' administrative dataset. Each dataset was subject to both a light cleaning process and a heavy cleaning process (details provided in a table within the article). The cleaning process, linkage method, and how the linkage quality was measured can't be easily summarised so I will skip to the results and conclusions.

Research results and conclusions

The expectation is that data cleaning would have significant benefits but their results contradict the accepted wisdom that data cleaning is a worthwhile procedure.

Overall, if there was any effect at all, data cleaning appeared to decrease linkage quality. For both datasets, a high level of cleaning resulted in a decrease in linkage quality while light cleaning slightly decreased the linkage quality for the synthetic dataset and no change in the administrative dataset.

The key reason why cleaning failed to improve quality was the reduced variability of each field. While the number of correct match results increased with data cleaning, the number of incorrect matches also increased, in most cases dramatically, leading to a poorer overall outcome. While removing missing values and uninformative values seemed to increase predictive ability, all the other cleaning mechanisms produced mixed or worse results. Using name variables that had nick-names and diminutive names replaced with their original names resulted in a large decrease in that variable's predictive value.

The article concludes that there are many different data cleaning techniques which may be suitable in specific settings, but care needs to be taken.

So, how do you go about linking (or matching) records?

Record linkage methodology

There are two record linkage methods commonly used, the deterministic method and the probabilistic method. The process used in practice may well be a mixture of these two approaches. In simple terms, the deterministic method relies on matching consistent identifiers such as a client number. It the client number (and possibly some additional information) matches then the records are considered to be about the same person.

The probabilistic method involves comparing various fields of information to calculate the likelihood or odds that different records relate to the same person. The match result is scored, and when the score meets a certain threshold the records are deemed to be about the same person. Probabilistic matching allows for inconsistencies and errors in the linking fields, while still coming up with a match result. Probabilistic matching has been found to be more effective at dealing with errors and results in better linkage quality than deterministic methods.

An example of the probabilistic matching method being used in information matching is the Customs/Justice Fines Defaulters Alerts programme that identifies serious fines defaulters as they cross the border. The Customs matching algorithm provides results that are expected to be correct in 9 out of 10 cases and is supported by manual steps to verify identity before any action is taken.

Data cleaning techniques

Data cleaning involves correcting, removing or changing fields based on their values. These new values are assumed to increase data quality and result in enhanced match results. In the absence of strongly identifying information, data cleaning is recognised as one of the key steps in the linking (matching) process. When trying to link records, data cleaning aims to reduce the number of false positives and false negative results. Without data cleaning, many true matches may not come to light.

There are a variety of data cleaning techniques used in record linkage and data matching:

Reformatting values: Sometimes data is represented in differing formats that make comparison impossible, differing date formats is a common one. For example, 25 December 2013 and 25/12/13 would need to be changed to a common format to enable successful linkage or matching.

Removing punctuation: Removing apostrophes, hyphens and spaces from names can reduce the likelihood of miss-matching two records that are about the same person

Name and address standardisation: Breaking a name or address into its individual components means each component can be individually compared. By creating multiple variables, small differences such as a differing order may have less impact on the ability to successfully match two records.

Nickname lookups: Using a file of common nicknames and diminutive names can be used to translate forenames to a common value, for example, Alex and Alexander, or Bill and William.

Publications

To find out more about information matching, check out some of our other resources and publications at http://privacy.org.nz/data-matching-introduction/

Contacts

Wellington
4th Floor, 109-111 Featherston Street
PO Box 10-094
Wellington, New Zealand
Telephone: 64-4-474 7590

Simon Rae
Team Leader, Policy and Technology
Direct Line: 64-4-494 7082

Neil Sanson
Data Matching Compliance Adviser
Direct Line: 64-4-474 7592

Colin Trotter
Senior Adviser, Data Matching Compliance
Direct Line: 64-4-494 7087

You can contact us by email. Our standard email format is first name.surname@privacy.org.nz

OPC/1795/A335286