Chapter Three: The Good Data Manifesto

Claire Trenham and Adam Steer

# Abstract

The Good Data Manifesto sketches out a set of principles which can be used to guide data collection, storage, and re-use. In an increasingly data-rich world, we have long failed to fully consider many implications of how data might be used, once collected and stored. On the technical front - how do we manage, move, and pay for the storage of data? How long can we assume that infrastructure-scale computing is a sustainable solution? And on the human side, how can we adequately protect our rights to privacy and individuality, or even to be forgotten? While some research fields consider ethics deeply in their data management, others have not traditionally done so, thus our aim is to develop principles that might broadly apply across disciplines, to address the question ‘*what makes data good?*’. We draw on examples from a number of fields, but with a general focus on geospatial data which is often large in volume, ubiquitous, and personal. We aim to help data collectors and managers consider more fully how they go about their task, and to help data users critically consider the applicability of datasets to their need.

# Introduction

Data exist everywhere. In a growing technological society, humans are increasingly recording, cataloguing and exploiting observations of the world and ourselves. We are also getting better at producing data - inferences, models and predictions of the world and our behaviour - driven by a growing ability to collect and collate observations, together with increasing computational power to integrate this data with complex theoretical models. Invariably, this leads to new problems: what do we do with all these data? How do we catalogue them? How should we use them?  
  
Less often we consider the questions: *should* we collect, aggregate, catalogue and exploit these data? If so, how? What would be ethical means for doing so? These questions have been deeply considered in health and human-related sciences (e.g. psychology, sociology). Driven by Europe’s implementation of General Data Protection Regulation (GPDR),[[1]](#footnote-1) these questions are also under active consideration in high-resolution earth observation,[[2]](#footnote-2) and wherever ubiquitous, commercialised personal data collection takes place (e.g. ride sharing apps).[[3]](#footnote-3) Historically, these questions are rarely asked before amassing collections of data containing explicit and/or implicit (e.g. location tracks) personal information, until something goes wrong: for example, the 2018 revelation of a major social media company’s collection and third party exploitation of user data.[[4]](#footnote-4)

The preparation of this book is timely, with a number of groups around the world considering what “good data” and data ethics mean.[[5]](#footnote-5) [[6]](#footnote-6) The authors of this chapter are not ethicists, nor data privacy experts. We are Australian data practitioners with experience managing and working with petabyte-scale data collections; advisors on and contributors to continent-scale data infrastructures. We love high quality data but want to make sure the data we produce and consume considers more than fidelity, precision, accuracy, and reproducibility. This chapter is written with a focus on Australian concerns. We focus often on geospatial data or data with a geospatial component, however our goal in this chapter is to touch on considerations across a broad range of data fields.

Boundaries between geospatial data and personal data are increasingly blurry. Data warehouses and practitioners must consider multiple data types and myriad layers of licensing, ethical and access concerns. In this chapter we present a generalised manifesto for ‘good data’, with the aim of creating a set of principles which can guide operations whilst avoiding harm.[[7]](#footnote-7) We ask, ‘*what attributes make data good?*’

# The Good Data Manifesto

There are international efforts devoted to various kinds of data collection ‘goodness’, for example the Research Data Alliance, the Open Geospatial Consortium, the ISO, and the IEEE. These organisations are typically concerned with (meta)data standards and focus on technical aspects with little attention to ethical aspects. We ask, ‘*should* the data be collected?’ If so, how and why? How long should it persist? What makes it useful/good? Aspects we consider that contribute to ‘good data’ are described in this chapter.

|  |  |  |
| --- | --- | --- |
| **Good data are…** | **Considerations** | **Questions we may ask** |
| Usable: fit for purpose | 3.1.1 Well described  3.1.2 Include uncertainties/limitations  3.1.3 Readable  3.1.4 FAIR (Findable, Accessible, Interoperable, Reusable)  3.1.5 Reproducible  3.1.6 Timely  3.1.7 Appropriately licenced | - Is the purpose of the dataset well defined?  - Are these the best data for the task?  - Are the data well described, including limitations and uncertainties?  - Is the dataset discoverable, accessible, and readable?  - Are the data reproducible?  - Is the method by which open data was produced also open? |
| Collected with respect to… | 3.2.1 humans and their rights  3.2.2 the natural world | - Was the data collected/produced for this purpose, not incidentally? |
| Published | 3.3.1 with respect to openness  3.3.2 maintaining privacy  3.3.3 carrying owner licensing | - Is the dataset published with a DOI and version?  - Does the data carry an appropriate licence? |
| Revisable | 3.4.1 Personal: opt-in/out alternatives  3.4.2 Long term accuracy: data may change over time  3.4.3 Older versions of data may be decommissioned | - For human-related data, could participants realistically opt-out?  - Are the data time dependent? |
| Form useful social capital | 3.5 Valuable to society  ‘FAIR’, persistent, open  Available for ethical use | - Have we considered ethics around the data? |

Table 1: Guidelines for ‘good data’.

The above questions might assist data producers and consumers to consider the impacts on their research outcomes.

## 3.1 Good Data are fit for Purpose

For data to be ‘good’ they must meet the consumer’s needs (and indeed, the data producer’s needs). This may include measuring the right things, measuring appropriately, and with an understanding of uncertainties in the data (that is, data is accurate, and data has an appropriate level of precision). Data should be internally consistent within the dataset (e.g. angles should consistently be measured in degrees *or* radians), or in the case of non-numeric data, language uncertainties should be minimised (e.g. is a survey conducted in the native tongue of the respondents? Could there be translation uncertainties between respondents and analysts?). When possible, Controlled Vocabularies - which provide taxonomies for cataloguing information in particular fields - should be used.

Data that is created to answer a specific question may be more useful than ‘incidental’ data. Use of data should be ‘defensible’, i.e. it is demonstrable that the data can be validly used for its primary, or secondary, purposes. For example, population biases exist in social media analytics that could be deliberately avoided in constructing random populations for surveying, but the data may be much harder to collect. Similarly, in marine science when using tracers of a measurable quantity to infer values of another quantity, how strongly correlated is the effect? Is measuring the desired variable possible? What is the uncertainty associated with use of a notionally equivalent metric?

### 3.1.1 Good Data are Well Described

Good data need to have a plan, and be curated according to that plan, which may change over time.[[8]](#footnote-8) Datasets contain, or are accompanied by, *metadata* describing the dataset. This metadata must contain a description of how the dataset was created - measurement techniques, associated uncertainties, and dataset provenance. It should also provide transparency: who funded and collected the data, what was the purpose, any post-processing steps that have occurred, when the original data was collected, and when the data product was created, as well as versioning information if the dataset has been updated or re-produced. Metadata must be accessible in a machine-readable format, but good metadata are also human-readable. Google have recently released a Dataset Search tool which relies on the use of open metadata standards in contributing datasets.[[9]](#footnote-9)

### 3.1.2 Good Data Include Uncertainty Estimates or Description of Limitations

Good data are self-deferential, datasets are open about limitations. Every observation has some uncertainty. Good data describe uncertainties, and are not asserted as canonical truth. Data collectors must consider all sources of error and uncertainty in their data, and ensure this information is available to downstream consumers of the data.

### 3.1.3 Good Data are Readable

Good Data are written in common Open formats with standards-governed specifications. If proprietary formats must be used to meet community standards, thought is given to how the data should be accessed in 10 years’ time when standard software may well have changed. Open formats include .txt, .csv, .html, .mp3, .gzip, .png, .geoTIFF, .nc, .hdf, and more. Because open formats adhere to published standards, maintaining support for these formats in the future is easier than when formats are proprietary.  
  
Examples of closed or proprietary data formats that may also be appropriate to use due to their ubiquitous community uptake include .pdf, .doc, .xls, .gif. The vast number of files written in these formats means that backward compatibility for these file types is likely to persist for some time.

### 3.1.4 Good Data do better than FAIR

Data should be *findable*. Digital Object Identifiers (DOIs) should be minted for datasets.

Data should be *accessible*. Not hidden behind paywalls or obscure metadata redirection servers.

Data should be *interoperable*. Data can be meaningfully read by multiple, ideally non-proprietary software products.

Data should *reusable*. Data are available for use.  
Data should be *ethical*. No entity will be harmed in the collection or use of the data.

Data should be *revisable*. Errata can be filed, corrections made, and updated versions released with older versions archived.

The FAIR data principles have been widely adopted among public data providers including data.gov.au, and online assessment tools have been developed to rate compliance with FAIR principles.[[10]](#footnote-10) However, FAIR data may not be ‘good’. Adding the terms ‘ethical’ and ‘revisable’ to make FAIRER data is a good step - but may still be applied to data which miss metadata; are not appropriately licensed; which do not describe uncertainties or whose definition of ‘ethical’ may differ from a common usage concept.

### 3.1.5 Good Data are Reproducible

If a dataset cannot be reproduced (other than time-dependent observations of a dynamic system), its value is severely limited. Reliability of data depends on its reproducibility.

* For medical and social research, have the populations tested been constructed such that the same results would be produced if the study was repeated using appropriately sampled individuals?
* For survey data, will the same people give the same answers if surveyed again (barring interventions designed to change responses)? Were populations appropriately randomly selected? Could language barriers contribute issues around reproducibility?
* For observed data, will instruments produce consistent results (i.e. are they well calibrated)?
* For modelled data, is the code that produced it open source and available to be run on other systems by other people; and can the code be associated with the data it produced?[[11]](#footnote-11)
* For data collected by ‘apps’, is the software used by the app open, versioned, and adhering to relevant standards? Are algorithms published?

Reliability outside of a laboratory context with controlled conditions takes different meanings for different fields. For example, a satellite image captured at position *(x,y)* at time *t=0* will never be repeated at a later time. But if a static object can be reliably identified, measurements of that object should return consistent results.  
  
The question of reproducibility is difficult in non-controlled, time-dependent data collection – a feature of many social media or personal geolocation platforms. In these scenarios, reliability may be cast as capacity to understand the conditions and algorithms which led to the data being collected/created.

Additionally, data which cannot be reproduced for ethical reasons (for example, experiments on humans in war time),[[12]](#footnote-12) could be open and published and may contribute important understandings of e.g. human physiology, but the data are limited by the fact that they must not be reproduced. In other words, data could be FAIR but that does not make it ‘good’.  
  
“Good data” can be recreated when conditions of their original formation can be controlled.

### 3.1.5 Good Data are Timely

Data should be released to primary consumers as soon as possible after collection. They should still be relevant when they are released to the wider community. It may be reasonable to issue a metadata record rapidly after data collection with the accompanying data to be published later. The risks of early release (e.g. needing to correct errors in the data) are important, and quality checking and control is crucial, as is anonymising human data. Sometimes the funding associated with data acquisition necessitates an embargo period for the funder to maximise return on investment in the early period, however there is likely to be broader benefits to releasing data to the community that will indirectly benefit the funder. Delaying release means the data may no longer be fit for use.

### 3.1.6 Good Data are Appropriately Licensed

Ambiguity about how data may be used (assuming ethical and anonymity criteria are met) is resolved by application of an appropriate license. In Australia, without a license, data are considered ‘all rights reserved’[[13]](#footnote-13) which may not be the intention of the data publisher. A license may not meet the desires of all parties wishing to re-use data - but it must, at least, resolve ambiguity about whether data can be used for a particular purpose.

A range of creative commons (CC-BY) licenses are available for canonical data (data which are generally immutable, e.g. observations of temperature). Where data may be edited, or version-controlled, software licenses may be more appropriate (e.g. https://choosealicense.com). The terms of software licenses are better designed for the case where the licensed collection may evolve over time – and may be more permissive of contributions. Licensing is the collective decision of the data producers, owners, and custodians, who should become familiar with the various licenses available.

## 3.2 Good Data Respects Rights

Good data are collected with respect to various rights, including human rights, property rights, and privacy rights.[[14]](#footnote-14) Here we take ‘rights’ to mean human rights in a broad sense (privacy, freedom to live without harassment) and the rights of the natural world, in the sociocultural context at the time of collection. Other principles that may be considered are ‘responsibilities’ of the data collector - sustainability (environmental and financial cost of maintaining the data), fairness/equitability, and path to impact; as well as ‘value’, which may have a range of meanings.

### 3.2.1 Human Privacy

The Australian Government recently tightened requirements around reporting of data breaches which could impact on privacy (via the *Privacy Amendment (Notifiable Data Breaches) Act 2017*).[[15]](#footnote-15) Organisations storing personal data are required to strengthen security and establish/review data breach reporting processes. While not protecting privacy directly, it enables citizens to know when their personal data may have been breached. Significant numbers of breaches have been reported to date across government and industry.[[16]](#footnote-16)

The European Union have legislated for greater control of personal data for their citizens via the General Data Protection Regulation (GDPR).[[17]](#footnote-17) The GDPR sets a high standard for data privacy. The changes have had wide-reaching impacts on mobile apps and websites that collect data.

### 3.2.2 The Natural World

Data collection may have implications for the natural world. Consider the possible impacts of increasingly high resolution remote-sensed imagery. There are implications of sensor data being used in detection and monitoring of threatened species: there are conservation benefits, but potential black-market risks. If it were deemed that fossil fuel extraction harms the planet, this affects models and observational datasets used to detect such deposits. This is of particular interest in heretofore unspoilt wilderness areas such as Antarctica.[[18]](#footnote-18)

The physical cost of holding, cataloguing, accessing and processing data is important. Infrastructure scale computing, data storage and data retention is expensive; as is large scale data transmission and energy infrastructure. Costs of unnecessary data collection and retention are ethical and environmental, as well as financial.[[19]](#footnote-19) We should aim to minimise the impact of data collection and retention on the natural world.

## 3.3 Good Data are Published

If data remain solely accessible by an individual, group, or corporation, then utility and defensibility of process and products is limited. Good data are ‘as open as possible’; ranging from CCBY-4 licensing to pay-per-access. The utility (or ‘goodness’) of any data is always restricted if nobody knows they exist, so publication of metadata, even without the accompanying data, is very important. There can be no means of assessing data quality and veracity if the data are not published and comparable with other datasets.  
  
There must be exceptions where privacy or ethics are concerned to maintain the rights of the data subject(s), but aggregated and de-identified data (which cannot be re-identified) should be published where possible.

Publishing data is not simply making it accessible via a web server. If datasets are formally published, DOIs can be minted, increasing the value of the data as a citeable resource. In turn this provides incentive for academic data collectors to release valuable data collections, as it helps researchers to accrue credit in the academic system. DOIs also provide a permanent metadata record for the dataset.

### 3.3.1 Open Data should be Published Openly

Data should be openly available in compliance with FAIR data principles. Licence restrictions may be required, but the barrier to access should be low, to encourage reuse. Data access should not have a monetary cost, as exemplified by the Australian Government’s Research Data Storage Infrastructure scheme (2010-2015), which provided infrastructure and support for nationally significant data collections. Cost recovery models are understandable, but the data should be accessible by all people.[[20]](#footnote-20) Publicly funded research should be published openly, both data and journal articles.[[21]](#footnote-21) Ideally, ‘for profit’ or ‘commercial’ data should also be available for discovery in an open fashion.

### 3.3.2 Published Data must Maintain Privacy

It is important that no individual (person, government, corporation, or the Earth itself) should have their privacy damaged as a result of data publication. Significant thought has been given to data privacy in medical contexts.[[22]](#footnote-22) However, emerging fields in ubiquitous data are only now facing these issues, prompted in part by the EU’s GDPR, and by the public response to data breaches and unexpected information sharing revealed in the media. Data released in Strava’s Heatmap demonstrated an unexpected privacy violation;[[23]](#footnote-23) similarly the Earth and humanity may be better served if the location of critically endangered species (e.g., the Wollemi Pine)[[24]](#footnote-24) is kept secret.

### 3.3.3 Published Data should Carry Owner Licensing

The data owner should be able to decide what license terms they apply to their data and may wish to control limitations about what users can do with the data, and whether they can profit from modifications to it. Attribution should not be lost when data is used for downstream applications, and derived data products should contain acknowledgement of or reference to the parent dataset(s). In Australia, CC-BY licensing does not compromise moral rights under copyright law.[[25]](#footnote-25)

## 3.4 Good Data are Revisable

Data may have a lifespan. It is important that datasets are maintained, reviewed periodically and retired to an archive if need be. In the built environment, Ingress (www.ingress.com) is an Augmented Reality mobile game built around user-identified ‘portals’ attached to real world objects, in which users are able to submit edits and corrections, and flag portals for removal if the object in the real world no longer exists.   
  
Humans have successfully asserted that their personal data be ‘forgotten’,[[26]](#footnote-26) arguing that the right to privacy over-rides historical record. In these cases, electronic records of history are revised – not necessarily removed – with aggregate or non-personal data and information potentially retained for historical purposes.

### 3.4.1 Personal Data should be Opt-in

Keßler and McKenzie construct a *Geoprivacy Manifesto* designed to address an internet-age trend of personal data collection *and storage* being the default for many web-based services.[[27]](#footnote-27) This leads to potential exposure of individuals through various inadvertent means - their location, which services they used, how they communicated. In a good data scenario, these data would be retained just long enough: exploited at the time of usage, then forgotten, unless the individual expressly opted for their data to be retained.

In the context of ubiquitous data collection about individuals, ‘good data’ respects the right to be forgotten.[[28]](#footnote-28) Should records of a person attending a specific place be retained? What if the person may be unaware that their location is being recorded, or if cultural sensitivity requires consideration of deceased persons? There may be ethical advantages to data retention which appropriately considers privacy, e.g. notification of people who may have been exposed to a toxin because they visited an infected location in a given timeframe.

### 3.4.2 Good Data may Change over Time

The world is not static. Considering geospatial datasets for example, population distributions change over time which may diminish accuracy and trustworthiness of an ecological dataset. Projects like the Atlas of Living Australia help researchers track these changes.[[29]](#footnote-29) The built environment changes at an even higher rate. In response, Australian state mapping authorities release regularly updated maps to ensure data is effectively versioned, and consumers can access the most accurate data available. Conversely, the use of an older map collection implies a level of uncertainty with respect to present day locations. In other words, the ‘goodness’ of geospatial data may decay over time.

Data formats and conventions can also change specification over time, and good datasets may need to be updated to meet new standards or be converted to new formats to maintain interoperability.

### 3.4.3 Older Versions of Data may be Decommissioned

Good data are versioned and timestamped, so that when data become un-useful they can be decommissioned. This does not mean erased - historical data contributing social capital may be retained - simply removed from active usage.

This highlights the need for versioned datasets. Real-world applications may be interested in the present state of geospatially referenced objects, or they may be interested in changes over time, for example land use, coastlines, or urban development (e.g. an historical New York cartography project and mapping of historical photos).[[30]](#footnote-30)

Dataset production should consider the valid lifetime of its data, and if it can change over time, how data should be marked as out-of-date. For example some states of Australia have a fuel price monitoring body, but due to rapid fluctuations in fuel prices, these products may flag some locations as being inaccurate due to age. Data should not be thrown away without good cause. Older versions of datasets should be kept for research purposes at a later date to study trends, for example.

## 3.5 Good Data form Useful Social Capital

‘Social capital’ aims to describe a collection of resources and networks that enable social interaction and understanding[[31]](#footnote-31). A concise definition might specify ‘*the shared values and understandings that enable people to trust each other.*[[32]](#footnote-32)

Good data are a social asset, a platform upon which trust and cooperation can be built, enabling a ‘social license to operate’. This may occur between scientists – for example field scientists collecting measurements and providing them to systems modellers; or anyone – for example, trusting a restaurant address is correct, or a weather forecast is mostly accurate.

Even private or sensitive data form a useful social asset - while they may not be openly accessible, ‘good data’ are reliable and their veracity can be examined by whoever has the appropriate permission to use it.

# Examples of Good Data and Conclusion

To the authors, a dataset is ‘good’ if it can reasonably address the questions suggested in Table 1. We do not expect a ‘good’ dataset to be able to ‘tick every box’, indeed that may be logically impossible for some data, our thesis here is aimed at helping data producers and consumers think qualitatively about the goodness of their data. Quantitative measures of FAIR data principles exist, but we hope to encourage data practitioners to step beyond these metrics.

The following four examples represent, by these guidelines, ‘good data’.

1. Data can have power in numbers. Not only in the literal sense, rather, just as repeatability is important, so aggregation and meta-analysis of repeated and comparable studies acts to reduce the uncertainty of individual studies. Cochrane reviews in medical research carry a good deal of weight for this reason and are considered ‘gold standard’.[[33]](#footnote-33) These reviews reduce the influence of individual companies or vested interests, and lead to more informed health policy.
2. A spatial dataset which meets all relevant criteria is the National Public Toilet Map,[[34]](#footnote-34) available through data.gov.au. This dataset contains historical versions allowing changes to be monitored over time. It meets ‘good data’ requirements around publication, licensing, availability, versioning, formats (both human- and computer readable), and it also forms useful social capital. This data enables apps to be built,[[35]](#footnote-35) which are of high impact to families with infants, people with medical continence issues for whom knowing where these facilities are can be vital, and accessibility information benefits those with reduced mobility. The dataset contains not only the location of public toilet facilities, but metadata about the facilities.
3. OpenStreetMap (www.openstreetmap.org) is a geospatial dataset forming a social asset. It is built by volunteers and geospatial professionals alike, maintained by an active community, editable by anyone, and governed by a code of conduct.
4. The Australian Electoral Commission (AEC),[[36]](#footnote-36) and state-based electoral commissions, provide polling data that is open and accessible down to polling place level (without compromising privacy, though a voter can choose to vote below the line in a Hare Clark election such that their vote may be uniquely identifiable, due to the possible permutations of preferences combined with small numbers of persons voting at each booth). Transparency in democracy is a powerful thing.

# Collected Thoughts on ‘Good Data’

Posing a question[[37]](#footnote-37) about what make ‘good data’ resulted in the themes ‘consistency’ – agreement about what to call things between data providers, ‘accessibility’ with regard to cost and licensing, and ‘provenance’ – knowing where data come from. In a largely geospatial cohort, ethical aspects were not widely discussed. One person we spoke to likened ‘good data’ to edge pieces of a jigsaw puzzle. Every piece of data is like part of the puzzle that is how the world works (and some are poorly formed!), but good data are the pieces that allow you to constrain the others.

In the context of developing countries,[[38]](#footnote-38) another contributor noted considerations including faith in the data collectors – data not hearsay; how money sponsoring data collection is spent – bias that might result in donor organisations being seen to do good things; is government intimidation a concern? Consistency within data records can be problematic, and survey responses can vary as those interviewed may say what they think the interviewer wants to hear.

# Conclusion

As well as the EU’s GDPR, we are aware of a Manifesto for Data Practises,[[39]](#footnote-39) a Code of Ethics for Data Science,[[40]](#footnote-40) a Geoprivacy Manifesto,[[41]](#footnote-41) and an Open Geospatial Consortium and World Wide Web Consortium note on ‘Spatial Data on the Web Best Practices’. [[42]](#footnote-42) The proliferation of these considerations suggests that in the near future, these ideas may crystallize into formal guidelines just as medical ethics did during the 20th Century.[[43]](#footnote-43)

The size and variety of data created and available continues to grow, and we are moving from ‘the 3 V’s of Big Data’ to 10 V’s[[44]](#footnote-44) (or even 17![[45]](#footnote-45)). In the context of ‘good data’ we need to consider validity, veracity, volatility, vulnerability, and value. This manifesto outlines the concepts that we believe act toward making data good.  
  
Why are all data not inherently good data? And what are the risks of creating data which are not designed with ‘good data’ principles in mind? These questions are critical, but the possible answers too numerous to be included in this introduction of what we think ‘good data’ could be. As you consider this book, we invite you to reflect on those questions in your own data environment.

As data producers and consumers, we challenge ourselves to consider the principles presented in Table 1 in our work and encourage others to do the same. We only see part of the complete ‘data picture’ ourselves; and the picture changes more rapidly than we can keep pace with. It is our hope that this chapter inspires discussion and reflection on what ‘good data’ means to you.

# Acknowledgements

We thank our employers for flexibility to contribute to this project, and the organisers for promoting the discussion. Thanks to those who provided valuable insights in writing this chapter, including Dr Rowan Martin-Hughes, Lanes Koeu, and Twitter users @SimoneMantovan3, @bestqualitycrab, and @ellenbroad. Thanks to John Morrissey, Gareth Williams, Tristan Kenderdine, and the chapter reviewers for critical feedback. Claire Trenham is supported by the CSIRO Climate Science Centre. We are thankful for experience in Open Data resulting from work with the National Computational Infrastructure (NCI) through the Research Data Services (RDS) project; both supported by the Australian Government.

# References

@adamdsteer, Twitter post, 18 January 2018, https://twitter.com/adamdsteer/status/953792786607742977.

Arockia, P. S., S.S. Varnekha and K.A. Veneshia. ‘The 17 V’s of Big Data’. *International Research Journal of Engineering and Technology (IRJET)*, 4.9 (2017): 329-333, https://irjet.net/archives/V4/i9/IRJET-V4I957.pdf.

Atlas of Living Australia, https://www.ala.org.au/.

Australian Electoral Commission, https://www.aec.gov.au.

Australian National Data Service [ANDS], ‘Licensing and copyright for data reuse’, https://www.ands.org.au/working-with-data/publishing-and-reusing-data/licensing-for-reuse.

Australian Research Data Commons [ARDC], ‘FAIR data assessment tool’, https://www.ands-nectar-rds.org.au/fair-tool.

Brody, H., S.E. Leonard, J-B. Nie and P. Weindling. ‘United States Responses to Japanese Wartime Inhuman Experimentation after World War II: National Security and Wartime Exigency’, *Cambridge Quarterly of Healthcare Ethics* 23.2 (2014): 220-230, DOI: 10.1017/S0963180113000753.

(Cambridge Analytica series), *The Guardian*, 2018, https://www.theguardian.com/uk-news/cambridge-analytica.

Claridge, T. *Social Capital and Natural Resource Management: An important role for social capital?,* Unpublished Thesis, University of Queensland, Brisbane, Australia, 2004, https://www.socialcapitalresearch.com/literature/definition/.

Cochrane Collaboration, http://www.cochrane.org.

data.gov.au, ‘National Public Toilet Map’, *Australian Government data.gov.au*,2018, https://data.gov.au/dataset/553b3049-2b8b-46a2-95e6-640d7986a8c1.

Dixon, S. and L. Quirke. ‘What’s the harm? ...’, *Teaching Sociology* 46.1 (2018):12-24, DOI: 10.1177/0092055X17711230.

DroneRules.eu, ‘8 Data Protection Guiding Principles under the GDPR for Drone Pilots’, 23 May 2018, http://dronerules.eu/en/professional/news/8-data-protection-principles-under-the-gdpr-for-drone-pilots.

European Commission, ‘Creating value through open data’, *EU Publications*, 2015, DOI: 10.2759/328101.

European Commission, ‘2018 reform of EU data protection rules’, 2018, https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules\_en.

Farrell, P. ‘Darknet sale of Medicare data ‘traditional criminal activity’, minister says’, *The Guardian*, 4 July 2017, https://www.theguardian.com/australia-news/2017/jul/04/federal-police-asked-to-investigate-darkweb-sale-of-medicare-data.

Firican, G. ‘The 10 V’s of Big Data’, *Transforming Data with Intelligence*, 8 February 2017, https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx.

Floridi, L., S. Kauggman, L. Kolucka-Zuk, F. La Rue, S. Leutheusser-Schnarrenberger, J-L Piñar and J. Wales. ‘The Advisory Council to Google on the Right to be Forgotten’, 6 February 2015,https://static.googleusercontent.com/media/archive.google.com/en//advisorycouncil/advisement/advisory-report.pdf.

IEEE, ‘Ethically Aligned Design’, IEEE Global Initiative, v2, December 2017, https://ethicsinaction.ieee.org/.

Irving, D. ‘A minimum standard for publishing computational results in the weather and climate sciences’, *BAMS*, 2015, DOI: 10.1175/BAMS-D-15-00010.1.

Katerbow, M. and G. Feulner. ‘Recommendations on the development, use and provision of Research Software’, *Zenodo*, 16 March 2018, DOI: 10.5281/zenodo.1172988.

Keßler, C. and G. McKenzie. ‘A geoprivacy manifesto’, *Transactions in GIS*, 22 (2018) :3–19, DOI: 10.1111/tgis.12305

Kidd, J. ‘Wollemi Pines: Secret plantation in Blue Mountains to ensure species’ survival’, *ABC News,* 21 September 2014. http://www.abc.net.au/news/2014-09-21/secret-bid-to-save-prehistoric-wollemi-pines/5758542.

Manifesto for Data Practices. 2018, https://datapractices.org/manifesto/.

Mapillary. ‘Accurate Privacy Blurring at Scale’, *The Mapillary* Blog, 19 April 2018, https://blog.mapillary.com/update/2018/04/19/accurate-privacy-blurring-at-scale.html.

National Computational Infrastructure Australia [NCI]. ‘NCI Annual Report 2016-2017’, 2017, https://nci.org.au/wp-content/uploads/2014/12/NCI-Report-2017-web-sm-updated.pdf, p.65.

Noy, N. ‘Making it easier to discover datasets’, *Google Blog*, 5 September 2018, https://www.blog.google/products/search/making-it-easier-discover-datasets/.

New York Public Library [NYPL]. ‘NYC Space/Time Directory’, *New York Public Library*, http://spacetime.nypl.org/.

\_\_\_\_\_. ‘OldNYC: Mapping historical photos from the NYPL’, *New York Public Library*, https://www.oldnyc.org/.

OAIC. ‘Notifiable Data Breaches scheme’, *Australian Government Office of the Australian Information Commissioner*, https://www.oaic.gov.au/privacy-law/privacy-act/notifiable-data-breaches-scheme.

\_\_\_\_\_. ‘Annual Report 2016-2017’, *Australian Government Office of the Australian Information Commissioner*, 2017, https://www.oaic.gov.au/annualreport2016-17, p.80.

\_\_\_\_\_. ‘Notifiable Data Breaches Quarterly Statistics Report: 1 April – 30 June 2018’, *Australian Government Office of the Australian Information Commissioner*, 2018, https://www.oaic.gov.au/resources/privacy-law/privacy-act/notifiable-data-breaches-scheme/quarterly-statistics/notifiable-data-breaches-quarterly-statistics-report-1-april-30-june-2018.pdf.

\_\_\_\_\_. ‘Health Information and medical research’, *Australian Government Office of the Australian Information Commissioner*, https://www.oaic.gov.au/privacy-law/privacy-act/health-and-medical-research.

OECD, ‘OECD Insights: Human Capital’, *Organisation for Economic Cooperation & Development*, 2007, https://www.oecd.org/insights/37966934.pdf.

O’Leary, D.E. ‘Ethics for Big Data and Analytics’, *IEEE Intelligent Systems*, 31.4 (2016): 81-84, DOI: 10.1109/MIS.2016.70.

Patil, DJ. ‘A code of Ethics for Data Science’, *Medium*, 1 February 2018, https://medium.com/@dpatil/a-code-of-ethics-for-data-science-cda27d1fac1.

Pérez-Peña, R. and M. Rosenberg. ‘Strava Fitness App Can Reveal Military Sites, Analysts Say’, *NYTimes*, 29 January 2018, https://www.nytimes.com/2018/01/29/world/middleeast/strava-heat-map.html.

Science Europe. ‘cOAlition S’, September 2018, https://www.scienceeurope.org/coalition-s/.

Secretariat of the Antarctic Treaty. ‘The Protocol on Environmental Protection to the Antarctic Treaty’, 1991, https://www.ats.aq/e/ep.htm.

Serwadda, D., P. Ndebele, M.K. Grabowski, F. Bajunirwe and R.K. Wanyenze. ‘Open data sharing and the Global South - Who benefits?’, *Science*. 359.6376 (2018): 642-643, DOI: 10.1126/science.aap8395.

Tandy, J., L. van den Brink, and P. Barnaghi (eds). ‘Spatial Data on the Web Best Practices’, *W3C Working Group Note*, 2017, https://www.w3.org/TR/sdw-bp/.

Thinh. ‘CC, Open Access, and moral rights’, *Science Commons Blog,* 7 November 2007, http://sciencecommons.org/weblog/archives/2007/11/07/cc-oa-moral-rights/.

Uber. ‘Privacy Policy’, https://privacy.uber.com/policy.

UMKC. ‘History of Research Ethics’, *UMKC Office of the Vice Chancellor*, date unknown, http://ors.umkc.edu/research-compliance-(iacuc-ibc-irb-rsc)/institutional-review-board-(irb)/history-of-research-ethics.

United Nations [UN]. ‘Universal Declaration of Human Rights’, *United Nations General Assembly*, 1948, http://www.un.org/en/universal-declaration-human-rights/.

Violet Blue. ‘Strava’s fitness heatmaps are a “potential catastrophe”’, *Engadget,* 2 February 2018, https://www.engadget.com/2018/02/02/strava-s-fitness-heatmaps-are-a-potential-catastrophe.

Wang, J., B. Evans, L.A.I. Wyborn, K. Gohar, I. Bastrakova, C. Trenham and K. Druken. ‘The A, B, C, D, E, F of Multipurpose Active Data Management Plans to Facilitate Multiscale Interdisciplinary Science’. *eResearch Australasia 2015,* Brisbane, Australia, 2015, https://www.researchgate.net/publication/309704990\_The\_A\_B\_C\_D\_E\_F\_of\_Multipurpose\_Active\_Data\_Management\_Plans\_to\_Facilitate\_Multiscale\_Interdisciplinary\_Science.

Wikipedia contributors. ‘The right to be forgotten’, https://en.wikipedia.org/wiki/Right\_to\_be\_forgotten, accessed 13 March 2018.

Yu, J. and S. Cox.‘5-Star Data Rating Tool’, v3, *CSIRO Software Collection*, 2017, DOI: 10.4225/08/5a12348f8567b.

1. European Commission, ‘2018 reform of EU data protection rules’, 2018, https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules\_en. [↑](#footnote-ref-1)
2. DroneRules.eu, ‘8 Data Protection Guiding Principles under the GDPR for Drone Pilots’, 2018 http://dronerules.eu/en/professional/news/8-data-protection-principles-under-the-gdpr-for-drone-pilots; ‘Accurate Privacy Blurring at Scale’, *The Mapillary* Blog, 2018, https://blog.mapillary.com/update/2018/04/19/accurate-privacy-blurring-at-scale.html. [↑](#footnote-ref-2)
3. ‘Uber Privacy Policy’, https://privacy.uber.com/policy. [↑](#footnote-ref-3)
4. Cambridge Analytica series, *The Guardian*, 2018, https://www.theguardian.com/uk-news/cambridge-analytica. [↑](#footnote-ref-4)
5. ‘IEEE Ethically Aligned Design’, *IEEE Global Initiative*, v2, 2017, https://ethicsinaction.ieee.org/. [↑](#footnote-ref-5)
6. D. E. O’Leary, ‘Ethics for Big Data and Analytics’, *IEEE Intelligent Systems*, 31.4 (2016): 81-84, DOI: 10.1109/MIS.2016.70. [↑](#footnote-ref-6)
7. S. Dixon and L. Quirke, ‘What’s the harm?...’, *Teaching Sociology*, 46.1 (2018): 12-24, DOI: 10.1177/0092055X17711230. [↑](#footnote-ref-7)
8. J. Wang et al, ‘The A, B, C, D, E, F of Multipurpose Active Data Management Plans…’, *eResearch Australasia,* Brisbane, 2015, www.researchgate.net/publication/309704990\_The\_A\_B\_C\_D\_E\_F\_of\_Multipurpose\_Active\_Data\_Management\_Plans\_to\_Facilitate\_Multiscale\_Interdisciplinary\_Science. [↑](#footnote-ref-8)
9. N. Noy, ‘Making it easier to discover datasets’, *Google Blog*, 2018, https://www.blog.google/products/search/making-it-easier-discover-datasets/. [↑](#footnote-ref-9)
10. ARDC, ‘FAIR data assessment tool’, https://www.ands-nectar-rds.org.au/fair-tool; ‘Accurate Privacy Blurring at Scale’, *The Mapillary Blog*, 2018, https://blog.mapillary.com/update/2018/04/19/accurate-privacy-blurring-at-scale.html. [↑](#footnote-ref-10)
11. D. Irving, ‘A minimum standard for publishing computational results in the weather and climate sciences’, *BAMS*, 2015, DOI: 10.1175/BAMS-D-15-00010.1; M. Katerbow & G. Feulner, ‘Recommendations on the development, use and provision of Research Software’, *Zenodo*, 16 March 2018, DOI: 10.5281/zenodo.1172988. [↑](#footnote-ref-11)
12. H. Brody et al, ‘United States Responses to Japanese Wartime Inhuman Experimentation after World War II: National Security and Wartime Exigency’, *Cambridge Quarterly of Healthcare Ethics*, 23.2 (2014): 220-230, DOI: 10.1017/S0963180113000753. [↑](#footnote-ref-12)
13. ANDS, ‘Licensing and copyright for data reuse’, https://www.ands.org.au/working-with-data/publishing-and-reusing-data/licensing-for-reuse. [↑](#footnote-ref-13)
14. UN, ‘Universal Declaration of Human Rights’, 1948, http://www.un.org/en/universal-declaration-human-rights/. [↑](#footnote-ref-14)
15. OAIC, ‘Notifiable Data Breaches scheme’, *Australian Government Office of the Australian Information Commissioner*, https://www.oaic.gov.au/privacy-law/privacy-act/notifiable-data-breaches-scheme. [↑](#footnote-ref-15)
16. OAIC, ‘Annual Report 2016-2017’, *Australian Government Office of the Australian Information Commissioner*, 2017, https://www.oaic.gov.au/annualreport2016-17, p.80; P. Farrell, ‘Darknet sale of Medicare data “traditional criminal activity”, minister says’, *The Guardian*, 2017, https://www.theguardian.com/australia-news/2017/jul/04/federal-police-asked-to-investigate-darkweb-sale-of-medicare-data; OAIC, ‘Notifiable Data Breaches Quarterly Statistics Report: 1 April – 30 June 2018’, 2018, https://www.oaic.gov.au/resources/privacy-law/privacy-act/notifiable-data-breaches-scheme/quarterly-statistics/notifiable-data-breaches-quarterly-statistics-report-1-april-30-june-2018.pdf. [↑](#footnote-ref-16)
17. European Commission, ‘2018 reform of EU data protection rules’, 2018, https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules\_en [↑](#footnote-ref-17)
18. Secretariat of the Antarctic Treaty, ‘The Protocol on Environmental Protection to the Antarctic Treaty’, 1991, https://www.ats.aq/e/ep.htm. [↑](#footnote-ref-18)
19. NCI, ‘NCI Annual Report 2016-2017’, 2017, https://nci.org.au/wp-content/uploads/2014/12/NCI-Report-2017-web-sm-updated.pdf, p.65. [↑](#footnote-ref-19)
20. European Commission, ‘Creating value through open data’, *EU Publications*, 2015, DOI: 10.2759/328101. [↑](#footnote-ref-20)
21. Science Europe, ‘cOAlition S’, 2018, https://www.scienceeurope.org/coalition-s/. [↑](#footnote-ref-21)
22. OAIC, ‘Health Information and medical research’, *Australian Government Office of the Australian Information Commissioner*, https://www.oaic.gov.au/privacy-law/privacy-act/health-and-medical-research. [↑](#footnote-ref-22)
23. R. Pérez-Peña & M. Rosenberg, ‘Strava Fitness App Can Reveal Military Sites, Analysts Say’, *NYTimes*, 29 January 2018, https://www.nytimes.com/2018/01/29/world/middleeast/strava-heat-map.html; Violet Blue, ‘Strava’s fitness heatmaps are a “potential catastrophe”’, *Engadget,* 2 February 2018, https://www.engadget.com/2018/02/02/strava-s-fitness-heatmaps-are-a-potential-catastrophe. [↑](#footnote-ref-23)
24. J. Kidd, ‘Wollemi Pines: Secret plantation in Blue Mountains to ensure species’ survival’, *ABC News.* 21 September 2014, http://www.abc.net.au/news/2014-09-21/secret-bid-to-save-prehistoric-wollemi-pines/5758542. [↑](#footnote-ref-24)
25. Thinh, ‘CC, Open Access, and moral rights’, *Science Commons Blog,* 2007, http://sciencecommons.org/weblog/archives/2007/11/07/cc-oa-moral-rights/. [↑](#footnote-ref-25)
26. L. Floridi et al. ‘The Advisory Council to Google on the Right to be Forgotten’, 2015,https://static.googleusercontent.com/media/archive.google.com/en//advisorycouncil/advisement/advisory-report.pdf. [↑](#footnote-ref-26)
27. C. Keßler and G. McKenzie, ‘A geoprivacy manifesto’, *Transactions in GIS* 22:3-19 (2018), DOI: 10.1111/tgis.12305. [↑](#footnote-ref-27)
28. (The right to be forgotten), https://en.wikipedia.org/wiki/Right\_to\_be\_forgotten. [↑](#footnote-ref-28)
29. Atlas of Living Australia, https://www.ala.org.au/. [↑](#footnote-ref-29)
30. NYC Space/Time Directory, *New York Public Library*, http://spacetime.nypl.org/; OldNYC: Mapping historical photos from the NYPL, https://www.oldnyc.org/. [↑](#footnote-ref-30)
31. T. Claridge, *Social Capital and Natural Resource Management: An important role for social capital?*, Unpublished Thesis, University of Queensland, Brisbane, Australia, 2004, https://www.socialcapitalresearch.com/literature/definition/. [↑](#footnote-ref-31)
32. OECD Insights: Human Capital, *Organisation for Economic Cooperation & Development*, 2007, https://www.oecd.org/insights/37966934.pdf. [↑](#footnote-ref-32)
33. Cochrane Collaboration, http://www.cochrane.org. [↑](#footnote-ref-33)
34. data.gov.au, ‘National Public Toilet Map’, *Australian Government data.gov.au*,2018, https://data.gov.au/dataset/553b3049-2b8b-46a2-95e6-640d7986a8c1. [↑](#footnote-ref-34)
35. Web: https://toiletmap.gov.au/, also available on iTunes. [↑](#footnote-ref-35)
36. Australian Electoral Commission, www.aec.gov.au. [↑](#footnote-ref-36)
37. @adamdsteer, Twitter post, 18 January 2018, 11:54AM, https://twitter.com/adamdsteer/status/953792786607742977. [↑](#footnote-ref-37)
38. D. Serwadda et al, ‘Open data sharing and the Global South - Who benefits?’, *Science* 359.6376 (2018): 642-643, DOI: 10.1126/science.aap8395. [↑](#footnote-ref-38)
39. Manifesto for Data Practices, https://datapractices.org/manifesto/. [↑](#footnote-ref-39)
40. DJ Patil, ‘A code of Ethics for Data Science’, *Medium*, 1 February 2018, https://medium.com/@dpatil/a-code-of-ethics-for-data-science-cda27d1fac1. [↑](#footnote-ref-40)
41. Keßler and McKenzie, ‘A geoprivacy manifesto’. [↑](#footnote-ref-41)
42. J. Tandy, L. van den Brink, and P. Barnaghi (eds), ‘Spatial Data on the Web Best Practices’, *W3C Working Group Note*, 2017, https://www.w3.org/TR/sdw-bp/. [↑](#footnote-ref-42)
43. UMKC, ‘History of Research Ethics’, *UMKC Office of the Vice Chancellor*, date unknown, http://ors.umkc.edu/research-compliance-(iacuc-ibc-irb-rsc)/institutional-review-board-(irb)/history-of-research-ethics. [↑](#footnote-ref-43)
44. G. Firican, ‘The 10 V’s of Big Data’, *Transforming Data with Intelligence*, February 2017, https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx. [↑](#footnote-ref-44)
45. P.S. Arockia et al, ‘The 17 V’s Of Big Data’, *IRJET* 4.9 (2017), https://www.irjet.net/archives/V4/i9/IRJET-V4I957.pdf. [↑](#footnote-ref-45)