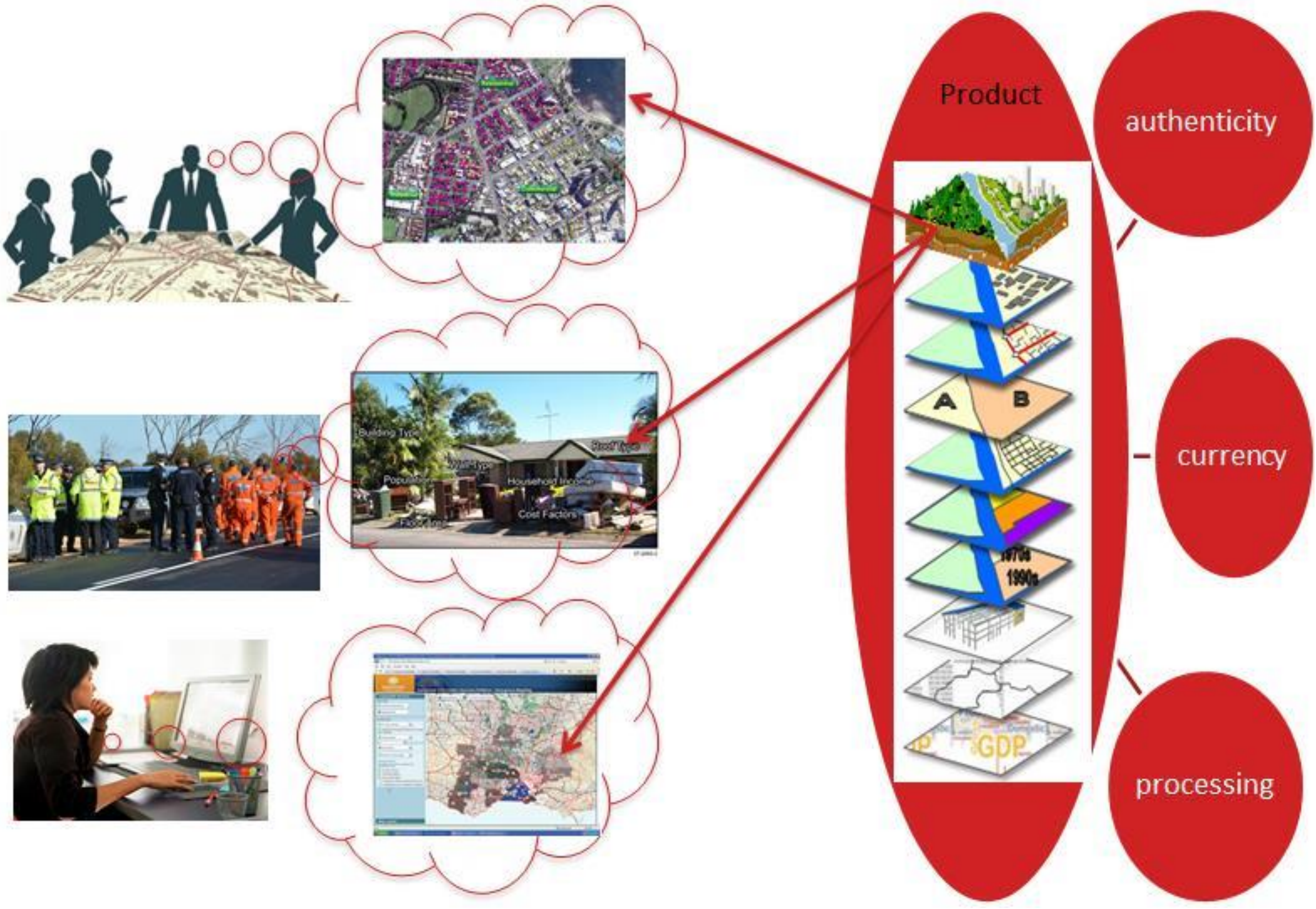
bnhcrc.com.au

CONSTRUCTING A DATA RELIABILITY FRAMEWORK FOR THE NATURAL HAZARD EXPOSURE INFORMATION SYSTEM

Yogi Vidyattama
University of Canberra
Bushfire and Natural Hazards CRC



UNIVERSITY OF CANBERRA
Australian Government
Geoscience Australia

Version	Release history	Date
1.0	Initial release of document	5/12/2018

Disclaimer:
The University of Canberra, Geoscience Australia and the Bushfire and Natural Hazards CRC advise that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, The University of Canberra, Geoscience Australia and the Bushfire and Natural Hazards CRC (including its employees and consultants) exclude all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Publisher:
Bushfire and Natural Hazards CRC

December 2018

Citation: Vidyattama, Y. (2018) Constructing a data reliability framework for the Natural Hazard Exposure Information System. Melbourne: Bushfire and Natural Hazards CRC

TABLE OF CONTENTS

1 ABSTRACT

1.1 CONSTRUCTING A DATA RELIABILITY FRAMEWORK FOR THE NATURAL HAZARD EXPOSURE INFORMATION SYSTEM

Yogi Vidyattama, *IGPA, University of Canberra, ACT*

Consistent and reliable exposure information is crucial for disaster mitigation and evidence-based decision-making for bushfire and other natural hazard risk. There are a few capabilities in Australia to provide nationally consistent exposure information, such as Geoscience Australia's National Exposure Information System (NEXIS), but they are not comprehensive enough to address the entire spectrum of disaster risk reduction. The existing capabilities were developed to provide information for their known clients and users. To manage disasters efficiently, there is the need for a nationally consistent framework that deals with the collection, collation and dissemination of exposure information for researchers and decision-makers.

The aim of this research project is to develop a framework that could provide a reliability assessment for exposure information that is available to various types of users. This assessment framework will play an important role in the disbursement of data and knowledge through a website and web portal because generally, the users take the information as it stands and assume the information will be appropriate for their usage.

In particular, the data and knowledge disbursed are intended to moderate the socio-economic impact from natural hazards and lifeline utility failures. In doing so, improved risk assessment tools underpin planning, preparedness, response and recovery in disaster management that enable well-informed decision-making. Exposure information systems obtain data from many sources with varying resolutions, quality, standards, aggregations, dis-aggregations, statistical approximations and estimations. The reliability assessment framework will help the users, providers and managers of the exposure data and information to communicate the variation in reliability or quality, and ensure they are used appropriately to assess the risk.

Building on the International Standards Organisation's criteria for data quality as well as a standardised data provenance framework, we propose a data

reliability framework for exposure information systems. One of the features suggested for this framework is for exposure information systems to start with classification systems for various reliability or quality criteria based on the provenance, spatial accuracy, currency and precision of the data. The framework then sets different thresholds of these quality criteria for different types of users. However, this is only an initial threshold within the system. It is recommended that in order for the framework to work, it needs to be open to user input to re-evaluate the standard being put in place. This is owing to the huge variation of users and hence their requirements in terms of data quality. Therefore, it is important to recalibrate the data element criteria, the assessment threshold and the grouping of the type of users based on their inputs.

2 END-USER STATEMENT

Mark Edwards, *Geoscience Australia, Canberra, ACT*

It can be challenging to translate detailed technical data into meaningful information for decision makers, particularly in response to disasters. Exposure information systems obtain data from many sources with varying resolutions, quality, standards, aggregations, dis-aggregations, statistical approximations and estimations. Much of this information is documented with accompanying metadata, however, the overarching question for users is just how reliable is the exposure information and how do I know it is fit for the purpose I need?

The Reliability Framework for the Natural Hazard Exposure Information Framework proposes a set of guidelines for data custodians to implement critical elements to enable a better understanding of the reliability of data for decision making purposes. Not surprisingly, these elements include: data provenance (history), currency, completeness and quality, which are normally described to varying degrees in metadata statements. The reliability assessment framework differs from a standard metadata guideline in two ways. Firstly, it describes the process for collecting and collating information about reliability indicators and seeks to use this information to express these indicators in a simple qualitative measure making it easier for an end-user to select the data that matches their requirements. Secondly, the framework includes the ability for end-users to provide feedback about the performance of the data for their particular purpose. Importantly, the end-user feedback loop enables data providers to better understand how their data is being used, what improvements could be programmed into the data supply chain and to clearly communicate the appropriate use of the data.

Implementation of this framework in the future is an important step in not only providing a simpler way of understanding whether data may be fit for purpose, but also in describing a mechanism for end-users to provide valuable feedback to data suppliers to improve the quality of the data for future use.

3 INTRODUCTION

Exposure information is fundamental in the development of risk assessment models for natural hazards, lifeline and infrastructure failures, and the consequences of climate change. Exposure data is also highly useful to underpin early warning systems and support national priority outcomes as described in the National Disaster Resilience Strategy (Council of Australian Governments (COAG), 2011). This exposure data is important in understanding risks, reducing the risks in the built environment, and supporting capabilities for disaster resilience. The information is particularly important for government at all levels to moderate impacts from natural hazards.

Comprehensive and consistent information is needed across the nation for emergency management in order to reduce risks (Harper, 2006). In 2002, COAG announced that it was committed to establishing 'a nationally consistent system of data collection, research and analysis to ensure a sound knowledge base on natural disasters and disaster mitigation' (COAG, 2004). A nationally consistent exposure information framework for natural hazard risk reduction forms the basis of an essential element for coordinated decision-making. The question is whether the information available for decision-making is sufficiently reliable and well understood. Therefore, the present study aims to develop a framework to assess and disclose data reliability. This is done by preparing a list of elements needed, standards to adopt and reliability parameters to address.

The proposed reliability framework can be used to improve existing exposure database capabilities, such as Geoscience Australia's (GA) NEXIS, as well as State Emergency Services systems used for disaster management and risk assessment. The framework can also underpin improvements to databases in support of strategic planning for disaster mitigation. This will assist government (national, state and local) and industry to better understand the reliability of exposure data for decision-making.

3.1 THE IMPORTANCE OF ASSESSING AND DISCLOSING DATA RELIABILITY

Exposure information is managed and delivered to users through an information system and commonly released through a website or web portal (in GA, it is done through NEXIS). The system is very useful for providing decision-makers with situational awareness, but the reliability of the data is not presented to users as there is little room left for such information (Evans, 1997). In this situation, it is difficult for the decision-maker to objectively analyse the real situation (Goodchild et al., 1994) and this may cause a bias in discussing potential solutions (Kobus et al., 2001). Information regarding data reliability can help decision-makers incorporate uncertainty when determining the 'best' answer to a problem (MacEachren et al., 2005).

Another and potentially more significant issue in data delivery is that it may lead some users to a false perception of reality by giving them no knowledge of data quality or placing the information on quality 'behind the screen' in metadata documents. This issue has grown as information systems are being used as decision-support tools. Poor quality data may lead to poor policy decisions (either allocation or treatment) and the impact of this could disadvantage or even be harmful to certain community groups (Onsrud, 1995). Information systems can lead users to think that their underlying databases give representations of objective truth while this may not actually be the case (Veregin, 1999). In reality, there is increasing reliance on secondary data sources such as estimates and predictions. This secondary data may have different characteristics (e.g. the estimate is based on a set of assumptions) that affect the overall reliability while the end users of exposure information systems often assume that their characteristics are uniform (Wong and Wu, 1995; Burrough, 1986). This is when the more conventional method of disclosing reliability, such as metadata, becomes less useful because the metadata needs to contain certain information, such as data provenance, which is hidden in the background by the system's data presentation.

Further use of this type of data is likely to increase this problem, especially owing to the increasing empirical use of spatial data. Linking data to external simulations has become more common and this increases the usefulness of the data (Fotheringham and Wegener, 2000) but the reliability of the data has a major impact on the output. The modeller needs to assess the reliability of the output based on a judgement of the reliability of the input, as well as the simulation model itself (Brimicombe, 2002). This is when data provenance (lineage) becomes important as it tracks changes to the data during processing, which provides a means to judge its reliability (Di et al., 2013). However, data provenance is not the only factor that affects data reliability assessments and the next section discusses these different factors.

4 LITERATURE REVIEW

4.1 WHAT AFFECTS DATA RELIABILITY?

Before reliability assessment can be discussed, it is important to know what constitutes data reliability. Veregin (1999) notes that to understand the quality of data, we need to understand intangible aspects of the data, including its accuracy and its scope as well as how it is produced. This is because the existence of something in an information system (recorded in the data) is not only dependent on space and time but also on the theme of the data (which determines the data scope). This theme will determine whether recording the existence of a particular feature is important or not. Without this understanding, data that could be considered good quality for one use may actually be useless for others as it is may not be fit for their use (Brimicombe, 2002; Heuvelink, 1998).

Important too is key metadata, such as author(s), release date and time represented by the data, title, version, archive (and/or distributor), locator or identifier and access to the data,. With this information, data users may judge if the data is authoritative and produced by an institution that has a mandate for its production or in maintaining its currency. To sum up, data reliability does not only depend on accuracy (spatial, temporal, representational) but also other factors detailed in the following sections.

4.1.1 Lineage or Provenance

Provenance means origin, and in databases, it relates to the process by which data is produced (Buneman et al., 2001). Provenance is a fundamental factor because it is often required to assess the trustworthiness of data (Di et al., 2013). Provenance information often indicates the level of reliability as it shows how the data is produced. An example in the exposure information system context would be whether the data is captured by satellite, surveys on the ground or both. There may be a large amount of information on provenance because data may be produced by custodians through varied and lengthy processes (Chebotko et al., 2011). The provenance information needs to capture all these different steps. For that reason, provenance is also defined as history (or lineage) of data in terms of workflow context as well as web context (Wong and Wu, 1996; Di et al., 2013).

There are two main components of lineage information that users need to know: the first is the source materials and the second is the method of derivation (Veregin, 1999; MacEachren et al., 2005). MacEachren et al. (2005) note that categorisation based on these two components is complex. This is because it may include various subcomponents of the derivation of source materials rather than the particular data itself. The subcomponents of method of derivation may include specification of processes and individuals involved in their creation. Therefore, the amount of lineage information can be very large if it contains material related to the process of creating the source materials. Furthermore, the subcomponents may also include information about other reliability aspects linked to the materials and method of derivation such as accuracy, currency, precision, and others that will be discussed below. Lineage information relevant to users may therefore need to be queried from, or assessed based on, the source materials (Di et al., 2013).

One of the issues in providing information on the provenance or lineage of a database is that the diverse types of data in different databases will require different provenance representations for different needs (Di et al., 2013). This affects the way the provenance information needs to be captured. Uniform components of the provenance information could lead to automated generation of this information. This large volume of information could then be stored in metadata and returned on query. This process is less complicated if the information needed by the end-users is similar'. A related issue is the visualisation of vast amounts of provenance information for users. Di et al. (2013) argue that the important thing in this regard is how to help users discover anomalies and evaluate the information they have received.

4.1.2 Accuracy and Currency

Accuracy in the exposure data context relates to the data error, which means the difference between observation and reality (MacEachren et al., 2005); therefore, accuracy can mean how good the data is at representing the reality that users need to know. Accuracy of data is so important that it was proposed as one of the standard components of metadata (Federal Geographic Data

Committee, 1994). The problem is that the accuracy of spatial data often differs, which means data from different locations may have different spatial accuracy (Wong and Wu, 1996). This is not only true for positional accuracy but also attribute accuracy.

Positional accuracy, or spatial accuracy, refers to the closeness of the spatial components (Veregin, 1999) in data and determines the difference in the recorded positions of objects and features compared with perfectly measured reality (Wong and Wu, 1996). As spatial information systems have introduced three-dimensional data, positional accuracy needs to be considered in both the horizontal and vertical planes. Positional accuracy of this sort can only be measured if the data can be observed and compared with on-ground reality (Chrisman, 1991). There are several methodologies to provide proxy for reality'. The preferred one is using independent sources of higher accuracy. This can then be used to determine the root-mean-square error of the data for a well-defined point. Other methods are deductive estimates based on knowledge of error, comparison with the source using check plots, and internal evidence such as journals or records of the existence of certain features in a particular location, for example, the fire map of a bushfire event matched with the activity report of the area at that time.

Attribute accuracy looks at the discrepancy in thematic element measurements (Veregin, 1999): it concerns error in the description of features. For quantitative attributes, similar procedures to those used with positional accuracy can be applied, and produce error measures such as the root-mean-square error. On the other hand, for non-quantitative attributes, deductive estimates or known error are a better option. Other options to measure attribute accuracy are by sampling or map overlay. One source of attribute inaccuracy is spatial aggregation, as the generalisation of the data makes the classification for a single sample point location no longer be aligned with what is on the ground (Wong and Wu, 1995).

Another type of data accuracy is temporal accuracy. According to Veregin (1999), this type of accuracy has not received much attention given that conventional geospatial data does not explicitly reveal time vectors. Nevertheless, this type of accuracy may actually equate to 'current-ness' (Thapa and Bossler, 1992) or currency (MacEachren et al., 2005), which is whether data is still valuable at the time it is used given the delay since its collection. MacEachren et al. (2005) remind us that the currency of the data also depends on context, and give an example of a car park in a factory, where the year-old data on the cars in the carpark is less likely to be current than the location of the factory.

4.1.3 Precision and Completeness

Precision refers to the exactness of measurement (MacEachren et al., 2005). In spatial databases, the term that is commonly used is resolution and this relates to how detailed the data is (Veregin, 1999). The level of aggregation and categorisation determines the precision of the data (Evans, 1997). Precision also depends on the measurement parameters and the estimation procedure or device (MacEachren et al., 2005).

Like accuracy, precision can also be seen from three aspects: spatial, temporal and thematic (Veregin, 1999). Spatial precision relates to the dimension of a picture element or pixel. It can also depend on the size of community that can be represented by a polygon. Temporal precision relates to the recording interval. The more often the data is updated, the more precise it is. However, similarly to temporal accuracy, context is important in temporal precision as different data may have different rates of change. Thematic precision relates to the measurement scale as well as the classification used in categorical data. The more refined the classification, the more precise the data will be. Although there are many similarities between precision and accuracy, a major difference between the two is that precision is about scale while accuracy is about correctness.

Both precision and accuracy may be unimportant if the data is incomplete. Completeness can be defined as the comprehensiveness of the data (MacEachren et al., 2005). It describes whether the information of interest is in the scope of the data. The completeness of the data depends on how the relationship between the object in the database and the 'abstract universe' is being represented (MacEachren et al., 2005; Veregin, 1999). This will determine which objects need to be covered by the database (Wong and Wu 1996). For example, if a person is standing beside a tree, precision may be detailed enough to put the tree in its place with good accuracy, but as the man is not covered by the database, then the database will not identify the existence of a person there.

4.1.4 Other Factors

Other factors may affect the reliability of a database and consistency is one of them. By definition, consistency in spatial information refers to 'the fidelity of the relationships encoded in the database' (MacEachren et al., 2005; Veregin, 1999; Wong and Wu, 1996). In other words, it refers to how well abstract reality is being transformed into code in the spatial database. This can be assessed from any apparent contradiction in the database (Veregin, 1999). Kainz (1995) notes that consistency means that the data follows topological rules such as that no two points are at exactly the same location or that polygons are fully bounded by lines.

The combination of different reliability factors determines the credibility of the database (MacEachren et al., 2005). It is important to note that the credibility of a database is often judged by the users' experience, and therefore, the credibility of the data also depends on the judgement of the user. The judgement of the data provider or constructor is also an important factor, because the construction of data involves some human interpretation and judgement and thus, there is some level of subjectivity in the data.

4.2 DATA RELIABILITY AND USERS

Throughout the discussion above, it becomes obvious that the reliability of data depends on the needs of the users. This follows Heuvelink's (1998) assertion that 'Data, even when it is high in quality, may not guarantee fit to be used by anyone'. As a consequence, the need for reliability information also varies for different users. Despite this, Evans (1997) found that most users can decide whether to use the data or not when presented with reliability indicators. This is also supported by MacEachren et al. (2005), who point out that research on data uncertainty of cartographic representations suggests that inclusion of this information is helpful to decision-makers but the lack access to real-world verification leads users to take for granted visual depictions from an information system.

MacEachren et al. (2005) also point out the fact that decisions (e.g. policy decision) have to be made despite reliability issues in the data. Some users depend on statistical analyses to decide how spatial information contributes to their decision. On the other hand, there are some who rely more on heuristic methods to decide whether to believe the information or not. MacEachren et al. (2005) argue that the decision to implement explicit reliability indications should be based on the way the decision is made and the likeliness of the outcome,. This is because the information could prevent certain users from seeing important patterns. Further, the sense of uncertainty in the data could lead to a sense of ambiguity when a decision is made and, therefore, decision-makers may end up being unable to make a convincing decision because of this uncertainty (Cliburn et al., 2002).

Another important factor is how immediately the decision needs to be made (Kobus et al., 2001). Some decision-makers do not have time to analyse the data and its reliability. In some cases, decisions have to be made continuously using new information that is also flowing in continuously. One example of such a decision-making process is at the front line of either a natural disaster recovery and rescue mission or in a military operation. In this case, it is important to not only consider how the notion of uncertainty could affect the decision but also whether the information itself is the most crucial factor that needs to be looked at.

There are several important points from MacEachren et al. (2005) that can really affect the framework for data reliability information. The first one is the relationship between the uncertainty and the information needed by certain users in certain domains. This relates more to how important the information is for one user compared with another. The second factor is how the uncertainty affects the analysis, decision-making and, hence, the outcome. These two factors affect the framework for presenting information as well as the framework that related to the visualisation of data reliability.

5 BUILDING THE FRAMEWORK

5.1 EXISTING FRAMEWORKS

In this section, we discuss available frameworks including data, metadata, and user identification frameworks for developing similar systems Australia.

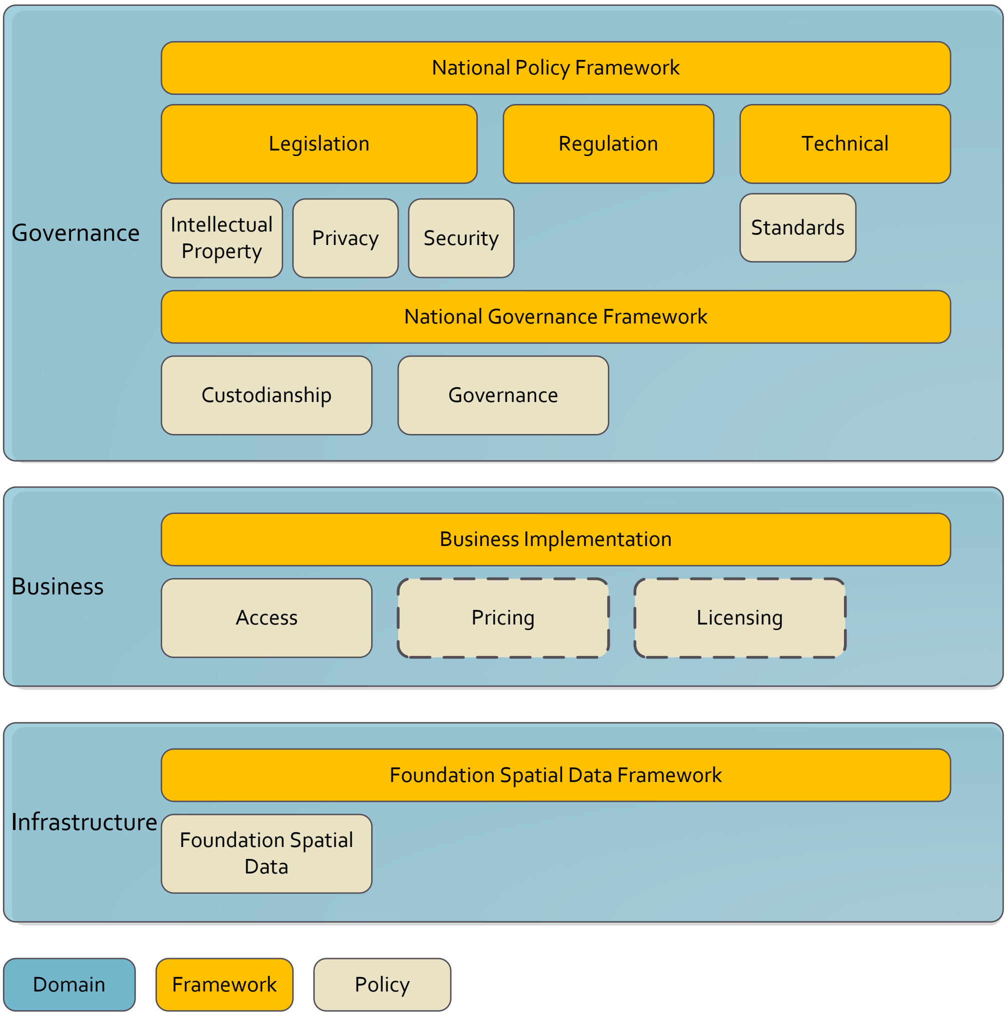
5.1.1 Foundation Spatial Data Framework (FSDF)

The Foundational Spatial Data Framework (FSDF) provides a common reference for the assembly and maintenance of Australian and New Zealand foundational spatial data in order to serve diverse users. The FSDF is an Australia and New Zealand Land Information Council (ANZLIC) initiative. The FSDF is part of an overarching policy framework related to spatial information in terms of governance, custodianship, standards, access, privacy, security and intellectual property. Figure 1 shows the position of FSDF in the management of spatial information in Australia.

The custodianship policy is an important part of a data reliability framework. This is because the custodian of the data has to ensure appropriate care in the collection, storage and maintenance of the information. In particularly, the dataset has to be 'collected and maintained according to certain specifications' and in 'a format that conforms to standards and policies established for the national spatial data infrastructure' (ANZLIC, 2014). This means if a custodian of the data has been appointed, then the responsibility to ensure reliability of information falls on them.

This work is still ongoing as currently the FSDF team is recording source datasets required from States and other jurisdictions to assemble the 'FSDF Datasets'. This includes information about the scope of responsibility or mandates and the funding for source datasets. Currently, the FSDF website (http://link.fsdf.org.au) provides information about how the national FSDF datasets are created (who contributes to making them), why these national datasets are important (focusing on related mandates, i.e. legislation associated with the data), use cases for the national datasets, and future requirements to ensure the FSDF datasets are relevant into the future.

FIGURE 1. SPATIAL INFORMATION MANAGEMENT POLICIES ACCORDING TO FSDF



NOTE: the pricing and licensing in dotted boundary box are not explicitly separated in the policy statement but included in the access (and intellectual property)

COURTESY: INTERGOVERNMENTAL COMMITTEE ON SURVEYING AND MAPPING (ICSM) as cited by ANZLIC FSDF document

The Data Reliability Framework as part of the Natural Hazard Exposure Information Framework should align with the FSDF, in particular the information management policies for custodianship and standards, which provide requirements, specifications, guidelines or characteristics that can be used consistently to ensure materials, products, processes and services are fit for purpose.

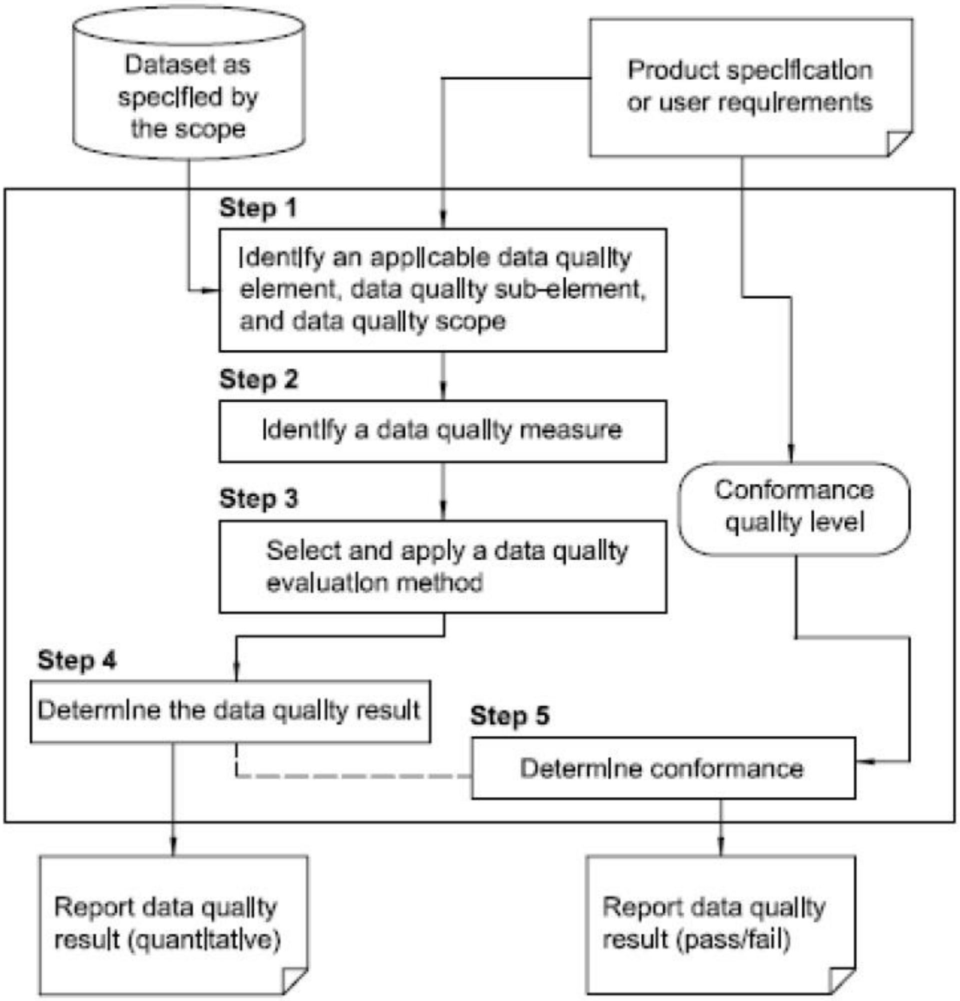
The use of metadata standards ensures that data is consistent with specifications and contains characteristics that enable the data to be exchanged between institutions. The FSDF references the International Organisation for Standardisation (ISO) Technical Committee 211 on

Geographic Information/Geomatics Standards (ISO 2003a; 2003b). Application of this standard, particularly to the data quality section, is important for building the data reliability framework.

5.1.2 ISO Geographic Data Framework

Similarly to the objective of this data reliability framework for existing exposure information systems, the aim of the geospatial data quality standard in ISO is to facilitate the selection of the geographic dataset best suited to application needs or requirements. Information on the quality of geographic data allows the validation of a dataset, which assists a data user in determining a product's ability to satisfy the requirements for their particular application (ISO, 2002). Nevertheless, the international standard will not be the minimum requirement for the data custodian, and instead serves as 'the principles for describing the quality of geographic data and specifies components for reporting quality information'. This description includes data quality elements that describe how well the data meets the specifications (which can be judged from the level of completeness, logical consistency, positional accuracy, temporal accuracy, thematic accuracy) as well as a data overview that provides general information about the data. These two components need to be provided in metadata (ISO, 2003a). The other standard that will be of benefit to the exposure data reliability framework is the procedure for determining and evaluating quality (ISO, 2003b). As discussed in the previous section, the framework must recognise the needs of different users and this is acknowledged in procedures that require both data producers and data users to contribute by expressing how well the product meets its specification and establishing the extent to which a dataset meets their requirements, respectively (Figure 2).

FIGURE 2. DATA QUALITY EVALUATION PROCEDURE


Dataset as specified by the scope
Product specification or user requirements
Step 1
Identify an applicable data quality element, data quality sub-element, and data quality scope
Step 2
Identify a data quality measure
Step 3
Select and apply a data quality evaluation method
Conformance quality level
Step 4
Determine the data quality result
Step 5
Determine conformance
Report data quality result (quantitative)
Report data quality result (pass/fail)

SOURCE: INTERNATIONAL ORGANISATION FOR STANDARDISATION (ISO, 2003b)

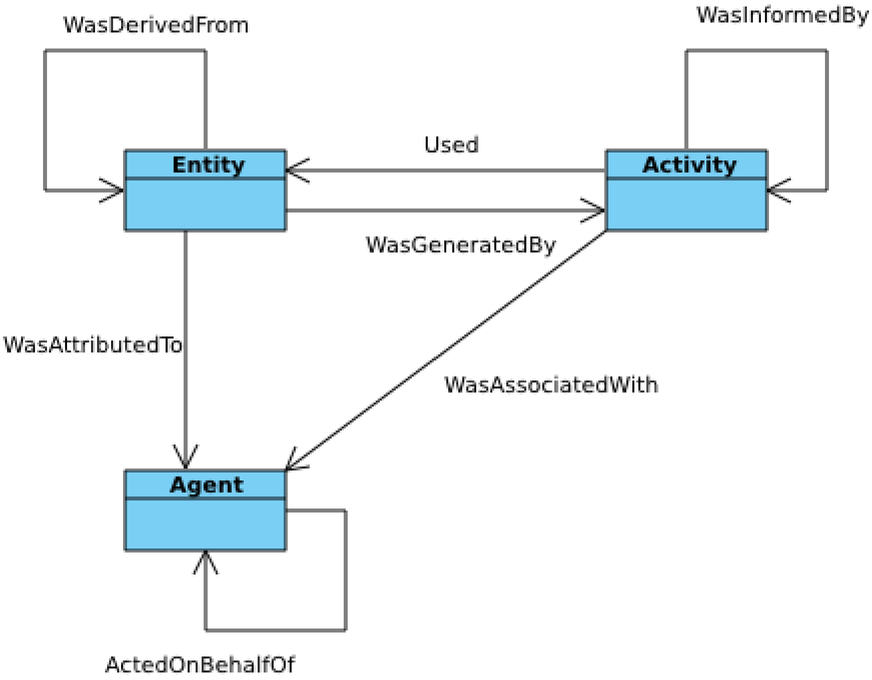
5.1.3 Data Provenance Framework

Data provenance is highly significant because it enhances the end users' trust and may include data reliability components within it. A provenance model is also a component of the FSDF. The general provenance model used for the FSDF is the PROV-DM (Provenance Data Model). The model is still under development but it has been planned to capture all additional metadata for datasets and custodians. The development will follow the formalisation of FSDF's metadata requirements.

PROV-DM is developed by the World Wide Web Consortium (W3C) to distinguish core structures and form the essence of provenance information. There are six components that PROV-DM deals with: Entities (data items); Activities (processes); Agents (custodians and other people, and organisations with roles in relation to Entities and Activities); the bundling process; the linking of Entities; and collections of logical structure. In this model, provenance records the people, institutions, data and processes involved in producing, influencing, or delivering data (Moreau and Missier, 2013).

The core structure of PROV-DM describes the use and production of Entities by Activities, which may be influenced in various ways by people or organisations (Agents). These core types and their relationships are illustrated in Figure 3.

FIGURE 3. PROVENANCE CORE STRUCTURES



SOURCE: MOREAU AND MISSIER (2013)

Figure 3 illustrates three core components of PROV-DM and their relationships in the core structure. A database could be modelled as an Entity in the diagram in Figure 3. However, the class Entity also contains any other conceptual, physical or digital things captured in a system. A data custodian is an Agent in this diagram. In addition to humans and organisations, systems may also be termed Agents, according to PROV-DM, if they cause actions to occur. Besides having the authority over particular data (Entities by virtue of Roles, something that is also modelled in PROV-DM), Agents also perform Activities, defined as something that was done, on or with Entities over a period of time. In summary, the PROV-DM models not only information about data but also causative agents that perform processes on data.

According to Car (2016), this framework can be used to produce an assessment of the reliability of the data. He offers two general options to assess the reliability of the data for which provenance is recorded. The first is by checking the history of all the provenance components and comparing them with some specific, desired criteria. Provenance about both the data ancestors and Agents who are responsible for data, as well as methods used to produce the data or information may all be relevant to a reliability assessment. The second option is to look at how data is perceived by other users of it. Statistics about use, who in particular used the data and how (the methods that need to be applied when using the data) may all be relevant.

5.2 DATA USERS

The two frameworks discussed above and the literatures reviewed indicate the importance of users in assessing the reliability of the data and relevant frameworks. The exposure information systems across the nation have recognised the diverse utilisation of their data and focus on end-user requirements. The Natural Hazards Exposure Information Framework categorises its users into three levels to reduce complexity (Nadimpalli and Mohanty, 2016). One important aspect in this differentiation is the information requirements of the different users and levels of disaster governance.

The first level of user identified by the exposure information framework is those who use the data for policy and planning at Commonwealth and State or Territory Government level. This level of use does not require highly precise data, either spatial, temporal or thematic. For these users, data can be aggregated for a defined geographic area with a combination of the necessary exposure information for buildings, population, business and infrastructure. For example, a combination of building elements such as building type, wall type and roof type, together with household income profile of people in a given geographic area would be aggregated and made available. The defined geographic area includes derivatives of some of the Australian Statistical Geographic Standard (ASGS) data such as those from Statistical Area 2 (SA2) or larger.

The second level of user uses the data for planning proposes in both tactical response and for strategic perspective. This level includes State, Territory and local government, researchers and the insurance sector. The precision of data required will be higher than that required by Level 1 users. The data could be aggregated for a defined geographic area with more detailed exposure classifications for buildings, population, business and infrastructure. This means the exposure information available at this level includes some of that found in the smallest ASGS area, Statistical Area 1 (SA1). Other exposure information available for this type of user is raster data in 1 × 1-km or smaller grid cells.

The third level of user is those who use the data for research and analysis at the asset or building level, data that may be needed to provide more detailed advice for decision-makers. These users require the most detail for analysis and emergency operations. Therefore, the data sourced for this level must be highly authoritative and reliable, and mapped at the asset or single-building level.

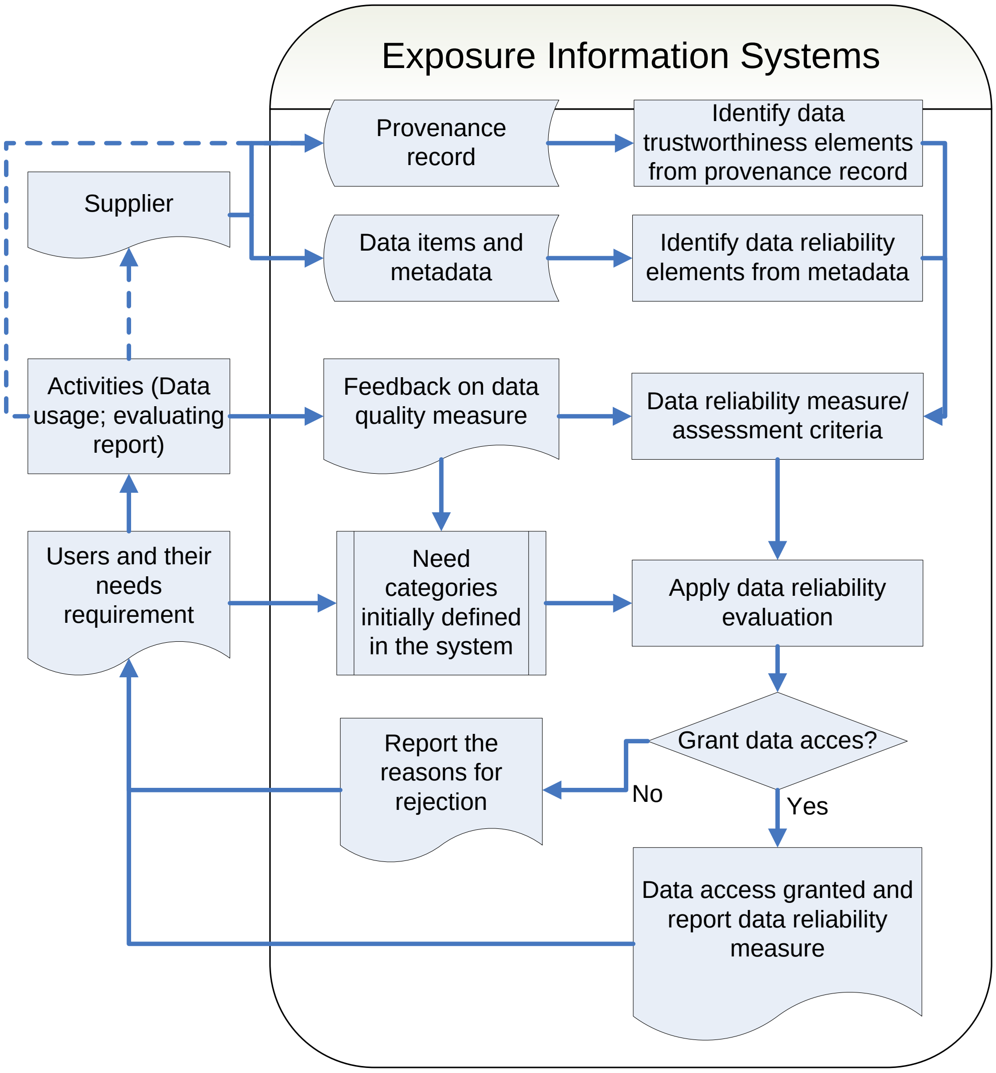
This categorisation is useful for determining what data could be released for particular users. Sometimes the same data will be used for various purposes and in various situations by the same user. The same data user may also act differently in these different situations. Therefore, the three user levels above should be based more on the need and requirements of the user at the time and less on the criteria of the user. This means each user needs to define whether their current need is related to Levels 1, 2 or 3. It should be noted that this categorisation from Level 1 to 3 will serve only in the initial stages of the framework. As input from users trickles in, the framework must communicate with users about whether those who have been grouped in the same level really have similar needs. If not, then the framework should be ready to introduce a new grouping to accommodate these users.

5.3 THE PROPOSED FRAMEWORK

The reliability framework to be built for exposure information systems should closely follow the ISO’s data evaluation process. This means the exposure information framework needs to provide specifics about how a dataset could be released as well as the necessary advice and metadata that accompany it. For this purpose, users have to submit their requirements regarding the quality level they require in order to assess the appropriateness of the datasets. Initially, this is fulfilled by the three categorisations of users discussed above. The framework also adopts the data reliability assessment framework discussed by Car (2016) by using the history (provenance) of the data as one component of reliability measure while also adopting the idea of 'forward provenance' by taking into account how the data will be used. To do so, the framework requires users to submit data and its metadata to information and provenance systems. Users can also re-submit their feedback as well as requirements to adjust the initial criteria that the exposure information framework sets out as default.

Figure 4 shows a process that adopts the two frameworks above and can be applied in an exposure information framework. The provenance framework is prominent here as it is recognised owing to the multiple ways in which provenance information is relevant to quality assessments. If this is not currently possible, then exposure information systems have to assist the supplier to fill in these components. In addition, the framework results for datasets should be able to change as the data supplied is updated and used by users. The framework should be able to capture the metadata from previous data to fill in the new metadata as well as record the provenance of data change, given that product is likely to be considered as new after each change. The two components – provenance and metadata – can be seen as one package, but it might be useful to separate them as provenance could be captured directly from activities. Nevertheless, as exposure information systems do not include the capability for users to change or create new data, suppliers should be required to submit their activity information as inputs into provenance.

FIGURE 4. PROPOSED DATA RELIABILITY FRAMEWORK FOR EXPOSURE INFORMATION SYSTEMS


Exposure Information Systems
Provenance record
Identify data trustworthiness elements from provenance record
Supplier
Data items and metadata
Identify data reliability elements from metadata
Activities (Data usage; evaluating report)
Feedback on data quality measure
Data reliability measure/ assessment criteria
Users and their needs requirement
Need categories initially defined in the system
Apply data reliability evaluation
Grant data acces?
Report the reasons for rejection
No
Yes
Data access granted and report data reliability measure

In the next stage, the exposure information framework will identify data reliability elements. These elements come from both the metadata and the provenance record. Given the provenance record may also contain series of metadata, the overall assessment of the provenance needs to be given first for each of the data items in the dataset. The exposure information framework is able to extract the reliability elements of each data item from metadata. This includes extracting the precision, accuracy, currency and completeness status. In this framework, the suppliers are required to fill in the metadata but the framework also needs to be ready to use the closest available metadata based on provenance (i.e. metadata of the data being used or updated). It is important to note that users who use more aggregated data, such as the

Level 1 user, are more likely to find more reliable data items for their purposes in a dataset.

5.3.1 Classification of Data Based on Reliability Information

After information is collected, it goes through the reliability measurement framework. The provenance of data items is assessed first. Initially, the framework focuses on the journey of the data, such as where it originates from, perhaps from a survey (census), satellite capture or administrative report, and then it will look at whether or not it has been estimated through sampling or other statistical techniques. It is also possible that the history of the data is unknown or unclear. An example of assessment using this classification methodology is as follows:

1. Original: survey, administrative data, satellite capture
2. Original estimate from big sample
3. Modification from original data
4. Modification from estimate
5. Modification from modified data (second in line from the origin)
6. Modification from modified data (third or more in line from the origin)
7. Unknown.

In this example, the provenance can be used to automatically update the status of the data every time it goes through modification. As stated above, the provenance may also contain historical metadata, information that can be included in the provenance assessment measure. However, the provenance classification does not carry over the information from previous metadata because this would produce complicated data with unlimited possibilities of a provenance quality level that may not be comparable with others. For example, it can use accuracy, currency or precision and produce categories such as 'modified from estimated data with 95% accuracy', 'modified from data older than 5 years ago' and 'modified from data at States or Territories level', respectively.

The next step in the framework is to look at the data reliability component. The main components are accuracy, currency and precision. Accuracy and precision can be differentiated into spatial, temporal and thematic. Some of the issues that cannot be captured by the provenance described above can be captured by presenting reliability components directly. This is because some of the historical metadata information can be referred to in current metadata. For example, the accuracy of the current data will depend on the accuracy of the input data. The historical metadata components may be reflected by a combination of the reliability components. For example, the output data can be disaggregated from the input data, which would provide a higher precision, with the cost of lower accuracy.

The challenge at this stage is for the exposure information framework to come up with measurement criteria for the different reliability components. As discussed in the previous section, the criteria will need to vary based not only on the type of component but also on the user and the data item itself. To deal with this issue, the framework provides the initial criteria while taking input from users for further consideration. In Figure 4, this is shown by the arrow from the user requiring feedback after activity or interaction with the data. Here are some examples of the measurement criteria:

Accuracy:

1. At least 95% of the data is correct;
2. 80% to less than 95% of the data is correct;
3. 70% to less than 80% of the data is correct;
4. and so on.

Currency:

1. Data represents conditions less than 1 month ago;
2. Data represents conditions around 1 year ago;
3. Data represents conditions 1 to 5 years earlier;
4. Data represents conditions 5 to 10 years ago;
5. and so on.

Spatial Precision:

1. Data is captured in almost exact location with 10 × 10-m grid cells;

2. Data is captured in 1 × 1-km grid cells or is captured by SA1 area;
3. Data is captured in 5 × 5-km grid cells or is captured by SA2 area;
4. Data is captured in 50 × 50-km grid cells;
5. and so on.

As can be seen above, the list contains the measures as absolute categories and criteria rather than filled according to the needs of specific users or, as discussed in the literature review section, relative to the expected time length for which certain data is deemed to be reliable. The reasons for this are twofold. The first is for flexibility and second is to enable the user to track the reason for including the criteria in data as well as to giving their input on these criteria. As a consequence, the assessment process will need to proceed to the next stage where a different assessment can be given by and for different users.

Table 1 illustrates the information card resulting from applying measurement criteria to data items. The illustration includes four variables or indicators. For example, these indicators could be the building location, various types of roof in an SA1 area, vehicle ownership at the SA1 area level and the location of parking. It is important to note that exposure information systems have to generate these initial criteria based on the information they have about the needs of the data end-user and how they use the data.

TABLE 1. AN ILLUSTRATED SCORECARD OF DATA ITEMS

	Data Items			
	Building Location	Various Types of Roofs in SA1	Vehicle Ownership at SA1 Level	Parking Location
Provenance	1	2	3	4
Accuracy	2	1	4	3
Currency	4	1	1	2
Precision	1	2	2	1

5.3.2 The Assessment of Reliability

The criteria set by exposure information systems not only serve varied needs but also anticipate potential problems that can arise in data utilisation. Therefore, the next stage of the framework, which is the reliability evaluation process, applies different assessment thresholds designating the quality of data': high, medium or low for different levels of user. Table 2A–C shows an example of how this differentiation can take place. For example, 'building location' data, derived from satellite data that may include clouds, would have a reduced accuracy'. The required quality of the data is high for Level 1 and 2 users who look at the data in aggregate; thus, the effects of clouds in the data for them is insignificant. For Level 3 users, this could be a problem as they may need to look specifically at the area under cloud.

Even if the data was collected more than 5 years ago, it may still be valuable for Level 1 users since at the aggregation level they are working at, the changes of 5 years may not substantially alter patterns. However, this quality of currency will likely be considered too low for both Levels 2 and 3. Currency is, of course, also subject to both the particular end-use and the particular data item as some uses and items are sensitive to time.

Table 2 also shows how the quality classification of high, medium or low can be allocated differently across different data items. While Table 1 captures and categorises the quality information immediately from the metadata, the assessment of specific users of various data items could be different and, therefore, data items may have different thresholds that need to be stated as high, medium or low. The currency of a parked car and factory location discussed earlier is one example of this issue: the knowledge about a parked car from a year ago is not as useful as the information from the same time about factory location. In particular, Table 2 shows different categories are given for Level 3 users in regard to the information of various roof types and the information about car ownership at the SA1 level. In this example, the user is hoping for more detailed information about which houses have vehicles while expecting that the roof type in a certain SA1 area is more or less uniform.

As can be seen in the car and factory example, setting the assessment threshold as the second step of the process provides flexibility. This is because the characteristic described in the first step can mean a different thing for different users as well as for different data items. Therefore, it is impossible to assess the reliability criteria for the data relative to an expectation at the first step because the expectation can differ. On the other hand, if the assessment is being done directly in the first step, then there is a possibility that the categorisation is being done randomly owing to lack of framing and the availability of too many options. This is especially relevant to the feedback loop expected from users for the advancement of assessment in the framework (Figure 4). Input from the users is likely to affect the assessment process in the second step, especially the threshold to categorise whether the data has high, medium or low reliability. This does not mean that the criteria in the first step cannot be changed but they are expected to be more rigid and sustainable.

TABLE 2. THE ASSESSMENT PROCESS FOR ILLUSTRATED DATA ITEMS IN THREE LEVELS

A

Level 1	Data Items			
	Building Location	Various Types of Roofs in SA1	Vehicle Ownership at SA1 Level	Parking Location
Provenance	High	High	Medium	Medium
Accuracy	High	High	Medium	Medium
Currency	Medium	High	High	High
Precision	High	High	High	High
Data release	Granted	Granted	Granted with warnings	Granted with warnings

B

Level 2	Data Items			
	Building Location	Various Types of Roofs in SA1	Vehicle Ownership at SA1 Level	Parking Location
Provenance	High	High	Medium	Low
Accuracy	Medium	High	Low	Medium
Currency	Low	High	High	High
Precision	High	High	High	High
Data release	Granted with warninsg	Granted	Not granted	Granted with warnings

C

Level 3	Data Items			
	Building Location	Various Types of Roofs in SA1	Vehicle Ownership at SA1 Level	Parking Location
Provenance	High	High	Low	Low
Accuracy	Medium	Medium	Low	Low
Currency	Low	Medium	High	Medium
Precision	High	High	Medium	High
Data release	Not granted	Granted with warnings	Not available	Not granted

The next stage in the framework is the decision about whether to grant the user access to the data and what warnings it needs to carry if given. The assessment of high, medium or low quality or reliability of the data should be the main reason for this decision. The criteria could again be set based on how its data quality elements behave or, at least, how we perceive the importance of the data element. If we assume provenance is the most crucial information for quality assessments, exposure information systems should not release data with low quality indicated by provenance. The important step is then to release a report to the user whether they granted access to the data or not.

5.3.3 User Feedback and Refinement of Assessment

As can be seen in Figure 4, the framework requires data users to always provide feedback to the system. This is important to recalibrate the data element criteria, the assessment threshold and the basis for granting the data. For example, if the data is assessed as unsuitable and the data user feels the reason for not granting the data is not strong enough, they can submit feedback arguing against the decision. The same thing would also be applied if the user felt that the standard for release of data was too low. Exposure information systems will eventually need to assess these submissions and reset the criteria and threshold to a more suitable level.

The next question for the framework is how to effectively manage and incorporate the feedback into the assessment system. The framework has an initial user level to differentiate the users who will be feedback contributors. However, within these grouping levels, there is a large variation of users regarding the frequency of usage, the variety of data required and the level of experience as well as other factors. All these can be taken into consideration in accommodating feedback using a weighting system or using an artificial intelligence process such as fuzzy logic. In doing so, the framework has to first capture these characteristics in the user profile. The user profile may also help in dealing with other issues such as determining the accuracy of feedback and reducing feedback bubbles if the usage is mostly coming from certain groups or a sub-section of users.

6 SUMMARY AND CONCLUSION

This study looked at a possible data reliability framework that could be applied to exposure information systems to ensure that the users are aware of the quality of the data they receive. We have reviewed the different data reliability elements that should be included as well as the currently available frameworks that can be the basis for putting together this data reliability framework. Building from ISO data quality evaluation procedures as well as data provenance models, we propose our data reliability framework for exposure information systems. A significant feature suggested in this framework is for exposure information systems to start with an initial threshold within the system but be open to user input to re-evaluate the standard put in place to better reflect their information requirements.

This reliability framework is generic and provides guidelines only. The data custodians have to address specific dataset issues and plan comprehensive reliability indicators at the micro level for implementation to take place. The framework only deals with the overall dataset and not the individual data elements that may have different spatial qualities within it. This condition can occur when data is partially updated. A reliability framework should be developed to assess the individual elements of the database. However, it is likely systems will be challenged while attempting to visualise quality differences at individual locations. Another issue that needs further investigation and development regards communication between users and data custodians. For a good response to feedback from users (e.g. in terms of reclassifying or weighting the feedback), it is necessary for the framework to know who the user is. Therefore, the framework will need a clear registration and sign-in system, which has not been discussed in this article.

7 REFERENCES

ANZLIC, (Australian New Zealand Land Information Council) (2014), The Australian and New Zealand Foundation Spatial Data Framework: FSDF Spatial Information Management Policies - Custodianship. Canberra: Commonwealth of Australia.

Brimicombe, A.J. (2002), *GIS – Where are the frontiers now? GIS Conference Proceedings*, GIS 2002-International Conference and Exhibition 11–13 March, Bahrain: The Bahrain Society of Engineers, p. 33–45.

Buneman, P., Khanna, S., Wang-Chiew, T. (2001), Why and where: A characterization of data provenance. *International Conference on Database Theory Proceedings*, Bussche J.v.d. and Vianu, V. (eds.) 8th international conference, London, UK, January 4-6. Berlin Heidelberg: Springer, pp. 316–330.

Burrough, P.A. (1986), *Principles of Geographical Information Systems for Land Resources Assessment*. Oxford: Clarendon Press.

Car, N. (2016), *Data Reuse Fitness Assessment Using Provenance*, Conference paper. SciDataCon2016 11-13 September, Denver, Colorado, USA Available at http://www.scidatacon.org/2016/sessions/53/paper/47/ [Accessed 13 July 2017]

Chebotko, A., Simmhan, Y., Missier, P. (2011), 'Guest editorial: Scientific workflows, provenance and their applications', *International Journal of Computer Application*, 18(3): 130–132.

Chrisman, N.R. (1991) 'The error component in spatial data.' In (Eds Maguire DJ, Goodchild MF, Rhind DW) *Geographical Information Systems*, New York: John Wiley & Sons Inc, pp. 165–174.

Cliburn, D. C., Feddema, J. J., Miller, J. R., Slocum, T. A. (2002). Design and evaluation of a decision support system in a water balance application. *Computers and Graphics*, 26(6), 931-949.

CoAG, (Council of Australian Governments) (2004), Natural disasters in Australia: reforming mitigation, relief and recovery arrangements. Canberra: Commonwealth of Australia.

CoAG. (2011), *National Strategy for Disaster Resilience: Building the resilience of our nation to disasters*. Canberra: Commonwealth of Australia.

Di, L., Yue, P., Ramapriyan, H.K., King, R.L. (2013), 'Geoscience data provenance: An overview', *IEEE Transactions on Geoscience and Remote Sensing*, 51(11): 5065–5072.

Evans, B.J. (1997), 'Dynamic display of spatial data-reliability: Does it benefit the map user?' *Computers & Geosciences*, 23(4): 409–422.

Fotheringham, A.S., Wegener, M. (2000), *Spatial Models and GIS: New Potential and New Models*. London: Taylor & Francis.

Goodchild, M.F., Chih-Chang, L., Leung, Y. (1994), 'Chapter 7: Visualizing fuzzy maps.' In (Eds Hearnshaw HM and Unwin DJ) *Visualization in Geographical Information Systems*. New York: John Wiley and Sons, pp. 158–167.

Heuvelink, G.B. (1998), *Error Propagation in Environmental Modelling with GIS*. London: Taylor & Francis.

International Standards Organisation (ISO). (2003a), *Geographic Information — Metadata*, No. 19115:2003

ISO. (2003b), *Geographic Information — Quality Evaluation* Procedures, No. 19114:2003

ISO. (2002), *Geographic Information — Quality Principles*, No. 19113:2002

Kainz, W. (1995), 'Logical consistency.' In (Eds Guptill SC, Morrison JL) *Elements of Spatial Data Quality*. Oxford: Elsevier Science, pp. 109–137.

Kobus, D.A., Proctor, S., Holste, S. (2001), 'Effects of experience and uncertainty during dynamic decision making', *International Journal of Industrial Ergonomics*, 28(5): 275–290.

MacEachren, A.M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., Hetzler, E. (2005), 'Visualizing geospatial information uncertainty: What we know and what we need to know', *Cartography and Geographic Information Science*, 32(3): 139–160.

Moreau, L., Missier, P. (2013), *PROV-DM: The PROV Data Model. W3C recommendation; The World Wide Web Consortium (W3C)*. Available at https://www.w3.org/TR/prov-dm/ [Accessed 18 May 2016].

Nadimpalli, K., Mohanty, I. (2016) *Natural Hazards Built Environment Exposure Information Framework*, Milestone Report, No. 165. Victoria, Australia: The Bushfire and Natural Hazard Cooperative Research Centre

Onsrud, H.J. (1995), 'Identifying unethical conduct in the use of GIS', *Cartography and Geographic Information Systems*, 22(1): 90–97.

Thapa, K., Bossler, J. (1992), 'Accuracy of spatial data used in geographic information systems', *Photogrammetric Engineering and Remote Sensing*, 58(6): 835–841.

Veregin, H. (1999), 'Data quality parameters', *Geographical Information Systems*, 1: 177–189.

Wong, D,W.S., Wu, C.V. (1995), 'Quality of aggregated spatial data – A guidance for decision'. In *4th International Conference on Computers in Urban Planning and Urban Management*, *Melbourne*. Melbourne: Anais, pp. 559–570.

Wong, D.W.S., Wu, C.V. (1996), 'Spatial metadata and GIS for decision support.' In *System Sciences, 1996. Proceedings of the 29th Hawaii International Conference on System Sciences*, 3-6 January Institute of Electrical and Electronics Engineers (IEEE), pp. 557–566