Guidelines for the Creation of Analysis Ready Data

Harriette Phillips[1*], Aiden Price[1], Owen Forbes[1], Claire Boulange[2], Kerrie Mengersen[1], Marketa Reeves[2], Rebecca Glauert[2]

[1]Queensland University of Technology–School of Mathematical Sciences 2 George St, Brisbane City QLD 4000, Australia
[2]University of Western Australia–School of Population and Global Health 35 Stirling Hwy, Crawley WA 6009, Perth 6009, Australia

Keywords: Analysis Ready Data; Digital Atlas; Good Practice; Data Collection

Citation: Phillips H., Price A., Forbes O., et al.: Guidelines for the Creation of Analysis Ready Data. Data Intelligence, 2026, Vol. 8, Art. No.: 20240007. DOI: https://doi.org/10.3724/2096-7004.di.2024.0007

ABSTRACT

Globally, there is an increased need for guidelines to produce high-quality data outputs for analysis. No framework currently exists that provides guidelines for a comprehensive approach to producing analysis ready data (ARD). Through critically reviewing and summarising current literature, this paper proposes such guidelines for the creation of ARD. The guidelines proposed in this paper inform ten steps in the generation of ARD: ethics, project documentation, data governance, data management, data storage, data discovery and collection, data cleaning, quality assurance, metadata, and data dictionary. These steps are illustrated through a substantive case study that aimed to create ARD for a digital spatial platform: the Australian Child and Youth Wellbeing Atlas (ACYWA).

1. INTRODUCTION

In almost every field of research there is an increased need to process quality-assured data for analysis [1, 2]. The production of analysis ready data (ARD) should adhere to any lawful and ethical requirements and be both interoperable with other datasets and throughout time [1, 3, 4, 5, 6]. ARD should also be created through minimal effort and allow for immediate analysis [1]. In response to the need for ARD, various criteria have emerged over the past five years [7, 8, 9]. These works typically focus on a specific subset of components of ARD development, without consideration of a holistic set of process

*Corresponding author: Harriette Phillips (Email: harriette.phillips@hdr.qut.edu.au; ORCID: 0000-0003-3001-8941).

© 2026 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

guidelines, causing issues in consistency and comparability [9, 10]. In this section, key articles in the literature are critically reviewed and summarised.

An article published in 2017 produced guidelines for three ARD steps: data discovery, data collection, and data cleaning [9]. These steps enable data to be collected and compiled into a standardised format before its final release. A later paper, published in 2018, largely focused on data discovery, data collection, and data storage addresses the ingestion and management of large volumes of data for social media analytics [11]. Ethics was also identified as an important issue in the production of ARD but was determined to be out of scope. In the following year, a book was published outlining guidelines for managing and sharing data [8]. This comprehensive resource includes extensive information on data collection, data cleaning, data governance, data management, data storage, ethics, and metadata. The book also outlined that ethics and legal requirements are dependent on the region in which the research is being conducted. Building on this work, a 2021 contribution focuses on the topics of metadata and data dictionary with the aim of creating sharable and transparent data that users can interpret easily and preferably be both human-readable and machine-readable [12]. This work can be further extrapolated through the principles that are aimed at making data findable, accessible, interoperable, and reusable [13]. An article published in the same year focused on building ARD through creating accessible data through open data products [2]. A recent article further extended the range of steps in the production of ARD, namely on the topics of general data collection and quality assurance [10]. They aimed to improve data quality by discussing quality assurance, data collection and ethics. The Committee on Earth Observation Satellites has also developed a framework for the development of ARD specifically for the use of satellite data [14]. This framework focuses on allowing for immediate analysis of ARD through a minimum set of specifications whilst following the core requirements for metadata.

Digital atlases are typically online, interactive tools that facilitate the display, inspection, and comparison of geospatial and temporal data [15]. The specific exemplar, used as a case study to develop the proposed guidelines, was the Australian Child and Youth Wellbeing Atlas (ACYWA) [16]. This digital atlas aims to improve the health and wellbeing metrics for Australian children and young people. Consequently, the production of high-quality ARD was essential for the production and interpretation of this digital atlas.

In summary, the literature to date suggests ten steps that should be considered in the creation of ARD: ethics, data governance, project documentation, data management, data storage, data discovery and collection, data cleaning, quality assurance, metadata, and data dictionary. Current literature does not consider these steps holistically to provide a comprehensive approach to producing ARD. This paper aims to develop ten key steps, summarised from current literature, which can then be used to create high-quality ARD outputs. The proposed guidelines will be investigated in 2 Guidelines for the Creation of Analysis Ready Data and contextualised, through a digital atlas, in 3 Guidelines in the Context of a Digital Atlas.

2. GUIDELINES FOR THE CREATION OF ANALYSIS READY DATA

The proposed guidelines were compiled into a comprehensive approach for processing, organising, developing, and producing ARD. Each principle (also referred to as steps) is discussed in detail in the subsequent subsections: 2.1 Ethics, 2.2 Project Documentation, 2.3 Data Governance, 2.4 Data Management, 2.5 Data Storage, 2.6 Data Discovery and Collection, 2.7 Data Cleaning, 2.8 Quality Assurance, 2.9 Metadata, and 2.10 Data Dictionary.

The ten principles that form the proposed guidelines for producing ARD are summarised in Figure 1 and Table 1. The sequential ordering of the steps was determined by the order necessary for the case study (ACYWA). The process typically starts with an ethics application, establishing project documentation, and leads into the other generation steps to create three outputs: metadata, data dictionary, and analysis ready data. Establishment of data storage is also required before data can be accessed. Thus, data storage needs to occur before data can be collected. Generation steps often occur concurrently with one another (e.g., data cleaning and quality assurance). These steps can also be repeated and updated as required. This repetition is represented by the dotted line shown in Figure 1.

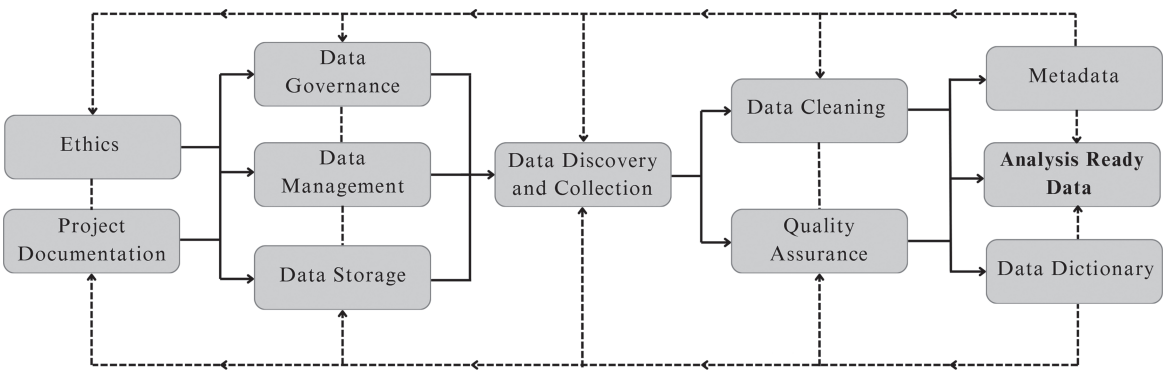


Figure 1. Visualisation of the Ten Guidelines for the Generation of Analysis Ready Data.

Table 1. Summary of Principles for the Guidelines in the Production of High-Quality ARD.

Principle	Consideration
Ethics	Ethics is typically established at the onset of a project where approval is sought for how methodologies will be conducted. This also governs other aspects of the project including, risks, participants, and research aims and outcomes. These legal and ethical requirements are region and context specific.
Project Documentation	Project documentation includes the research aims, decisions, and outputs for a given project which should be established at the onset of a project and maintained throughout the entirety of a project's progress.

Table 1. *Continued.*

Principle	Consideration
Data Governance	Data governance focuses on the protection of information surrounding data use through compliance with strategies to safeguard both personal and non-personal information from misuse and mishandling. This includes compliance with the safe use of data sharing and collecting.
Data Management	The management of data and all steps related to the creation of ARD (e.g., data storage, data collection, data cleaning, quality assurance, metadata, and data dictionary).
Data Storage	Safe storage of materials is related to the data, where backup, retention, and deletion strategies should be implemented.
Data Discovery and Collection	Data discovery is the initial phase of aligning data with project's methodologies and aims. Data is then collected through appropriate sources. These processes are well documented and should be guided by a data management plan.
Data Cleaning	Data cleaning occurs to compile data into a standardised format that does not contain identifiable information and is free of errors, inconsistencies, uncertainty, and noise.
Quality Assurance	Quality assurance is a process that occurs concurrently with other data processing steps to detect and resolve anomalies in data. This process ensures that all processes have been consistently applied during the data processing steps.
Metadata	Metadata refers to the creation of supplementary information about dataset characteristics, preferably in human and machine-readable format. This information is collated using both automatic and manual processes. This information is used to complement the data to increase its visibility and reusability.
Data Dictionary	Data dictionaries are created to also provide supplementary information, ease of understanding and interpretation according to commonly used terms or data vocabulary within a given field. As opposed to metadata, this information is often supplied on an individual data indicator level to provide context about data processes, data sources, and any other necessary contextual information required for any further data analytics.

2.1 Ethics

Ethics should be at the forefront of project development. Projects are obligated to conduct work in alignment with regional-specific legal and ethical requirements [1]. This is showcased through gaining ethical approval or clearance through an Ethics committee. Ethics often governs the entire process of project development (e.g., methodologies, data, participants, etc) where the approved ethics application typically outlines how the research should and will be conducted. The legal and ethical requirements should be researched before the commencement of a research project and abided by throughout the entire research process [8, 17]. Typically, ethics applications require review from a committee and adhere to an extensive list of criteria [18]. This is to ensure that the scope is achievable, often striving for improve-

ments in a given field, whilst safeguarding participants. These criteria can be used for ARD creation. This provides projects with the necessary resources to develop a project in alignment with the region-specific legal and ethical requirements. The region-specific criteria also vary depending on the field of research (e.g., human research). These criteria ensure that ARD is being created in line with the specifications needed for the field of research the project is being conducted in. Thus, ensuring that specific legal and ethical requirements of each criterion can be adhered to effectively and ethics can be subsequently approved for individual projects. Examples of region and field-specific criteria for the creation of an ethics application include The Belmont Report on the Ethical Principles and Guidelines for the Protection of Human Subjects of Research (Belmont Report) and The National Statement on Ethical Conduct in Human Research (National Statement). These criteria have been detailed below.

The American standard, The Belmont Report, focuses on three main topics for conducting ethical research involving human participants: respect for persons, beneficence, and justice. Respect for persons details that participants are able to make informed decisions about their involvement in the research process and their own best interests [19]. This is understood through two ethical convictions. The first, is that participants are autonomous and are entitled to be protected. Second, any participants with diminished autonomy are also entitled to be protected. Beneficence is aimed at minimising any harm and risks whilst maximising the benefits for participants who participated in the research. This is aimed at securing the well-being of participants through acts of kindness and charity that go beyond strict obligation. The last topic, justice, is that participants should be respected and treated equally.

Similarly, an Australian standard, the National Statement, can be used to develop, design, review, and conduct ethical human research. The guidelines in this approach require a review of values, judgement, principles, and context, depending on the level of risk to participants [20]. There are six topics in the national statement: merit and integrity in research, justice, beneficence, respect, risks and benefits, and general requirements for consent which are documented in Appendix 1. These topics provide a flexible framework for the design and review of ethical research.

2.2 Project Documentation

Projects need explanatory or contextual information surrounding data, procedures, decisions, and decision making processes to effectively conduct research [8]. The information surrounding data and processes is often referred to as project documentation [8, 21]. Project documentation is most easily maintained throughout the entirety of the research project when established at the outset of a project [8]. Furthermore, the inclusion of comprehensive documentation on how and why data were collected is critical for the short-, medium-, and long-term longevity of projects. Project documentation needs to be sufficient enough to ensure that data and project related materials are understandable, usable, and accessible to all users. For example, a data management plan (DMP) can be used to document the key information [22, 23]. This can include general project documentation such as documentation of decision making. Using a DMP to guide project documentation established a clear outline of project protocols. DMPs have been explored in further detail in 2.4

Data Management and Appendix 2. This documentation should also support effective strategies for maintaining the security of project information [21]. Effective security measures should include implementing version control when deleting and modifying records. These strategies should also include destruction methods after the project has been completed.

2.3 Data Governance

Data governance is achieved through the authoritative control and compliance with privacy, security, quality and ongoing data management in accordance with region-specific regulations [24, 25, 26]. These regulations aim to minimise risk and cost involved in data use whilst maximising the value of the data; and how it is used, shared, and collected [24]. Data governance is adopted to safeguard information minimising privacy concerns by preventing data from being misused or mishandled [25, 27]. Data governance is important for safeguarding both personal and non-personal information. Regulations differ depending on the type of data and the region in which the research is being conducted in [8, 25]. These regulations can be used to govern the quality of ARD. The regulations ensure that ARD is created and maintained at a high standard. Examples of these regulations include the General Data Protection Regulation (GDPR) and the Australian Code for the Responsible Conduct for Research (the Code) which are detailed below.

The GDPR is a European standard for regulating data governance strategies [28]. The GDPR was introduced to minimise the risk and cost involved in the collection, use, and sharing of personal data. This regulation provided a set of principles for data compliance and data protection, including data processing, lawfulness, consent rights, responsibilities, and documentation of activities [24, 26, 28]. The introduction of these stricter guidelines resulted in revisions by European researchers to how data is currently collected, shared, and used [25, 27, 28]. However, whilst the GDPR does provide great benefits and protections for personal data, the cost of compliance cannot be met at times due to insufficient resources [29, 30]. The GDPR is also a region-specific regulation, where there is the possibility of contextualising this approach to various regions, but compliance would not be required.

An alternative approach for managing the risk and cost relevant to data use is the Code. In contrast to the GDPR, this framework provides principles for the responsible conduct of research but does not incorporate laws, regulations, or guidelines [31]. The Code is reliant on the integrity of researchers and the expectation that principles and responsibilities for conducting honest and ethical research will be followed. However, failure to comply with the principles outlined in the Code can result in a breach or research misconduct. There are eight key principles outlined in the Code: honesty, rigour, transparency, fairness, respect, recognition, accountability, and promotion which are summarised in Appendix 3. The Code does not provide a comprehensive guideline for all aspects of data use. Howerver, it does use supporting documentation to provide researchers with expectations for thier research [31, 32]. There is limited current literature on the use of the Code where a recent study found that the principles contained within the Code were strongly endorsed [32].

2.4 Data Management

Effective management of large volumes of data is critical for individuals and organisations [22]. A Data Management Plan (DMP) is an effective way to summarise the key processes in the production of data [22, 23]. DMPs are often a research output requirement and having a well-managed and effective DMP can have a long-term term impact on the sustainability of a project. A detailed DMP enables the strategic implementation of good practices for generating data [8]. They provide documentation on the actual or expected aims so that they can be implemented with greater clarity of expectations and tasks. If a DMP is submitted to an ethics committee and approved this can validate that the project's aims will be ethically and legally sound and can be carried out for the duration of the project. Furthermore, developing a DMP can generate avenues for discussing issues, milestones, and progress. This will save time in the projection of ARD by reducing errors; duplication of efforts; assisting in the production of sharable data; improving data protection, security, quality; and promoting compliance with legal and ethical standards [22, 23]. A DMP generally involves key processes for ARD generation that should be documented and considered before the commencement of research. Appendix 2 summarises some considerations for the creation of an effective DMP based on current literature. This includes considerations for data discovery, data description, data collection, data organisation, and data preservation procedures, data storage, data volume, data loss procedures, data privacy and confidentiality, data ownership, quality assurance, documentation, dissemination, and metadata. It is also recommended that once a DMP has been established at the commencement of research and updated at least annually. The DMP outputs should be easily shareable with other researchers [33].

2.5 Data Storage

Implementation of effective storage plans and backup strategies can prevent the need to replicate research, minimise reputational risk, and maintain the privacy of research subjects [8]. Supporting documentation for The Code (see 2.3 Data Governance) outlines key storage, retention, and disposal strategies and should occur regardless of whether the project is occurring internally or externally. These policies include that projects should comply with current licensing and copyright requirements, and be conducted in accordance with research specific practices and standards [34]. The research should also be conducted in compliance with any ethical; privacy; and publication requirements, as well as compliance with laws; guidelines; and regulations. Furthermore, data storage should also be considered when completing a DMP [22, 23]. A DMP considers a wide range of items that capture how data should be stored. This ensures that, from the onset of a project, the data storage for a given project is apparent. The full list of considerations for a DMP has been detailed in Appendix 2.

Effective data storage, including avoiding data loss and maintaining data retention, is necessary beyond the immediate scope of a project [8]. This is crucial for a project's long-term success. Cloud-based storage is most commonly used with advanced technology and when large amounts of data are ingested. As such, it has recently gained popularity for use by both researchers and end users where data, documentation and information related to the research can be reliably accessed from any device, and large-scale data ingest is

supported [35]. Cloud-based storage also provides a flexible solution for cross-institutional access, is cost-effective, and provides larger storage capacities than other alternatives.

2.6 Data Discovery and Collection

Discovering and sourcing appropriate data should be achieved in alignment with project methodologies and aims. This step should provide adequate justification for the inclusion of given data before it is collected [11]. For example, this justification can include alignment of data with methodologies and aims by identifying appropriate sources with a standardised set of data characteristics (e.g., when and where the data were collected, spatial and temporal granularity, or characteristics of participants). This discovery process is key in any given research to describe, measure or reflect the condition of interest through the use of data [36]. Data can then be collected through the appropriate sources with reference to the pre-determined set of data characteristics. This process should be well documented and considered in the initial phases of creating, organising, and managing data that is continued through the entire research process [8]. Data collection and discovery is typically guided by a clear DMP, where the data characteristics, expected volume of data, start date, and end date of data collection should be documented when compiling a DMP (see 2.4 Data Management and Appendix 2) [36]. The inclusion of data discovery and collection in a DMP allows for the roles, responsibilities, and expectations of data collectors to be clearly defined. This ensures that all parts of the data discovery and cleaning process have been considered, which can include items such as the expected volume of data, expected data characteristics, start date of collection, and end date of collection [33].

Data are produced in different formats of varying quality [2, 37]. These data can be produced in several formats such as numerical, text, or multimedia. The varying degree of quality can be a result of noise, inconsistencies, and incompleteness in the data which can limit its use [2, 38]. This can also be referred to as uncertainty, where the absence of lack or absence of information, created by noise, inconsistencies and incompleteness, in data creates uncertainty during interpretation (see 2.8 Quality Assurance) [38, 39]. Subsequently, this varying degree of quality and formatting as well as the presence of uncertainty means data is often not readily usable in its original form [2]. Processes are required after collection occurs to ensure that data is in a fit-for-purpose format. This occurs during a phase in the generation of ARD called data cleaning (see 2.7 Data Cleaning) and quality assurance (see 2.8 Quality Assurance).

2.7 Data Cleaning

The generation of consistent data sets and using a standardised format is a requirement for the generation of ARD. During the process of data cleaning, inconsistencies, and errors are removed [40]. These include missing values, typographical errors, mixed formats, and replicated entries. The removal of these errors increases the overall quality of the data. Through this cleaning process, these inconsistencies and errors can be inspected, removed, or replaced in a consistent, replicable, and justifiable manner for each unique data set. [31, 32]. Data cleaning can be conducted by using open-source software that is widely available and distributed [41]. This provides a fast, reliable, robust, and adaptable way to prepare

outputs in a standardised format [42]. Examples of open-source programming languages with libraries widely used for data cleaning are R, Python, and Structured Query Language [43].

2.8 Quality Assurance

Quality assurance is the process of detecting and resolving anomalies in data. This process should be considered when creating DMP and can be conducted simultaneously with data cleaning processes (see Figure 1) [10, 44]. A DMP ensures that the procedures for quality assurance have been established from the onset of a project (see 2.4 Data Management and Appendix 2). Quality assurance is a crucial step to improving and guaranteeing the overall quality of the data. A quality assurance process can ensure that ARD are accurate, complete, consistent, relevant and maintain their longevity. Quality assurance is used to identify any missed or unidentified errors in the data cleaning process before outputs are finalised. After the ARD cleaning has been initially complied into a common format, several supplementary quality assurance processes are useful for ARD to be prepared for deployment: data alignment, uncertainty, and privacy preservation.

Data alignment occurs so that the ARD is converted to align across different characteristics so that datasets can be compared with one another. This process typically occurs after the data is cleaned and compiled into a standardised format. In the case of spatial data, data alignment might entail the conversion of geographic or time series data to a standardised areal or temporal scale to facilitate visualisation and comparison. For example, the Australian Bureau of Statistics (ABS) releases regional boundaries that describe population density and communities known as the Australian Statistical Geographical Standard (ASGS) [45]. These regional boundaries are updated every five years and are represented in a nested hierarchical structure of regions. The ASGS is comprised of several different geographical regions which are summarised in Appendix 4. As these boundaries are updated semi-frequently, data needs to be compiled into a standardised format across a standardised areal and temporal scale. This is achieved through correspondence. ABS also releases machine-readable correspondence files for mathematical conversion between most geographical regions [46].

Uncertainty is often attributed to collection processes (e.g., variance in sampling techniques), variance (e.g., conceptual difference across projects) or multimodality (e.g., data being noisy, inconsistent or incomplete) (see 2.6 Data Discovery and Collection) [2, 38, 39]. This can impact the effectiveness and accuracy of results leading to negative effects on the overall analytics and decision-making processes [37, 38]. If uncertainty is present in data it can also limit its use [2, 38]. Controlling for uncertainty within the quality assurance process will improve the effectiveness and accuracy of results [37, 38]. Uncertainty should be managed and removed, where possible, during the data cleaning and quality assurance process. If uncertainty cannot be removed it should be acknowledged and documented.

When conducting research using sensitive or identifying information, particularly involving human participants, ARD processes should ensure that adequate steps are taken to provide privacy and protection during the research process. Privacy preservation is the process of removing identifying attributes that

are associated with participants [47]. These attributes can include the degree to which an individual is directly identified, typically through an individual's characteristics demonstrated in the data. The risk of re-identification can be minimised through three commonly used techniques: suppression, pseudonymisation, and randomisation. The aim of using these techniques is to achieve an acceptably low risk of re-identification before the data are released to users [48]. The first approach, suppression, removes or suppresses values at a standardised threshold across all records [47, 48, 49]. For example, suppressing all cells with a count less than five is a common threshold for suppression [48, 49]. Pseudonymisation is the implementation of techniques to replace unique identifiers with pseudonyms. These pseudonyms cannot be associated with the data without adequate information (e.g., encryption key) regarding the encryption methods [50]. These can be randomly generated and are used to replace identifiers including names, government issues identifiers or addresses [47]. The final technique, randomisation, adds random values that are not predictable to an existing dataset.

2.9 Metadata

An umbrella term for the information used to describe, manage, and relate to the collection of data is metadata. This information can be generated through automatic or manual processes where a machine-readable output is preferred [51]. Metadata is collected to increase the visibility and reusability of the ARD collected. This visibility and reusability is also considered for Metadata when creating a DMP for a given project. This allows for the metadata creation to be guided by clear and established aims for its creation (see 2.4 Data Management and Appendix 2).

Metadata principles have been devised to guide the promotion and distribution of data on a broader scale [8, 51]. For example, the Australian Data Research Commons (ARDC) has devised principles for creating standardised metadata. These principles are made up of four components: findable, accessible, interoperable, and reusable (FAIR) data principles [51]. These principles created by the Australian Data Research Commons (ARDC) can be used to maximise data sharing capabilities as well as data use and reuse [8, 51]. ARD can be made findable by providing local and international users with a unique or persistent identifier for each dataset [52]. This allows data to be findable by both humans and computers [51]. ARD can be made accessible through the use of a standardised communication protocol even if the data is not open or available for public use. This allows for there to be clarity and transparency around the access, reuse, and authentication processes required for access to ARD [51, 52]. Interoperable ARD is achieved by providing outputs in a standardised approach. This includes the format, and vocabulary in the ARD and metadata so there can be comparisons between other data and metadata. ARD is established as reusable by providing accurate information to users about licencing agreements and detailed provenance information.

2.10 Data Dictionary

Individual data variables are distinct structures and in order to understand their meaning additional information is required [53]. A data dictionary is a central repository of information about variables

present in a dataset and their meaning. This information includes the name, origin of the data, data format, usage, cleaning processes, quality assurance processes, and data management tasks [12, 53]. Information present in data dictionaries should be developed and presented in a standardised format. Data dictionaries should also developed in conjunction with metadata, to ensure the interoperability and reusability of ARD [52, 53]. In comparison to metadata (see 2.9 Metadata), data dictionaries provide information on the individual level of data indicators, providing users with supplementary information about the data processes, data sources, and necessary contextual information about the data and support clarity for further research or analytics using a given collection of ARD. Data dictionaries are also useful for the input of new data or updating existing data sets [53]. Additionally, this allows for findings to be conveyed to a wider audience by increasing interpretability and reducing misperceptions [12, 53].

3. GUIDELINES IN THE CONTEXT OF A DIGITAL ATLAS

In this section, the proposed guidelines are considered in the context of ARD for a digital atlas. These guidelines were implemented for the Australian Child and Youth Wellbeing Atlas (ACYWA) to map data for children and young people's health and wellbeing across Australia [16]. Data relating to children and young people aged 0-24 years were collated for ACYWA. These data were collected from 2006 to 2022 and were aggregated by gender, age, and geographical region for reasons of privacy. Decisions based on the implementation of the ARD guidelines have been noted in the subsequent sections.

3.1 Ethics

In alignment with the proposed guidelines in this paper, ethics approval was gained for ACYWA. This supported the research project creation and design as it abided by the legal and ethical requirements for conducting research in Australia [8, 17]. Furthermore, ACYWA was upheld to high-quality criteria that maintained the security of participants' personal information for the duration of the research process [18]. Appendix 1 outlines key parts of Australia's National Statement, including aspects of the National Statement related to maintaining the security of children and young people. Ethical approval was achieved by the ACYWA development team submitting a Human Research Ethics Application (HREA) through the Human Research Ethics Committee at the Queensland University of Technology. Subsequently, ethics approval was gained for the use of aggregate de-identified data for the purposes of mapping ARD at different geographical levels across Australia using a spatial platform (digital atlas). The HREA submitted for the current project was created and approved in accordance with the National Statement. This specific criteria was chosen as the research involved Australian Universities using data for human participants (Australian children and young people aged 0-24 years old). Consent was not sought for the participants in this project since all data were either publicly available or accessed via request (permission was granted for access and subsequent public release). These data were then de-identified and aggregated at small area levels. These processes were approved by the Human Research Ethics Committee at the Queensland University of Technology.

3.2 Project Documentation

A free cross-institutional cloud-based storage system is currently being used for ACYWA data. Project documentation has ensured that there has been effective version control and modification of records for the duration of the research project [21]. Project documentation was established from the onset of research so that scope, research team, risk, communications, liabilities, processes, decisions, data, and other relevant materials for the project were meticulously recorded. This documentation was recorded and shared with and by the core ACYWA development team. These documents are crucial for rare scenarios where handovers of responsibilities may not be feasible, so that researchers will have step-by-step guides from previous researchers through the processes of the past, ongoing, and upcoming work. Project documentation also provided supplementary information during the handover process. This ensures that researchers, current, returning, and future, have sufficient information to maintain this national data asset efficiently and successfully [8, 21]. At current, ACYWA has been designed as an enduring data asset with no foreseeable end date, where updates to security, ethics and data protection methods will be made as required. As the project is currently ongoing, destruction methods have not been implemented. Considerations need to be made by future researchers on ACYWA about the destruction and deletion of project files once the project has been completed. This approach has been approved in the Human Research Ethics Application.

3.3 Data Governance

Data governance was employed for ACYWA to safeguard the personal information of participants used in this research. In alignment with the proposed guidelines, the Australian Code for the Responsible Conduct of Research was used to create strategies to protect data from misuse and mishandling as well as adopt safe strategies for collecting, sharing and use of the data [25, 27]. The eight principles of honesty, rigour, transparency, fairness, respect, recognition, accountability, and promotion were considered and followed for this research [31]. The approach developed by the ACYWA team was approved in the ethics process.

3.4 Data Management

A Data Management Plan (DMP) was completed in alignment with institutional protocols for ACYWA. This DMP underwent an approval process through ethics and was subsequently approved [22, 23]. A summary is available in Appendix 2. As ACYWA is currently ongoing, the DMP is scheduled to be revisited annually [33].

3.5 Data Storage

As previously stated, ACYWA utilises a free cloud-based storage system that allows cross-institutional access [35]. The cloud-based storage adopted was OneDrive as it allowed for secure access, editing, and organisation of project documentation and data across multiple institutions [54].

OneDrive was flexible, adjusted to advances in technology, and supported large amounts of data ingestion. This storage system has been reliably accessed by researchers on the project across institutions for data access, upload and project documentation [35, 54]. This has ensured reliable access to shared resources, minimised the risk of loss of data, the safety of data and research information, and that participants' information is secure [8].

3.6 Data Discovery and Collection

ACYWA's raw data was collected by first discovering available data resources that aligned with the project participant's characteristics. These characteristics include Australian children and young people aged 0-24 years old. These data were collected from 2006 and 2022 (children and young people born in Australia between 1982 and 2022) across the available spatial and temporal aggregations. This data was also aggregated by age and sex when available. De-identified data was collected for this project where further de-identification techniques were applied to keep the participant's confidentiality (see 2.8 Quality Assurance and 3.8 Quality Assurance). This was achieved by devising a set of potential resources that aligned with the evidence-based framework developed by the Australian Research Alliance for Children and Youth's (ARACY) The Nest. This framework aims to inform health and wellbeing metrics in order for children and young people to thrive and reach their full potential [55, 56]. The data that was identified as aligning with the desired participant characteristics was then assigned one of six key domains with The Nest: healthy; material basics; valued, loved, and safe; learning; participating; and positive sense of identity and culture. This provided a validated approach to creating, organising, and managing from the onset of the project [8].

These data were then collected from Australian data custodians and were obtained through custom requests, public access, and customised data downloads (e.g., TableBuilder [57]). The custom requests through data custodians were often subject to significant wait times (the maximum wait for this project was roughly twelve months, from submission of a request to data received). Data formatting, requirements, and costs were also discussed at length and these decisions were documented throughout the ACYWA research process. The public data was accessed through resources made available online and the ACYWA development team recorded the stages of the collection process to avoid double handling (e.g., downloading data more than once). This process was followed for all customised data sets to create uniform data sets using the same specifications. These data were downloaded from resources such as TableBuilder, and other government and non-government websites (e.g., Australian Institute of Health and Welfare, Australian Bureau of Statistics, and Australian Government) and were downloaded using a standardised set of specifications.

Documentation surrounding the ARD was collected for the duration of the project. This was beneficial as multiple development team members were working on ACYWA and this allowed them to have sufficient information to conduct research efficiently and successfully [8, 21]. The resulting documentation was compiled into standardised formats and was released alongside the data. This supports the guidelines proposed in this paper, in both the initial phases of research and throughout the entire research project [8].

3.7 Data Cleaning

Standardised cleaning processes were used to compile the data for ACYWA into a standardised format. Data was cleaned using open-source software R and RStudio as it provided specialised packages for data cleaning, visualisation, and data analysis [41]. This programming language was chosen as it most closely aligned with researchers' skillset, the ACYWA atlas aims, and methodologies. The data was compiled into a standardised form for the ingest into a custom-built platform. This data format is outlined in Table 2. This required adherence to strict regulation of data standards to properly ingest data into the custom-built atlas. Examples of packages used for data cleaning and management include dplyr, readxlsx, and tidyr [58, 59, 60]. The files for cleaning processes were stored on a cloud-based repository for version control of code and error detection. The resulting data was high-quality ARD that were as free as possible from missing values, typographical errors or replicated entries. Data for ACYWA was compiled in a standardised format that also supported ingest into a custom-built spatial platform. This process improved the quality of the data removed and identified inconsistencies and errors [40]. To date, it is approximated that the current version of ACYWA contains an estimated 430 individual data items which were collected from over 30 different data sources [16]. This data was collected using the ASGS at Statistical Area Two, Statistical Area Three, Statistical Area Four, State and Territory, Australia (National), and Local Government Area (see Appendix 4).

Table 2. Standardised Data Format used for Geographical Data in ACYWA.

Format	Geography	Filter	Filter	Filter	Data
Example	SA3CODE_16	CALENDAR_YEAR	AGE_GROUP	SEX	0

3.8 Quality Assurance

Quality assurance processes were used to check ARD quality and data cleaning processes concurrently for ACYWA. If any inconsistencies were found, the cleaning process was revised, and then the quality assurance was reapplied. This process was continued until the researchers were certain that the data was high-quality, accurate, complete, consistent, and relevant for the long term ingest into the digital atlas [10]. This process was implemented for all data sets in ACYWA. This process was recorded through reproducible documents for reporting and documenting the quality assurance in Rmarkdown [61]. The supplementary quality assurance steps (data alignment, uncertainty, and privacy preservation) were also carried out by the ACYWA development team. Data custodians were also given the opportunity to review the data after ARD processes were applied. Custodian approval was sought on the final ARD output, and any necessary changes were made as per the request of data custodians.

Data for ACYWA was received at several different iterations of the Australian Statistical Geographical Standard (ASGS) boundaries (ASGS2006, ASGS2011, ASGS2016, and ASGS2021) due to the constraints of data availability [62]. As such, correspondence methods, and the standardisation of areal and temporal scales for data comparison, were needed to compile the datasets in a standardised format before the final

release onto the digital atlas. ASGS2016 was chosen as the standardised regional boundary as it was the most reliably available at the time of collection. All other regional boundaries (ASGS2006, ASGS2011, and ASGS2021) were converted to ASGS2016 for the standardised comparison of data [1, 3, 4, 5, 6]. Correspondence was carried out through two different techniques: forward correspondence and backward correspondence. Forward correspondence and backward correspondence are explained in more detail in Appendix 5. Additionally, during these correspondence processes, uncertainty was created. This uncertainty is discussed below.

Uncertainty was present at two phases of the data creation process: uncertainty that was already present at the time of collection and uncertainty that was created during ACYWA's alignment processes. Uncertainty present at the time of collection is often due to noisy, inconsistent, or incomplete data (see 2.8 Quality Assurance) [38]. Certain aspects of ACYWA's cleaning and quality assurance processes (e.g., data alignment and correspondence) resulted in uncertainty, as a lead-on effect of the original noisiness, inconsistencies, and incompleteness in the original data. As such, some of the ARD could not be interpreted with absolute certainty in the results and were subsequently removed. This was achieved by first identifying all types of uncertainty and assigning a standardised numerical value depending on how it would impact the accuracy and effectiveness of results [38, 39]. These numerical values were as follows, zero indicated low uncertainty, one indicated medium uncertainty, and two indicated high uncertainty. High levels of uncertainty were then removed from all datasets. Low and medium uncertainty were deemed not to impact the results enough to be removed. If any uncertainty was present in the data, the level was recorded in the data dictionary for ACYWA on an individual data indicator level.

Data for ACYWA was safeguarded through de-identification and suppression techniques. First, the data was provided and downloaded as aggregate summarises (counts, percentages, rates, etc) for each specified geospatial area. This meant individual information was held by the data custodians and that the ACYWA development team did not have access to participants' personal information at any stage during the research process. Second, suppression techniques were applied by both the data custodians and the ACYWA development team. Suppression was applied across all data sets wherever cell counts of less than five were present [48, 49]. These approaches were documented, included in the DMP, and submitted to ethics for subsequent approval. These techniques ensured that the data risk of misuse and mishandling of ACYWA data was minimised through the protection of participants and their personal information [25, 27, 48, 49].

3.9 Metadata

Data used in ACYWA has been made accessible with increased visibility and reusability by implementing the FAIR principles created by ARDC [8, 51]. The principles adopted were devised to create metadata specifically for geographical and spatial data that aided in the data sharing capabilities for use in a digital atlas [8, 51]. ACYWA metadata was created using both manual and automatic processes [51]. An automatic process through Posit's Rmarkdown was used to produce a summary table with the necessary information to be included in the metadata. This information was then manually transferred to a

standardised metadata document, alongside other information that was manually inputted that could not be sourced from Rmarkdown. A standardised output was then produced for individual datasets. ACYWA's metadata aligned with the four key aspects of ARDC's FAIR principles. This was achieved by utilising a customised version of UWA-PURE that includes the core mandatory elements of additionally the metadata standard ISO19115. For each data set a metadata document was created and as required, was sent for approval and feedback to the appropriate data custodians. These documents were then compiled into a single document which is freely available online and accessible via the ACYWA website. The key aspects of the ACYWA's metadata that aligned with the principles have been outlined in Table 3.

Table 3. Summary of the Key Metadata Elements.

ARDC FAIR Principles	Examples of ACYWA elements that align
Findable	• Title (e.g., name of original source) • Digital Object Identifier (DOI) • Metadata reference
Accessible	• Legal and ethical requirements for data use • Access rights and visibility for data sharing
Interoperable	• Metadata has been compiled in a standardised approach using broadly applicable language and information
Reusable	• Licence • Geographical coverage • Temporal coverage • Fields of research • Socio-economic objectives

3.10 Data Dictionary

The data dictionary for ACYWA captures the necessary information that provides insight to users into the creation, usage, and format of the data [12]. The data dictionary for each variable was created manually by the researchers on ACYWA and has undergone a quality assurance process. These items will be provided to two different end users: users of the data and members of the ACYWA development team. These items have been summarised in Table 4, where the data dictionary items for users are planned for future release. The information provided to researchers is stored securely, to maintain the integrity of the project, minimise reputational risk and does not risk the identities of the participants in the project [8]. A similar process was followed in the creation of metadata, where Posit's Rmarkdown created a summary output with the key information and the ACYWA development team manually filled in the entries for the data dictionary. The entries not covered by the Rmarkdown output are then entered manually by the researchers. These entries then undergo a quality assurance process. This allows for the data to have maximised visibility, reach a wider audience and avoid any misperceptions when the data is being interpreted [12, 53]. The key items that have been included in the data dictionary for ACYWA have been summarised in Table 4.

Table 4. Key Elements in the ACYWA Data Dictionary for Users and Researchers.

Data Dictionary Users	Examples of Data Dictionary Elements in ACYWA
Users of Data (published)	• Data variable name • Data variable definition • Variable type (count percentage, etc) • Data source • Application of temporal correspondence • Presence of uncertainty

Researcher-Only (not published)	• Links for
	  ◦ Cleaning code
	  ◦ Data files
	  ◦ Project documentation

4. CONCLUSION

The current paper aimed to contextualise current approaches into a new set of guidelines that can be used to generate ARD. A total of ten steps were developed that can be used to produce ARD: Ethics, Project Documentation, Data Governance, Data Management, Data Storage, Data Discovery and Collection, Data Cleaning, Quality Assurance, Metadata and Data Dictionary. These guidelines have effectively combined previous literature to provide a comprehensive approach to ARD creation. Together, these principles form the first focused framework for producing ARD. The guidelines were also successfully contextualised to produce ARD for the digital platform ACYWA.

ARD processes are crucial for facilitating the comparison, inspection, and display of different geospatial and temporal data [15]. The paper showcased an example of effective the implementation of this guideline in the context of a digital atlas (ACYWA). Thus, achieving the aim of compiling current literature in a comprehensive guide for the creation of ARD. The resulting output has created a data asset that is free and publicly available to a range of users, and stakeholders, for general public use which can be used to inform children and young people's health and wellbeing metrics in Australia [16]. This digital atlas provides users with high-quality ARD on multiple different spatial and temporal levels across Australia that were not previously available. The guidelines also provided the development team with meaningful insights into the overall creation of the project. This includes the generation of supplementary materials such as project documentation, a data management plan, metadata, and a data dictionary. Overall, the culmination of the output of high-quality ARD and associated documentation by adhering to stricter generation processes promotes more trust in the data and the resulting product (i.e. a digital atlas).

The final product for the ACYWA is online, free, and publicly available for users, but establishment, collection, and cleaning had associated costs. This should be factored into future projects where data sources will need to be chosen subject to the availability of funding. For example, should data be collected from only free and public resources only or is there availability for paid resources as well. Data custodians for ACYWA also had the final decision on the inclusion or exclusion of data in the digital platform. Data

could meet all ARD requirements, but might not meet a data custodian's individual, internal approval processes. As such, a few custodians did not allow the ARD to be shared further than the ACYWA development team and could not be shared publicly. This decision could only be made after all ARD processes had been applied and the final product could be reviewed by data custodians. Albeit, this was a rare occurrence, where an estimated 430 individual data items were included in ACYWA. Nonetheless, the exclusion of ARD outputs due to custodian review processes should still be considered as a possibility for future research. Additionally, the standardised data format was created with the specifications required for geographical data and ingested into a digital atlas. This standardised format can be contextualised to other digital atlases and geographical data that require the same specifications. The use of this standardised approach allows the data to align with FAIR data principles, especially if the format is adopted by other projects. Future work could explore the use of this standardised approach to different types of data.

Strategies could be used to evaluate the compliance of researchers when implementing these guidelines in future studies. This research could also involve exploring the efficacy of the different steps and whether all steps are needed to produce the same high-quality ARD output. For example, if either a data management plan or quality assurance is excluded will the same output be produced and is the output of high-quality. The reordering of steps could also be explored. Different research fields may require a varied order of steps. This can be explored in future case studies if implementation of the steps in a different order yields successful production of ARD. Future work could also be undertaken to modify the ARD generation steps to focus on different core principles depending on the field the research is being conducted. For example, satellite data will require different generation steps than data using human participants. More specifically, satellite data may not require ethics, but would likely place a larger focus on data cleaning, quality assurance, and metadata processes where data needs to be aligned temporally, and be made findable, accessible, interoperable, and reusable [1, 3, 14]. The current guidelines provide a framework and a general guide needed for the generation of ARD.

These guidelines can be used not only to improve the overall quality of ARD through the stages such as discovery and collection, cleaning, as well as quality assurance; but also to ensure that the overall project is conducted at a high standard by considering steps such as ethics, project documentation, and data storage. This ensures that the data that was initially limited in its use is made accessible to its desired users [2]. As a result, this allows for the expansion of use through the trust in high-quality ARD outputs by users (e.g., the general public, stakeholders, users, researchers, etc) [16]. These uses can include new policy decisions, analytics, or general research. This also promotes further discussion in the scientific community and broader fields of research for the development of ARD through comprehensive guidelines.

The new guidelines build on and synthesise existing literature and have been showcased using a digital atlas. These guidelines allow for further consideration of processes underpinning the creation of ARD and steps, allowing avenues for innovation, novel, and future research. Overall, these guidelines provide a systematic approach to ARD creation that incorporates insights from existing literature and leveraging digital atlas technology that was not previously available.

AUTHOR CONTRIBUTIONS

Harriette Phillips: conceptualization, methodology, writing original draft, investigation, validation, project administration, software, data curation, writing review, and editing

Aiden Price: writing review and editing, data curation, and supervision

Owen Forbes: writing review and editing, and data curation

Claire Boulange: writing review and editing, and data curation

Kerrie Mengersen: writing review and editing, and supervision

Marketa Reeves: writing review and editing, project administration, supervision, and data curation

Rebecca Glauert: writing review and editing, and project administration

ACKNOWLEDGEMENTS

This work was supported by the Queensland University of Technology's Centre for Data Science and Research Training Program Stipend Scholarship. The Australian Child and Youth and Wellbeing Atlas also supported the work.

DATA AVAILABILITY STATEMENT

The ARD that has been produced is free and publicly available through the ACYWA platform. The digital platform can be found here: https://australianchildatlas.com/.

REFERENCES

[1] P. Potapov et al., "Landsat analysis ready data for global land cover and land cover change mapping," *Remote Sens.*, vol. 23, no. 4, pp. 497–514, 2021.

[2] D. Arribas-Bel et al., "Open data products: A framework for creating valuable analysis ready data," *J. Geogr. Syst.*, vol. 23, no. 4, 2021, doi: 10.1007/s10109-021-00363-5.

[3] A. Baraldi et al., "Innovative analysis ready data (ARD) product and process requirements, software system design, algorithms and implementation at the midstream as necessary-but-not-sufficient precondition of the downstream in a new notion of space economy 4.0 - part 1: Problem background in artificial general intelligence (AGI)," *Big Earth Data*, vol. 7, no. 3, 2023, doi: 10.1080/20964471.2021.2017549.

[4] A. Baraldi et al., "Innovative analysis ready data (ARD) product and process requirements, software system design, algorithms and implementation at the midstream as necessary-but-not-sufficient precondition of the downstream in a new notion of Space Economy 4.0 - Part 2: Software developments," *Big Earth Data*, vol. 7, no. 3, 2023, doi: 10.1080/20964471.2021.2017582.

[5] D. Huth, "A pattern catalog for GDPR compliant data protection," 2017. Accessed: Jul. 23, 2023. [Online]. Available: https://ceur-ws.org/Vol-2027/paper4.pdf.

[6] P. Baumann et al., "Datacubes: Towards space/time analysis-ready data," in *Service-Oriented Mapping, Lect. Notes Geoinf. Cartogr.*, 2019, pp. 269–299.
[7] G. C. Banks et al., "A review of best practice recommendations for text analysis in R (and a user-friendly app)," *J. Bus. Psychol.*, vol. 33, no. 4, pp. 445–459, 2018.
[8] L. Corti et al., *Managing and Sharing Research Data: A Guide to Good Practice*. SAGE, 2019.
[9] G. Giuliani et al., "Observations data cube: Lessons learned from Swiss data cube (SDC) on generating analysis ready data (ARD)," *Big Earth Data*, vol. 1, nos. 1–2, pp. 100–117, 2017.
[10] P. Derakhshan et al., "The impact of data quality assurance and control solutions on the completeness, accuracy, and consistency of data in a national spinal cord injury registry of Iran (NSCIR-IR)," *Spinal Cord Ser. Cases*, vol. 7, no. 1, 2021.
[11] S. Stieglitz et al., "Social media analytics: Challenges in topic discovery, data collection, and data preparation," *Int. J. Inf. Manage.*, vol. 39, pp. 156–168, 2018.
[12] E. M. Buchanan et al., "Getting started creating data dictionaries: How to create a shareable data set," *Adv. Methods Pract. Psychol. Sci.*, vol. 4, no. 1, 2021.
[13] A. Jacobsen et al., "FAIR principles: Interpretations and implementation considerations," *Data Intell.*, vol. 2, nos. 1–2, pp. 10–29, 2020.
[14] P. Harrison and M. Thankappan, "CEOS-ARD product self-assessment user guide." Accessed: Feb. 26, 2024. [Online]. Available: https://ceos.org/ard/files/User%20Guide/CEOS_ARD%20User%20Guide%20v1_3_Final_06152023.pdf.
[15] F. Ormeling and C. Elzakker, "Digital atlas user requirements and use scenarios." Accessed: Nov. 23, 2023. [Online]. Available: https://icaci.org/files/documents/ICC_proceedings/ICC2009/html/nonref/14_7.pdf.
[16] R. Glauert et al., "Australian child and youth wellbeing atlas." Accessed: Oct. 25, 2023. [Online]. Available: https://australianchildatlas.com/.
[17] K. Lõuk, "Relevance and challenges of ethics committees," in *Ethics and Responsible Research and Innovation in Practice, Lect. Notes Comput. Sci.*, 2023, pp. 125–136.
[18] G. Scherzinger and M. Bobbert, "Evaluation of research ethics committees: Criteria for the ethical quality of review process," *Account. Res.*, vol. 24, no. 3, pp. 152–176, 2017.
[19] United States Department of Health and Human Services, "The Belmont report: Read the Belmont report." Accessed: Dec. 14, 2023. [Online]. Available: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html.
[20] Australian Government National Health and Medical Research Council, "National statement on ethical conduct in human research." Accessed: Oct. 23, 2023. [Online]. Available: https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2023#mm-0.
[21] M. Das, "A blockchain-based integrated document management framework for construction applications," *Autom. Constr.*, vol. 133, 2022.
[22] R. Abraham, "Data governance: A conceptual framework, structured review, and research agenda," *Int. J. Inf. Manage.*, vol. 49, pp. 424–438, 2019.
[23] D. Paparova et al., "Data governance spaces: The case of a national digital service for personal health data," *Inf. Organ.*, vol. 33, no. 1, 2023.
[24] S. Meier et al., "A process model for systematically setting up the data basis for data-driven projects in manufacturing," *J. Manuf. Syst.*, vol. 71, pp. 1–19, 2023.
[25] M. Godinho and I. Adjerid, "Consumer consent and firm targeting after GDPR: The case of the large telecom provider," *Manage. Sci.*, vol. 68, no. 5, pp. 3330–3378, 2022.
[26] European Union, "General data protection regulation (GDPR)." Accessed: Jul. 28, 2023. [Online]. Available: https://gdpr-info.eu/.

[27] S. Sirir, J. R. C. Nurse, and H. Webb, "Are we there yet? Understanding the challenges faced in complying with the General Data Protection Regulation (GDPR)," 2018, pp. 88–95.
[28] T. G. Almeida, S. M. Mira da Silva, and R. Pereira, "The critical success factors of GDPR implementation: A systematic literature review," *Digit. Policy Regul. Gov.*, vol. 21, no. 4, pp. 402–418, 2019.
[29] Australian Research Council, "Australian code for the responsible conduct of research 2018." Accessed: Aug. 4, 2023. [Online]. Available: https://www.arc.gov.au/about-arc/program-policies/research-integrity/australian-code-responsible-conduct-research-2018.
[30] Ong et al., "Responsible research practices could be more strongly endorsed by Australian university codes of research codes," *Res. Integr. Peer Rev.*, vol. 8, art. no. 5, 2023.
[31] S. Bera et al., "Designing attribute-based verifiable data storage and retrievable scheme in cloud computing environment," *J. Inf. Secur. Appl.*, vol. 75, art. no. 103482, 2023.
[32] W. K. Michener, "Ten simple rules for creating a good data management plan," *PLoS Comput. Biol.*, vol. 11, no. 10, art. no. e1004525, 2015.
[33] Queensland University of Technology, "QUT-Data management plan." Accessed: Sep. 5, 2023. [Online]. Available: https://dmp.qut.edu.au/faq.
[34] Australian Research Council, "Management of data and information in research: A guide supporting the Australian code for the responsible conduct in research." Accessed: Feb. 24, 2023. [Online]. Available: https://www.nhmrc.gov.au/sites/default/files/documents/attachments/Management-of-Data-and-Information-in-Research.pdf.
[35] J. M. Cavanillas, E. Curry, and W. Wahlster, Eds., *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*. Springer International Publishing, 2016.
[36] L. M. Sylvia and M. F. Terhaar, Eds., *Clinical Analytics and Data Management for the DNP*. Springer Publishing, 2023.
[37] J. D. Hand, "Statistical challenges of administrative and transaction data," *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 181, no. 3, pp. 555–605, 2018.
[38] D. W. Rowe, "Understanding uncertainty," *Risk Anal.*, vol. 14, no. 5, pp. 743–750, 1994.
[39] H. R. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: Survey, opportunities, and challenges," *J. Big Data*, vol. 6, no. 1, art. no. 44, 2019.
[40] X. Chu et al., "Data cleaning: Overview and emerging challenges," in *Proc. 2016 Int. Conf. Manage. Data*, 2016, pp. 2201–2206.
[41] F. Kronthaler and S. Zöllner, "R and RStudio," in *Data Analysis with RStudio*. Berlin, Germany: Springer Spektrum, 2021, pp. 1–12.
[42] E. Kaya et al., "Spatial data analysis with R programming for environment," *Hum. Ecol. Risk Assess.*, vol. 25, no. 6, pp. 1521–1530, 2019.
[43] A. Badia, *SQL for Data Science: Data Cleaning, Wrangling and Analytics with Relational Databases*. Springer Nature, 2020.
[44] J. J. Gassman, "Data quality assurance, monitoring, and reporting," *Control. Clin. Trials*, vol. 16, no. 2, pp. 104–136, 1995.
[45] Australian Bureau of Statistics, "Australian Statistical Geography Standard (ASGS)." Accessed: Aug. 23, 2023. [Online]. Available: https://www.abs.gov.au/statistics/statistical-geography/australian-statistical-geography-standard-asgs.
[46] Australian Bureau of Statistics, "Correspondences." Accessed: Nov. 15, 2023. [Online]. Available: https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/access-and-downloads/correspondences.

[47] Y. Y. Heung, "An overview of de-identification techniques and their standardization directions." Accessed: Dec. 15, 2023. [Online]. Available: https://www.researchgate.net/publication/342606728_An_Overview_of_De-Identification_Techniques_and_Their_Standardization_Directions.

[48] K. Wilkinson et al., "Less than five is less than ideal: Replacing the 'less than 5 cell size' rule with a risk-based data disclosure protocol in a public health setting," *Can. J. Public Health*, vol. 111, pp. 761–765, 2020.

[49] Information and Privacy Commissioner of Ontario, "De-identification guidelines for structured data." Accessed: Dec. 15, 2023. [Online]. Available: https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf.

[50] T. Neubauer and J. Heurix, "A methodology for the pseudonymization of medication data," *Int. J. Med. Inform.*, vol. 80, no. 3, pp. 190–204, 2011.

[51] Australian Data Research Commons, "Metadata." Accessed: Dec. 14, 2023. [Online]. Available: https://ardc.edu.au/resource/metadata/.

[52] Australian Data Research Commons, "FAIR data." Accessed: Sep. 13, 2023. [Online]. Available: https://ardc.edu.au/resource/fair-data/.

[53] M. S. Rashid et al., "The semantic data dictionary: An approach for describing and annotating data," *Data Intell.*, vol. 2, no. 4, pp. 443–486, 2020.

[54] Microsoft, "Personal cloud storage-Microsoft OneDrive." Accessed: Feb. 27, 2023. [Online]. Available: https://www.microsoft.com/en-au/microsoft-365/onedrive/online-cloud-storage.

[55] Australian Research Alliance for Children and Youth, "In ARACY-Australian research alliance for children and youth." Accessed: Jul. 12, 2023. [Online]. Available: https://www.aracy.org.au/the-nest-in-action/in-aracy.

[56] Australian Research Alliance for Children and Youth, "The Nest in action-Australian research alliance for children and youth (ARACY)." Accessed: Dec. 6, 2023. [Online]. Available: https://www.aracy.org.au/the-nest-in-action.

[57] Australian Bureau of Statistics, "TableBuilder." Accessed: Oct. 4, 2023. [Online]. Available: https://www.abs.gov.au/websitedbs/censushome.nsf/home/tablebuilder.

[58] H. Wickham et al., "dplyr: A grammar of data manipulation." Accessed: Nov. 17, 2023. [Online]. Available: https://cran.r-project.org/web/packages/dplyr/index.html.

[59] H. Wickham et al., "readxl: Read Excel files." Accessed: Feb. 14, 2024. [Online]. Available: https://cran.r-project.org/web/packages/readxl/index.html.

[60] H. Wickham et al., "tidyr: Tidy messy data." Accessed: Nov. 15, 2023. [Online]. Available: https://cran.r-project.org/web/packages/tidyr/index.html.

[61] Rmarkdown, "Rmarkdown from RStudio." Accessed: Dec. 7, 2023. [Online]. Available: https://rmarkdown.rstudio.com/.

[62] Australian Bureau of Statistics, "Australian statistical geography standard (ASGS) edition 3." Accessed: Nov. 15, 2023. [Online]. Available: https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/latest-release.

AUTHOR BIOGRAPHY

Harriette Phillips is a master's student currently completing a Master of Philosophy at the Queensland University of Technology. The aim is to implement statistical methodologies for dimension reduction of aggregate data. This work will provide vulnerability indices for children and young people's health and wellbeing in Australia. She has previously worked as a Data Collection Project Manager on the Australian Child and Youth Wellbeing Atlas, and obtained a Bachelor of Behavioural Science. Harriette has experience in data science, data collection, project management, and statistical programming.
E-mail: harriette.phillips@hdr.qut.edu.au; ORCID: 0000-0003-3001-8941

Dr Aiden Price is currently a Research Associate at the Queensland University of Technology and Project Manager of the AusEnHealth Initiative. Aiden's work largely focuses on Environmental health and systems. He is also a Post-Doctoral Fellow for the Healthy Environments and Lives (HEAL) Network, Environmental Health Domain Specialist at the Australian Research Infrastructure Network (AURIN), and a ECR Co-lead Government Systems Domain Centre Chief Investigator.
E-mail: a11.price@qut.edu.au; ORCID: 0000-0001-5904-4110

Dr Owen Forbes has recently completed his PhD in novel statistical methods for analysing complex bio signal datasets. This was applied to electroencephalography (EEG) analysis and relating data-driven phenotypes to mental health and cognitive function. Owen has over seven years' experience in the areas of public health, epidemiology, mental health, and applied statistics. He has significant experience in applied data science, database engineering, and statistical programming, and a broad set of competencies in qualitative and quantitative research.
E-mail: owen.forbes@connect.qut.edu.au; ORCID: 0000-0002-7634-8279

Dr Claire Boulange currently works at the University of Western Australia as a data analyst. She has previously worked at KPMK Melbourne and GTA Melbourne as a Senior Consultant.
E-mail: claire.boulange@uwa.edu.au; ORCID: 0000-0002-0610-7496

Distinguished Professor Kerrie Mengersen is an Australian statistician. Her work focuses on developing new statistical and computation methods that aid in solving complex problems in environmental, genetics, health, medicine, and industry research.
E-Mail: k.mengersen@qut.edu.au; ORCID: 0000-0001-8625-9168

Marketa Reeves is currently the project manager for the Australian Child and Youth Wellbeing Atlas. She has work previously worked at the Commissioner for Children and Young People WA as a Manager Wellbeing Data and Research. Marketa has extensive knowledge on Childrens Health and Wellbeing and their data.
E-Mail: marketa.reeves@uwa.edu.au

Associate Professor Rebecca Glauert is currently a Principal Research Fellow at the University of Western Australian. Rebecca is also the Scientific Director of the Raine study and the Telethon Kid's Institute. Rebecca has extensive knowledge on children's health and wellbeing as well as their data. Her expertise is in linked administrative data across government sectors and is passionate about good data practices.
E-Mail: rebecca.glauert@uwa.edu.au; ORCID: 0000-0002-6272-8932

APPENDIX

APPENDIX 1

Overview of the General Consideration and ACYWA Considerations for the National Statement of Ethical Conduct of Human Research [20].

Considerations of National Statement Overview	
Merit and Integrity in Research	• Research outputs can be justified by research benefits. • Research conducted, developed, or designed using an appropriate methodology that aligns with the aims of the given research proposal. • Research is based on a thorough study of current literature and previous studies. This includes research in quick response to unforeseen situations and possible novel research. • The research aims of the research, how it is conducted, and the results should ensure that conducted in such a manner that respects the participants of the research. • Facilities where the research is conducted are appropriate for the given topic. • The qualifications or competence of researchers is sufficient to supervise or conduct research. • When conducting research, the researchers should be researching for knowledge and understanding of the topic, follow the principles of research conduct, disseminate results whether favourable or unfavourable. • Once a peer review of the research has been completed, research merit is no longer dictated by those completing the ethical review.
Justice	• The results of the research have been accurately and fairly described. • The scope, research objectives, and inclusion and exclusion criteria have been considered. • Participants are recruited fairly, are not burdened, or exploited in the research process. • Research outputs are disseminated in a timely and clear manner, and any benefits are made available to participants
Beneficence	• Any risk of harm or discomfort to participants must be outweighed by the benefits of the research. The research must be likely to show benefit to either the participants, their community, or both. • Researchers should design research to minimise risk and discomfort, as well as make any potential risks or benefits known to participants, whilst keeping the welfare of participants in mind.
Respect	• The confidentiality, cultural sensitivities, and privacy of participants and their communities be respected by researchers and institutions when conducting research. If necessary, any specific agreements should between participants or communities should be fulfilled during research. • All research should take into consideration the autonomy of humans, and their ability to make decisions on their own. If participants have a diminished capacity for decision-making, researchers should empower participants where possible to make decisions and provide them with the necessary protections.

Continued.

Considerations of National Statement Overview	
Risk and Benefits	• Any risks, harm, discomfort, or inconvenience (risk) should be identified, assessed for likeliness to occur, identified who the risk is likely to affect, how the risk will be minimised, potential benefits, and who the benefits are likely to affect. • The risks and severity of risk should be reviewed based on available evidence. • Research can only be deemed "low risk" if the risk identified in the research is "discomfort." Research is identified as "negligible risk" if the risk of harm or discomfort is not foreseeable.
General Requirements for Consent	• Research participants should participate in research voluntarily. Participants should also be provided with sufficient information about the research project (e.g., methods, risks, outputs, demands, or potential benefits) and the implications of the research to make an informed decision about their involvement. • Participants must be provided with information that is presented in a manner that is suitable for all. • Consent needs to be gained for inclusion in research projects where participants gain an understanding of their involvement in the research process. Consent can be gained in the following ways: ◦ Verbally ◦ Written ◦ By other means (e.g., Survey)
Considerations Specific to ACYWA	
Children and Young People	• Consent should be gauged for the inclusion of children and young people participation in research. • Researchers should describe how children and young people will describe the research and how it will be relayed to them at a level they can comprehend. • If consent cannot be obtained, inclusion in the research must be justified by advancement in knowledge or children and young people's involvement are indispensable to the research. Research should be conducted in the best interests of children and young people. • Research should be conducted in a manner that provides emotional and psychological security, safety, and wellbeing for children and young people.

APPENDIX 2

Data Management Plan Topics for Consideration

Key Processes	Overview
Data Discovery	• What is the expected type of data that is going to be collected for the project?
Data Description	• Describe of the data being collected for the data (i.e., what are the specifications for the data, who are the participants, etc).

Continued.

Key Processes	Overview
Data Collection	• Who will be in charge of collecting the data? • Where will the data collection occur? • Are there any equipment requirements for the collection of the data? • What are the expected characteristics of the data? • What is the start date of the data collection? • What is the end date of the data collection? • What is the expected volume of data?
Data Organisation	• How will the data be stored once it is received? • How will this be catalogued and documented? • Are the quality control measures in place?
Data Preservation Procedures	• Are the data preservation procedures beyond in place the scope of the project?
Data Storage	• Where will the data be stored? • Does this storage account for data loss (see data loss procedures)?
Data Volume	• What is the expected volume of data for the research project? • Can the chosen data storage handle the expected volume of data?
Data Loss Procedures	• Are there multiple storage options to account for and prevent potential data loss? • If data loss occurs, are there appropriate measures in place to negate the loss of data?
Data Privacy and Confidentiality	• Have participants' privacy and confidentiality been considered? • Is there no/ low risk to participants? • How have risks been minimised?
Data Ownership	• Which organisation or individual has ownership of the data collected?
Quality Assurance	• What quality procedures are in place?
Documentation	• Where are decisions made regarding the data collection processes stored? • Who is documenting decisions? • Are decisions about the data being documented? • What decisions should be documented?
Dissemination	• Is the data being disseminated? • If so, how is it being disseminated? • Have legal, licencing, and ethical requirements been considered?
Metadata	• Have metadata principles been completed to make the data more accessible?

APPENDIX 3

Outlines of the Key Eight Principles in the Australian Code for the Responsible Conduct of Research

Principle	Overview
Honesty	• Truthful and accurate information presented when preposing, conducting, and reporting research
Rigour	• Robust methodology used, attention to detail to avoid biases which are otherwise acknowledged
Transparency	• Research methodology, data, and findings are communicated and shared openly • Research is communicated responsibly and accurately • Conflicts of interests are managed or disclosed
Fairness	• Researchers and others involved in the research should be treated with respect and fairly • Works of others should be cited and referenced • Appropriate credit should be given to contributors to the research
Respect	• Conducting research for minority and vulnerable people should be given appropriate considerations, ensuring that participants and communities affects by research are care for and respected
Recognition	• Conducting research with Aboriginal and Torres Strait Islander Peoples should be engaged in the research which affects or is significance to them ◦ Aboriginal and Torres Strait Islander Peoples connection to land, diversity, heritage, knowledge, and cultural property is respected, valued, and recognised ◦ Prior to undertaking research, Aboriginal and Torres Strait Islander people are engaged with so that they may freely make decisions about their involvement • The outcome of the research is reported to the Aboriginal and Torres Strait Islander people who engaged with the research
Accountability	• Guidelines, policies, and legislations should be abided by • Good stewardship used when conducting research with public resources • Consequences and outcomes considered prior to communication
Promotion	• Responsible conduct of research is fostered and promoted

APPENDIX 4

Overview of the Structure of the Regions in the Australian Statistical Geographical Standard (ASGS) [62].

Australian Statistical Geographical Standard	
Mesh Blocks	• Mesh Blocks are the smallest defined geographical area and are used to define the larger regions in the ASGS • Mesh Blocks typically contain 30-60 dwellings
Statistical Area Level 1 (SA1)	• Used for the description of census data • SA1 typically contains 200-800 people
Statistical Area Level 2 (SA2)	• SA2s are used to represent communities, and how they interact with one another socially and economically • SA2s typically contain 3, 000 to 25, 000 people
Statistical Area Level 3 (SA3)	• SA3s are designed for the outputs of regional data • SA3s typically contain 30, 000 to 130, 000 people
Statistical Area Level 4 (SA4)	• SA4s function as defining key areas, such as capital cities, representing regional data and the labour markets • They typically contain over 100, 000 people
State and Territory (STE)	• The legally defined, cartographic representation of the state and territory boundaries
Australia	• A representation of the entirety of Australia
Indigenous Structure	• Used for the analysis and publication of statistics for the Aboriginal and Torres Strait Islander population in Australia
Remoteness Structure	• Five classes of remoteness (major cities, inner regions, outer regional, remote, and very remote) based on relative access to services
Non-ABS Structures	• Structures that are not defined by the ABS. ◦ Local Government Areas (LGA) ◦ Destination Zones ◦ Australian Drainage Regions ◦ Tourism Regions ◦ Postal Areas ◦ State Electoral Divisions ◦ Suburbs and Localities (formerly State Suburbs)

APPENDIX 5

Uncertainty Processes for ACYWA

Backward correspondence is the process of estimating count data for ASGS2016 regions using count data from ASGS2021 regions. Forward correspondence is the process of estimating count data for ASGS2016 using count data from ASGS2006 and ASGS2011 regions. That is, this approach achieves standardisation of the data backward through the temporal scales across the same areal levels. Consider a region, A, which has been split into two regions, B and C, through known ratios, a and b, respectively (see Figure 2a). Then forward correspondence is calculated as follows for regions B and C.

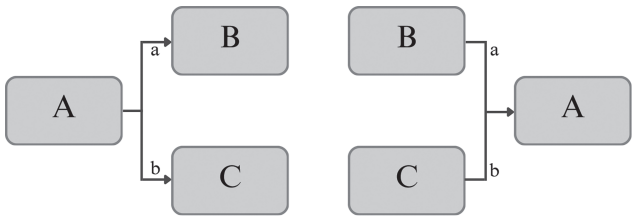


Figure 2. Visualisation of the Forward Correspondence Process.

$$B = a \times A$$
$$C = b \times A$$

with $a + b = 1$ such that region A is distributed entirely to ASGS2021 regions B and C. To derive ASGS2016 region counts from ASGS2021 regions, i.e., perform backward correspondence, consider the original distribution of counts in reverse. In Figure 2b, for example,

$$A = a \times B + b \times C,$$

where all of regions B and C combine to form region A. i.e., $a = b = 1$. Consider now the more complex example in Figure 3a, where two ASGS2016 regions, A and B, are distributed to three ASGS2021 regions, C, D and E. Region distributions between ASGS years, or forward correspondence, is given by:

$$C = a \times A,$$
$$D = b \times A + c \times B,$$
$$E = d \times B.$$

Backward correspondence for this scenario is visualised in Figure 3b, with counts for regions A and B calculated as follows.

$$A = a \times C + b \times D,$$
$$B = c \times D + d \times E.$$

In this process, regions C and E distribute uniquely to regions A and B, respectively, and so $a = d = 1$. However, region D in ASGS2021 was formed by two ASGS2016 regions, A and B, and so must redistribute its count data to both regions, with $b + c = 1$. In this case, however, we do not know either b or c. That is count estimates distributions are not known throughout region D for the purpose of redistribution to regions A and B. To provide correspondence measures when this occurs, the above calculations are considered using the forward correspondence ratios in Figure 3a as follows.

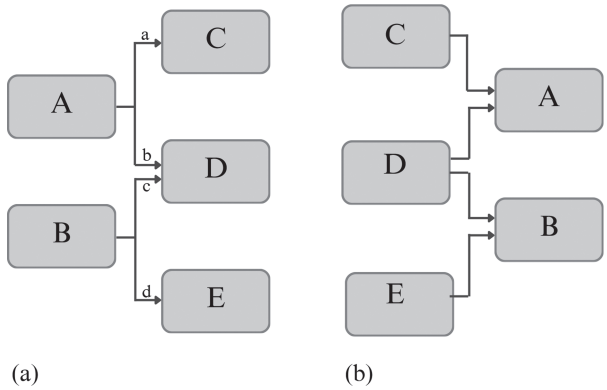


Figure 3. Visualisation of the Backward Correspondence Process.

$$A = \begin{cases} C, & b < 0.1 \\ Suppressed, & b > 0.1 \end{cases}$$
$$B = \begin{cases} E, & c < 0.1 \\ Suppressed, & c > 0.1 \end{cases}$$

In this way, insignificant contributions are discarded (less than 10%) from any backward correspondence calculation. As a result, the backward correspondence process described above can occasionally introduce uncertainty into data assets. This uncertainty is discussed in 2.8 Quality Assurance and 3 Guidelines in the Context of a Digital Atlas.