WingArc 1ST

WHITE PAPER

THE STATE OF OPEN DATA

wingarc.com.au

At WingArc Australia we have been helping government agencies and other organisations to explore, analyse and share their data for over three decades.

WingArc solutions are used by leading government departments and statistical agencies worldwide, and we strongly believe that opening up government data to citizens as widely as possible can bring huge benefits to society.

In this white paper, we look at the history of the open data movement, exploring what has worked well and where there are gaps between the promise and the reality.

We also look at how the right tools can allow data to be opened up in ways that are most useful to the community, and allow consumers of the data to extract maximum value.

Finally, we look at how having the right confidentiality capability can enable the release of datasets that might otherwise have been deemed too sensitive or difficult to release.

We're very pleased to present this white paper on the state of government open data. We hope you find it useful.

Steven Hulse is CEO of WingArc Australia

WingArc 1ST

WingArc Australia

360 Collins Street
Melbourne
Victoria, 3000
Australia

+61 3 9615 5200

Over the past twenty years, governments around the world have come under increasing pressure from researchers, statisticians, journalists and ordinary citizens to make more of their data available. While the term "open data" was first coined in the mid 1990s, the movement really gathered pace towards the end of the first decade of the new century.

But after an initial flurry of activity, has open data lived up to its promised potential?

There is widespread agreement that opening up government data is good for everyone. Some of the major promised benefits include:

- Easier access to government services and information.
- Fostering innovation and enabling new solutions and services to be built.
- Easier interaction and sharing of knowledge between government departments and agencies.
- Improved transparency and accountability.

In theory the opening up of data should lead to economic growth, as new businesses and industries spring up to take advantage of and build on top of these new data streams. Navigation apps that are powered by live transit data feeds from a city's public transportation authority are just one example of the innovation that industry can bring to government data.

Open data can also drive improvements in public services, by allowing easier identification of gaps and areas for improvement.

And yet despite the progress that has been made, there are undeniably issues with the current state of open data.

In many countries open datasets are primarily delivered through vast data portal sites. Typically these sites organise datasets either chronologically, by file type, or under broad thematic categories. Often it can be difficult to search or sort through large numbers of datasets of somewhat limited value to find the gems. While the objective of releasing as much data as possible is consistent with the overall goals of the open data movement, quantity does not necessarily equate to quality. There is a need for data curation, as well as engagement activities with the community to help more people to take advantage of the data, and to help the more useful or interesting datasets to find their audience.

In addition, simply uploading large quantities of datasets to these portals does not in and of itself make that data truly open. Complex datasets released in their raw form are only really open to a small group of technical specialists with the skills and tools to use that data. That data is not really open unless governments also provide tools that make it easy for

the more general audience to use and understand what is being released—tools that give end users the power to self-serve.

Open data also suffers from fragmentation and a lack of standardisation across states and cities. There might be really strong datasets for a particular topic covering one city or state, but no equivalent data available for other jurisdictions across the country to allow for comparisons.

Or a time-series dataset might be released only for a limited period, perhaps driven by a diligent individual data custodian within a particular department. There is no guarantee that the dataset will continue to be updated over the long term, and without that reassurance we miss out on the promised economic growth from commercial enterprises building solutions on top of those datasets: what business can justify the investment in research and development without guarantees around continued availability of key datasets?

And some datasets just don't get released at all.

Typically these are the more sensitive—and potentially more interesting and useful—datasets. These are the ones that don't easily fit the model of "raw data files on a data portal", due to the confidentiality implications. For these datasets, different tools are required to enable their release.

The COVID-19 pandemic presents another example of gaps in government data releases. During an unprecedented global health emergency there was a widespread desire amongst the community for accurate and up-to-date data on aspects such as case numbers, hospitalisations, and, later, vaccination uptake. And yet here in Australia there was a striking lack of data transparency during the pandemic, and a huge disparity between what was made available by different states.

Ultimately the community stepped in to fill the void, with individual volunteers devoting their spare time to scrape whatever data they could and develop websites such as covid19nearme.com.au and covidlive.com.au. This included scraping values out of PDF dashboards to infer case numbers in the absence of a consistent official API or other data stream in a machine-readable format being made publicly available across all states and territories.

The Australian experience with COVID data is in stark contrast to other countries around the world. The UK, for example, embraced an open approach to its COVID statistics with a world leading dashboard with APIs for anyone to access[1]. The technical lead behind this project was recognised with an MBE in the 2023 New Year's Honours, but even there, rather than heralding the start of new advances in open data, the team has subsequently been disbanded, in a move commentators have described as a missed opportunity: "The [UK] COVID dashboard should've been the start of something, not the end".[2]

OPEN DATA FOR ALL

At WingArc Australia, we believe that for government data to be considered truly open, it must be opened in a way that meets the needs of all the different types of data consumers. For example, that includes:

- Internal government users, both trusted and untrusted.
- Trusted and untrusted external parties such as other agencies or researchers.
- Industry.
- Data journalists and data scientists.
- The general public.

Our tools are designed to meet the needs of all different user types by providing a powerful, programming-free interface to explore and analyse data. Not just raw data, but a true self-service platform that anyone can access. We also support interactive visualisations using charts and maps that users can easily share and embed in reports.

Traditionally, governments only made data available in aggregated form, or though pre-defined reports. While there is value in summary data, that format also limits its usefulness. When building the report or the aggregated table, the data publisher must anticipate the questions that users will want to ask and decide which specific combinations to release.

But what if users want to ask different questions?

As data providers around the world know only too well, the same piece of data can mean something different to each stakeholder, and the questions they will want to ask are different. It is impossible for data providers to offer a pre-defined report that can anticipate all possible uses of that data.

In fact, we believe that the real potential of open data is in enabling innovation by allowing the users of that data to build new solutions and use the data in ways that we haven't even thought of yet.

Our web portal tools are designed around this self-service model, which provides all the advantages of unit record data without actually releasing the underlying unit records at all. End users can build their own *ad hoc* queries to ask any question they like and see the aggregated results, rather than being limited to predefined summary information and views that have been created for them.

On top of that, our Open Data API enables direct integration of the data into other tools, such as R or dashboarding platforms, apps, visualisations, interactive data stories and more.

THE IMPORTANCE OF DISCLOSURE CONTROL

One of the reasons why governments have historically released summary data, rather than unit records, and the reason why some datasets simply don't get released at all, has been due to concerns around protecting the privacy of the individuals and organisations in that data.

A breach of confidentiality occurs when a person or entity is recognised in a dataset, allowing an attacker to find out new information about that person.

While there is a risk of identification from aggregated data, the risk is clearly much higher when the underlying unit records are released.

Even with aggregations, merely removing personally identifiable information (PII) is not enough to fully anonymise a dataset. While removing independently sensitive fields such as names, credit card numbers or IP addresses is an absolute minimum level of protection, this alone will not always hide the identity of the individuals in the data.

This is especially true with web access interfaces that allow multiple queries to be submitted and easy ways to digitally combine datasets. There is still a risk that specific characteristics might allow an individual to be located even without the inclusion of explicit identifiers. True privacy protection requires an integrated and systematic approach.

The critical task of protecting privacy is far too important to leave to manual control. Human error can be catastrophic both for the department releasing the data and the individuals within that data. Proven confidentiality routines are therefore essential, and the application of this protection must be automatic.

Traditionally, open data solutions have been forced to compromise. To choose between utility of data and risk of disclosure. Often this means that some data simply doesn't get released.

Our product suite is designed to offer highly useful data with a low risk of disclosure. The applications sit on top of the unit record data, providing on-the-fly aggregation from the data in those unit records, with confidentiality protection that is robust and automated. This offers a "best of both worlds" solution that allows both confidentiality protection and data utility.

In the following section we look in more detail at various disclosure control solutions and their advantages and disadvantages.

> **The government's open data agenda allows us to find out more than ever about the performance of public bodies.**
>
> **However, there is also a risk that we will be able to piece together a picture of individuals' private lives.**
>
> **With ever increasing amounts of personal information in the public domain, it is important that organisations have a structured and methodical approach to assessing the risks.**
>
> Christopher Graham
> UK Information Commissioner[3]

Data providers must maintain the privacy of the individuals and organisations that contribute to their data, and many countries have laws and regulations that require and reinforce this.

It is also important to note that malicious identity theft is not the only consequence of confidentiality breaches. If survey respondents do not believe they are adequately protected from a possible disclosure risk, they are less likely to comply with requests for information, or even worse they may simply respond with inaccurate or fake details. For organisations who need the underlying data for public policy and planning, this can have far reaching consequences.

On the other hand, if the only data that can be released has to be confidentialised to such an extent that it becomes less useful and less accessible, then there is a risk of negative reactions from the end users of that data. This can also increase staff workload, as they struggle to serve a growing number of more complex data queries from consumers and stakeholders. For the modern data collection and dissemination agency, balancing this risk–utility equation is one of the toughest tasks in designing disclosure control.

HOW IS CONFIDENTIALITY BREACHED?

A breach of confidentiality can happen when multiple fields or variables are combined to uniquely identify an individual or enterprise. For example, while "occupation", "postcode" and "number of dependants" might not on their own be considered identifying information, what if there is only one person in the dataset that matches a specific combination of all three?

Typical confidentiality breaches involve combining information in the released data with some known or easily discoverable information.

For example:

- An attacker searches for a specific individual in the dataset, based on information the attacker already knows about that individual. Only one record matches all known criteria.
- An attacker starts with a record in the anonymised dataset and then tries to identify that individual by matching them with publicly available information.

In some cases it can take surprisingly few pieces of additional information to re-identify an individual. A 2014 study of anonymised credit card data belonging to 1.1 million people found that just four pieces of external information were enough to match a person with their anonymised credit card record, 90% of the time.[4]

Some typical types of attacks are described below.

DIFFERENCING ATTACKS

A differencing attack involves generating two related tables and comparing the results.

For example, one for all employees and one for employees earning less than $150,000. By subtracting the results of the two tables an attacker can produce a third, "differenced" table containing information about a subset of interest (in this case, employees earning *over* $150,000).

Differencing attacks are commonly used to breach suppression algorithms, which are rules that hide cells with a count below a certain value or with fewer than a certain number of contributors. In a differencing attack, the tables created have results that are large enough not to trigger the suppression rules, allowing the attacker to infer results that would otherwise be suppressed.

HOMOGENEITY BREACHES

Department	Sales
Gender	Male
Age Band	50-54
Diabetes	0
Something Embarrassing	5
High Blood Pressure	0

Sometimes confidentiality breaches occur without even needing to isolate a single record in the data. Consider a simple example table showing medical conditions for males employed by a particular organisation.

In this case, even though there is a count value of 5 for the embarrassing condition, all males in the selected subgroup have that condition. This group's confidentiality has been breached through homogeneity of their attributes.

INFERRED BOUNDING

Inferred bounding is the name given to a targeted attack that attempts to calculate the value of a suppressed cell to within set "bounds". This typically involves sophisticated linear programming techniques to deduce the maximum and minimum possible values of a protected cell. If the difference between the upper and lower bound is less than one count, then the suppressed value is discovered.

Although bounding can commonly be used on count datasets to deduce the count with a particular combination of attributes, inference can also successfully be used to deduce magnitude data.

Consider two competing firms that dominate industry in particular region. A simple query regarding the monetary value of government grants to each region immediately tells each firm roughly what its competitor is receiving.

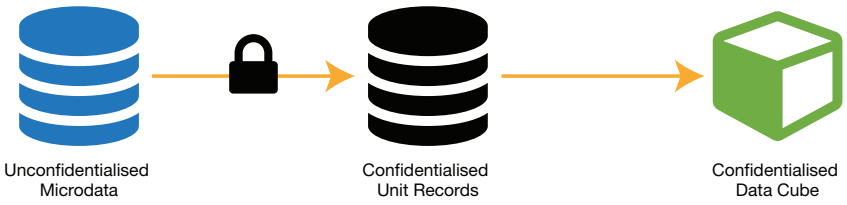
Protecting against this risk is not as simple as enforcing a minimum number of contributors (say, three) to each cell. In the above scenario, there could easily be three firms, but one is much smaller than the others, in which case a good bounding estimate can still be generated from the report. This type of attack is prevalent in commercially sensitive information, particularly financial or otherwise tactically sensitive magnitude data.

DISCLOSURE CONTROL SOLUTIONS

When considering an effective disclosure control solution, many factors need to be taken into account, including the type of data, how the data is reported, the end users and the overall sensitivity of the dataset. Another consideration is where to apply disclosure control: at the microdata level or post query?

MICRODATA CONFIDENTIALISATION

Microdata confidentiality involves pre-processing individual unit records, and is most suited to organisations that need to release microdata cubes (such as those with a high proportion of research users).



Typical methods include attribute swapping, generalisation techniques or tuple suppression, and involve finding suitable matches within the record set that need to be adjusted.

However, the number of records and classifiers in modern sets makes this a challenging task, one that can rapidly outclass realistic computational abilities.

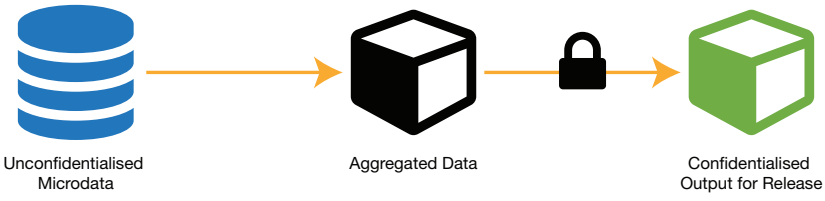
Great care also needs to be taken to ensure that the overall statistical properties of the data are not compromised. Any bias or modified variance in the data must be kept to a minimum.

PRE-AGGREGATION

This technique involves pre-aggregating the unit records in some way, to protect access to the microdata.

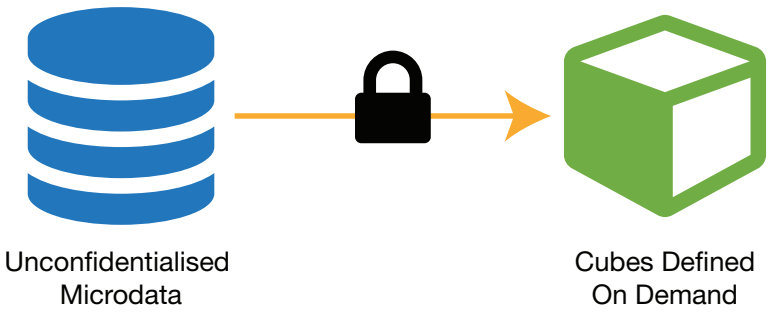
Business rules are applied to generate a “safe” level of aggregation below which the query engines may not drill.

Further disclosure control is applied to predefined output datasets, which are then released.



QUERY-BASED TABULAR LEVEL CONFIDENTIALISATION

In the examples above, end users can submit queries against predefined, approved output. A variant on this approach is query-based access, which relies upon the routines being sufficiently robust that any tabulation query can be submitted by an end user. This approach reduces and sometimes replaces the need to predefine output, delivering greater flexibility to the user.



Query-based access requires a tabular confidentiality routine that can work for a limitless number of *ad hoc* queries. This scenario removes the need for the data provider to predict what combinations of variables need to be combined in summary outputs, delivering much greater flexibility to the end user.

SOLUTIONS FOR DATA DISSEMINATION

Typically, disclosure control techniques fall into three categories:

1. SUPPRESSION

The most visible method is simply to hide sensitive values, replacing them with a symbol or a 0 value. However, a high level of protection can only be achieved when both sensitive cells and related cells are suppressed.

For example, if one value in a row is sensitive and is suppressed, then other values in the row, or the total for that row, may also need to be suppressed otherwise attackers may be able to calculate the suppressed value.

This can lead to very low utility, as many values in a table may be rendered unavailable. Suppression methods are also subject to breach through bounding or differencing attacks.

2. GENERALISATION

This method suppresses some values to a degree, by only reporting more general values (for instance, reporting a more general 4 digit region code, rather than a 5 digit postal code).

It can be applied at both the microdata or at the tabular level, depending on the individual requirements, and is typically used in conjunction with other confidentiality mechanisms.

3. OBFUSCATION

Obfuscation techniques hide information by adjusting the true value of any given cell in the table and reporting a slightly different value instead. With these techniques it is important to ensure that the resulting table preserves the same statistical characteristics as the original table, and that no bias is introduced by changing values.

Some obfuscation techniques, such as randomly rounding cell values, offer high utility and are fast and easy to implement, but may reduce the usefulness of the data, if the rounding is too aggressive. This approach also may introduce a high level of bias, as the rounding is not controlled.

It is also possible that different users may see different results for the same query.

OUR SOLUTION: PERTURBATION

WingArc Australia's perturbation algorithm is a form of obfuscation. It makes adjustments to cell values to ensure that individuals cannot be identified. However, these adjustments are both controlled and repeatable.

This offers a good balance between utility and protection, and ensures that no bias is introduced. It also ensures that the same cell is always adjusted in exactly the same way, no matter how the table query is constructed.

It supports both count and magnitude data and has become a natural choice for many high risk datasets, including population census data.

CHOOSING THE RIGHT SOLUTION FOR YOUR DATA

Disclosure control methods are developing rapidly to accommodate the increasing demand for online data. The benefit of each approach must be weighed against the potential cost to data accessibility and utility.

To do this, we recommend that a data confidentiality assessment be carried out. This process involves developing a risk profile before data is released that takes into account factors such as:

- Data collection factors, such as whether the data comes from a census or survey, and any risks from survey frame or post-collection weighting.
- The type of data (count or magnitude).
- The size of the dataset.
- The range of end users and the type of access required.
- The likelihood and consequences of breaches, considering the end users, the sensitivity of the data, the existence of previous or similar datasets and the required level of detail in the data.
- Implementation and usability requirements.

There are a range of tabular and microdata confidentiality methods available for implementation. Some are already in regular use internationally, whilst new methods are being proposed and tested constantly.

A confidentiality assessment process can help organisations to:

- Build a customised risk profile for your data.
- Suggest the most appropriate disclosure control solution.
- Understand and communicate the benefits and risks of the preferred solution.

LEARN MORE

If you would like to learn more about our solutions for government data dissemination, then please visit us our website at wingarc.com.au/gov where you can read more about our solutions and request a demo.

SELF-SERVICE TABLE BUILDER

FLEXIBLE

CONFIDENTIAL

DISSEMINATION

Try It Now

wingarc.com.au/gov

REFERENCES

1. https://coronavirus.data.gov.uk/details/developers-guide/generic-api

2. https://www.spectator.co.uk/article/pouria-hadjibagheri-and-the-uks-abandoned-open-data-revolution/

3. https://ico.org.uk/media/1061/anonymisation-code.pdf

4. *"Unique in the shopping mall: On the reidentifiability of credit card metadata", Science Magazine, 30 Jan 2015 (Vol. 347, Issue 6221, pp. 536-539).* https://science.sciencemag.org/content/347/6221/536