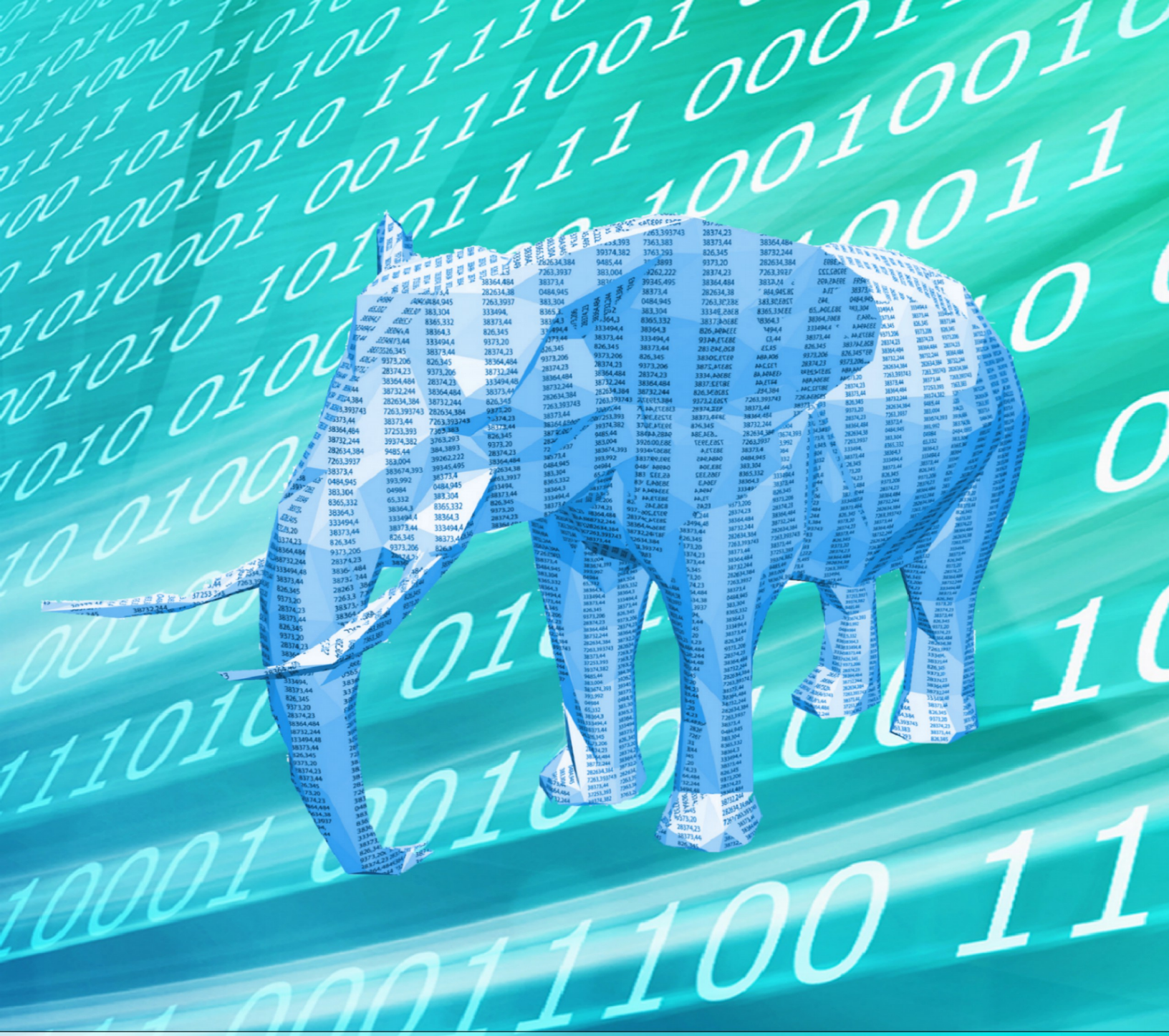


Data Safety Guidance

The Data Safety Initiative
Working Group (DSIWG)

ISBN-13: 978-1519533579

ISBN-10: 1519533578

The SCSC is the UK's professional network for sharing knowledge about safety-critical systems. It brings together engineers and specialists from a range of disciplines working on safety-critical systems in a wide variety of industries, academics researching the arena of safety-critical systems, providers of the tools and services that are needed to develop the systems, and the regulators who oversee safety. It provides, through publications, seminars, workshops, tutorials, a web site and, most importantly, at the annual Safety-critical Systems Symposium (SSS), opportunities for them to network and benefit from each other's experience in working hard at the accidents that don't happen. It focuses on current and emerging practices in safety engineering, software engineering, and product and process safety standards.

This document was written by the Data Safety Initiative Working Group (DSIWG), which is convened under the auspices of the SCSC. The document supports the DSIWG's vision, which is to have clear guidance on how data (as distinct from the software and hardware) should be managed in a safety related context, which will reflect emerging best practice. It was formally released at SSS'16, 2-4 February 2016.

Comments on this document are actively encouraged. These should be sent to: dsiwg@hotmail.com.

Data Safety Guidance

The Data Safety Initiative
Working Group [DSIWG]

January 2016

Change History

Version	By	Status	Date
1.0	The DSIWG Team	First draft for external review	31-JAN-2014
1.1	The DSIWG Team	(Internal edition for DSIWG use only)	09-DEC-2014
1.2	The DSIWG Team	For publication at SSS'15	23-JAN-2015
1.3	The DSIWG Team	For publication at SSS'16	29-JAN-2016

Foreword

Data is here. Data is growing. Data is causing harm.

Data is here: The way that systems are being designed and built is changing. Whereas data was once used simply to configure (and leave configured) a system, its use is rapidly expanding as developers seek to exploit the flexibility provided by data-driven systems. Organisations now make significant decisions (including safety-related decisions) based solely on data held in systems. There is now a clear gap in the way those organisations manage, control and process their data. Key data properties that preserve safety have not been actively managed. Consider, for example, the importance of data defining the layout of Britain's railway signals, data which indicates the position of underwater obstructions in nautical channels or data that records a patient's treatment history.

Data is growing: There are at least two reasons why the use of data has grown and, equally important, why it is expected to continue to grow. The first relates to the rapid expansion of the area loosely termed "Big Data". The second is the growing use of systems of systems, where data is the lifeblood that connects together the disparate elements and allows a cohesive capability to be built. Put simply, the need to address data-related issues is a pressing problem.

Data is causing harm: Strictly speaking, this is not accurate; by itself data can neither cause nor prevent harm. However, mistakes introduced in data, or inappropriate use of data, within safety-related systems have been factors in a number of documented incidents and accidents. Examples include: aircraft attempting to take off from the wrong runway (and consequently crashing); ships running aground; and patients being exposed to higher than planned doses of radiation.

Against this background, the Data Safety Initiative Working Group (DSIWG) was established under the auspices of the Safety Critical Systems Club. The DSIWG's aim is to develop clear, cross-sector guidance on how data (as opposed to software or hardware) should be managed in a safety-related context. For the most part, this guidance is based on well-established techniques. What is new, however, is the explicit and relentless focus on data, making it a "first class citizen" within system safety analyses. By doing so, this guidance should help organisations identify, analyse and treat data-related risks, thus reducing the likelihood of data-related issues causing harm in the future.

Quick Start Guide

- Systems are changing. The role of data is becoming more prominent. Hence, data needs to be considered as a "first class citizen" in system safety analyses. This will help mitigate organisational, and system level, risks associated with the use of data.
- A pan-sector collection of data types has been produced, and a generic set of data properties has been described. These should help identify uses of data within systems. They are discussed in **Section 2**.
- Tools have been developed to support a generic risk management process, including:
  - Assessments to help establish the appropriate context (**Section 4**);
  - Methods for identifying risks, for example Hazard and Operability (HAZOP) study guide words (**Section 5**);
  - An approach for analysing risks, which is based on Integrity Levels (**Section 6**);
  - Support for the evaluation of risks (**Section 7**); and
  - An illustration of methods and approaches that can be used to treat risks (**Section 8**).
- A partial worked example is provided (**Section 9**);
- A collection of appendices provide more detail, including:
  - A brief summary of previous accidents and incidents in which data was potentially a causal factor (**Appendix A**);
  - An Organisation Data Risk assessment questionnaire (**Appendix B**);
  - A Data Safety Culture questionnaire (**Appendix C**);
  - A description of the Dataware Assessment Framework (**Appendix D**).
  - The suggested contents of a Data Safety Management Plan (**Appendix E**);
  - A list of definitions, acronyms and glossary entries (**Appendix F**);
  - A collection of references (**Appendix G**).

Contents

1 Introduction

> *We're entering a new world in which data may be more important than software.*
> ***Tim O'Reilly***

1.1 Motivation

On 9 May 2014 a Qantas 737 was leaving Perth en-route to Canberra. During take-off the aircraft appeared nose heavy. Significant back pressure was required to rotate the aircraft and lift off from the runway. As a consequence, the aircraft exceeded the calculated take-off safety speed by about 25 kt. Fortunately, the rest of the flight passed without incident. Later investigations showed that incorrect passenger data was the main reason for the unbalanced aircraft. Whilst, in this specific case, a data-related problem did not lead to an accident, there are other situations where the outcome has been less benign. A number of these are discussed in Appendix A.

Through incidents like the one outlined above, it has become increasingly clear that system safety depends not only on the system's hardware and software, but also on the data it accepts, generates, processes and produces. As a result, the UK Safety Critical Systems Club (SCSC) has promoted efforts to investigate the issue of data in safety-related systems[1]. This data takes many forms; examples include: Data used by an application, e.g. patient medical records in a hospital; Data about a system itself, e.g. configuration data for a satellite navigation system; Data about users of a system, e.g. operator competence data in a nuclear power plant.

One reason for the growing importance of data is a change in the way that systems are designed and built. Traditionally, components were engineered to specific standards and then configured by data to perform a bespoke role. Today, safety-related systems are becoming more data-intensive and/or data-centric. A key feature of this type of system is that the criticality is inherently with the data, rather than being in a directly controlling function. These data-intensive, safety-related systems are often used as decision support or advisory systems, which support a trained and experienced operator, who may be able to detect and correct data problems. However, data is now so complex and of such a large volume it is increasingly unlikely that a user would spot the data errors, and it would be unreasonable to expect them to do so.

There are also industry trends which make a data safety initiative very timely: the drive towards “Big Data” systems means that safety-related data is being used in more and varied ways, and as part of very large aggregated databases, often via the Internet. Increasing use of Systems of Systems (SoS) technology means that data systems are becoming more highly connected using data from a variety of sources. Hence, the mapping and translation of data between diverse systems becomes more important and more challenging.

Despite the ever growing importance of data, current safety standards and regulations focus strongly on systems, hardware and software development. Data-related aspects are covered comparatively poorly, if they are covered at all. In most sectors, data as separate entity has hardly been considered; the aviation domain is a notable exception. Sometimes certain types of safety-related data have been identified, but then little guidance on how to determine or manage the associated risks is provided. In many standards data is treated in a similar way to software, although the properties it exhibits can be very different. There are also several sectors which now produce and manipulate safety-related data in vast quantities, and

1 The term “safety-related systems” is used throughout this document. It includes systems where failure will lead to immediate harm, as well as systems where failure may not lead to immediate harm but could contribute to its likelihood of occurring. A hospital records system is an example of the latter type.

which are not thought to be covered by any existing safety standards; the Police and Criminal Justice sector is one such example.

The significant, and continually growing, importance of data, combined with the apparent lack of guidance on how associated risks can be managed, provide the background to the SCSC's data-safety initiative. The main motivation of this document is to begin the process of raising the prominence of data to that of a "first class citizen" in the design and implementation of safety-related systems.

1.2 Aim and Scope

This guidance document aims to:

- Describe the data safety problem;
- Provide methods for establishing levels of risk; and
- Recommend strategies and approaches for managing and mitigating those risks.

It should be noted that, whilst they are considered mature enough to be useful, the contents of the document represent current thoughts on what is a complex and evolving area. Furthermore, in order to allow it to be produced within a reasonable timescale, this edition focuses on key data types; it is not intended to be exhaustive.

This document has been written for a wide readership. Its target audience covers all those who have an interest in, or a responsibility for, safety-related data within systems, including: Managers, Developers, Safety Engineers, Assurers (including Independent Safety Auditors), Regulators and Operators.

The breadth of readership is also intended to cover a number of different sectors. As such, the document identifies a wide spectrum of safety-related data that exists in many forms within systems: from specification and requirements data, to maintenance and disposal data, and everything in between. In particular, this document is not just concerned with numerical or well-structured data used during system operation.

The document is not intended to replace or supersede any existing material, whether that be sector-specific or of a more general nature. Nor does it (currently) provide an acceptable means of compliance to any particular standard. Instead, the intent is for this guidance to be used as a supplement to current standards. In the longer-term the hope is that future standards documents take up relevant concepts, approaches and methods from those described here. Ultimately, a separate Data Safety document may not even be required.

1.3 Document Structure

The remainder of this document is structured as follows:

- Section 2 introduces different types of data. Data properties, which provide a way of describing the characteristics data must exhibit to support safe system operation, are also discussed, as are key Data Safety Principles.
- Section 3 contains a general introduction to the process of managing risks. In particular, it establishes a structure that is used to shape subsequent discussions. This section also outlines the links between data safety and security.

- Section 4 is concerned with establishing the context, including the scope of any system assessments that are required and the risk appetite of the system stakeholders.
- Section 5 discusses ways of identifying data-related risks, including a selection of keywords for a Hazard and Operability (HAZOP) study.
- Section 6 provides a means of analysing risks, which is achieved by assigning Data Integrity Levels (DILs).
- Section 7 is relatively brief. Its focus is on evaluating risks; that is, comparing the identified risks (Sections 5 and 6) with the risk appetite (Section 4).
- Section 8 contains a range of methods and approaches that can be used to control different types of data-related risk in different contexts.
- Section 9 provides a partially worked example of applying the Data Safety guidance.
- Section 10 contains conclusions.
- A series of appendices provide more detail on specific aspects, including: incidents and accidents (Appendix A) in which data was a causal factor; an Organisational Data Risk assessment (Appendix B); a Data Safety Culture Questionnaire (Appendix C); Dataware Framework Assessments (Appendix D); the Data Safety Management Plan (Appendix E); definitions, acronyms & glossary entries (Appendix F); references (Appendix G); the history of the DSIWG (Appendix H); contributors (Appendix I); and acknowledgements (Appendix J).

2 Data

> *Data is a precious thing and will last longer than the systems themselves.*
> ***Tim Berners-Lee***

2.1 Purpose

This section discusses the different types of safety-related data that occur within systems. It also highlights key data properties, which can help establish what aspects of the data (e.g. timeliness, accuracy) need to be guaranteed in order that the system operate in a safe manner. These data-related facets provide a way of thinking about, and talking about, data, which supports the activities discussed later in this document. A collection of key Data Safety Principles, which have been derived from the related principles for software, are also provided.

2.2 Data Types

The full set of data types which can have safety implications is large: to date some twenty-three types (and one meta-type) have been identified; these are documented a little later in this section. In order to allow this document to be produced in a timely manner, more detailed work (e.g. consideration of appropriate methods and approaches, in Section 8) has focussed on just five data types:

1. Verification (data used to test and analyse the system);
2. Infrastructure (data used to configure, tailor or instantiate the system);
3. Performance (data collected or produced about the system during trials, pre-operational phases and live operations);
4. Dynamic (data used in the system during operations);
5. Justification (data used to justify the safety position of the system).

These five types were selected based on two considerations: they were judged as being most likely to allow progress to be made in the data safety arena; and, in general, they should be easily identifiable within safety-related systems. The intent is to gradually integrate the other data types into future editions of the document.

The table below gives the current view of the types of safety-related data that contribute to, are used by, produced by or affected by safety-related systems. They are roughly organised into a number of categories, which aim to cover all aspects of the system lifecycle.

No.	Type	Description	Explanation	Typical containers
			Context	
1	Predictive	Data used to model or predict behaviours and performance	Data for studies, models, prototypes, initial risk assessments, etc. This is the data produced during the initial concept phase which subsequently flows into further development phases.	Prototype results, evaluations, analyses, etc.
2	Scope, Assumption & Context	Data used to frame the development, operations or provide context	Restrictions, risk criteria, usage scenarios, etc. explaining how the system will be used and any limitations of use.	Concepts of Operation, Safety Case Report part 1
3	Requirements	Data used to specify what the system has to do	Data encompassing requirements, specifications, internal interface or control definitions, data formats, etc.	Formal specifications, Interface Control Documents, User Requirements documents, Safety Case Report part 1
4	Interface	Data used to enable interfaces between this system and other systems: for operations, initialisation or export from the system	Data that exists to enable exchange between this system and other external systems. Covers start-of-life operations (data import or migration), end-of-life operations and ongoing operational exchange of data between systems.	Protocols, Schemas, Interface Control Documents, Transition Plans, Extract-Transform-Load tool specifications, Cleansing and Filtering rules
5	Reference or Lookup	Data used across multiple systems with generic usage	Data comprising generic reference information sets used by multiple systems (i.e. not produced solely for this system). Typically updated infrequently, and not specific to this system.	Dictionaries, materials information, sector data reference sets, encyclopedias, etc.
			Implementation	
6	Design & Development	Data produced during development and implementation	This is data encompassing the design & development process artefacts: everything from design models and schemas to document review records. It also includes test documents (specification and results) but not the test data itself.	Design documents, Review records, Hardware, Software and design, Test scripts, Code inspection reports, etc. Safety Case Report part 2
7	Software	Data that is compiled and executed to achieve the desired system behaviour	From some perspectives it is helpful to consider software (e.g. source code) as another type of data.	Text files, configuration management systems

No.	Type	Description	Explanation	Typical containers
8	Verification	Data used to test and analyse the system	This is data comprising the test values and test data sets used to verify the system. It may include real data, modified real data or synthetic data. It includes data used to drive stubs, and any data files used by simulators or emulators.	Test data sets, Stub data, Emulator and Simulator files
			Configuration	
9	Infrastructure	Data used to configure, tailor or instantiate the system itself	Data used to set up and configure the system for a particular installation, product configuration, or network environment.	Network configuration files, Initialisation files, Hardware pin settings, Network addresses, Passwords, etc.
10	Behavioural	Data to change the functionality of the system	Data to enable / disable or configure functions or behaviour of the system.	XML configuration files, Comma Separated Variable data, schemas, etc.
11	Adaptation	Data to configure to a particular site	Data used to tailor or calibrate a system to a particular physical site or environment, incorporating physical or environmental conditions.	Configuration files
			Capability	
12	Staffing & Training	Data related to staff training, competency, certification and permits	Data which allows staff to perform a function within the wider context of the safety-related system. This may include training records, competency assessments, permits to work, etc.	Human Resources records, training certificates, card systems
			The Built System	
13	Asset	Data about the installed or deployed system and its parts, including maintenance data	Data related to location, condition and maintenance requirements of the system under consideration. This may cover hardware, software and data.	Inventory, asset and maintenance database systems
14	Performance	Data collected or produced about the system during trials, pre-operational phases and live operations	Data produced by and about the system during introduction to service and live service itself. Includes fault data and diagnostic data. This may be the results of various phases of introduction and may include trend analysis to look for long-term problems.	Field data, Support calls, Bug reports, Non-Compliance Reports, Defect Reporting and Corrective Action System data
15	Release	Data used to ensure safe operations per release instance	Explanation of particular features or limitations of a release or instance. May include specific time-limited workarounds and caveats for a release.	Release notes, Certificates of Design, Transfer documents, Safety Case Report part 2 or part 3

No.	Type	Description	Explanation	Typical containers
16	Instructional	Data used to warn, train or instruct users about the system	This is data that explains to users the risks of the systems and gives any mitigations that may be required to be implemented by users, e.g. by process, procedure, workarounds, limitations of use.	Manuals, Standard Operating Procedures, On-line help, Training courses, etc. Safety Case Report part 3
17	Evolution	Data about changes after deployment	This is data that covers enhancements, formal changes, workarounds, and maintenance issues. It also covers data produced by configuration management activities, such as baselines or branch data.	Change Requests, Modification Requests, Issue and version data, Configuration Management system outputs
18	End of Life	Data about how to stop, remove, replace or dispose of the system	This is data covering all activities related to taking the system out of service or mothballing / storage / dormant phases.	Transition, Disposal and decommissioning plans
19	Stored	Data stored by the system during operations	This is the data stored or utilised within the system which has end-user meaning. It may be displayed an used within the system or may be for transfer and distribution to other systems or downstream users. It is data that has some real domain meaning.	May be stored internally within the system (e.g. in databases or text files), or transferred into or out of the system through interfaces (e.g. Ethernet)
20	Dynamic	Data manipulated and processed by the system during operations	This is the data processed, transformed or produced by the system which has end-user meaning. It may be displayed an used within the system or may be for transfer and distribution to other systems or downstream users. It is data that has some real domain meaning.	May be manipulated within the system in data structures or transferred into or out of the system through interfaces
Compliance and Liability				
21	Standards and Regulatory	Data that governs the approaches, processes and procedures used to develop safety systems.	This is data predominantly in the form of documents that describe and dictate the activities, processes, competencies etc. to be used for a particular development in a particular sector.	Standards documents, guidelines, legal directives and laws
22	Justification	Data used to justify the safety position of the system	Data used to justify, explain and make the case for starting or continuing live operations and why they are safe enough. Often passed to external bodies (regulators, Health and Safety Executive, Independent Safety Auditors) for their review.	Safety Case report, Certification case, Regulatory documents, COTS Justification file, Design Justification file

No.	Type	Description	Explanation	Typical containers
23	Investigation	Data to support accident or incident investigations (i.e. potential evidence)	This is data collected or produced during an incident or accident investigation which may be used in investigation reports, lessons learnt or prosecutions. This can be process data, trace data, site data (e.g. photographs of crash site) or may be derived (accident simulations, analyses, etc.).	Incident/accident Investigation reports and supporting documents
Meta-Property				
+1	Trustworthiness	(Meta) data which tells us how much the system can be trusted	This is data which provides assurance or confidence about the other data within or about the system under consideration. This may be some of the data mentioned in the other types, but may be different.	Data audits, data quality index measures, sign-off sheets traceability records, model database

2.3 Data Properties

James Inge's work [1] produced a useful taxonomy of data types, and went on to look at faults in data. He concluded that a rigid taxonomy of data types was unhelpful due to various properties or characteristics of the data which vary independently. In short, it is the combination of data type with the required properties that facilitates safety analysis.

To support such analyses a collection of data properties has been produced; this is documented in the following table. Typically speaking, it is the loss of one of these properties that presents a hazard. Furthermore, this notion of "loss" is dependent on the intended use: for example, what is "timely" for one use may not be for another.

Property	Description
Integrity	the data is correct, true and unaltered
Completeness	the data has nothing missing or lost
Consistency	the data adheres to a common world view, e.g. units
Continuity	the data is continuous and regular without gaps or breaks
Format	the data is represented in a way which is readable by those that need to use it
Accuracy	the data has sufficient detail for its intended use
Resolution	the smallest difference between two adjacent values that can be represented in a data storage, display or transfer system
Traceability	the data can be linked back to its source or derivation
Timeliness	the data is as up to date as required
Verifiability	the data can be checked and its properties demonstrated to be correct
Availability	the data is accessible and usable when an authorized entity demands access
Fidelity / Representation	how well the data maps to the real world entity it is trying to model
Priority	the data is presented / transmitted / made available in the order required
Sequencing	the data is preserved in the order required
Intended Destination/Usage	the data is only sent to those that should have them
Accessibility	the data is visible only to those that should see them

Property	Description
Suppression	the data is intended never to be used again
History	the data has an audit trail of changes
Lifetime	when does the safety-related data expire
Disposability / Deletability	the data can be permanently removed when required

2.4 Data Safety Principles

2.4.1 Purpose

Hawkins *et. al.* established some generic software safety assurance principles, which are commonly referred to as "4 + 1" [2]. Given the close links between software and data it is prudent to consider these principles from a data-safety assurance perspective. The results are detailed in this section, with each principle being considered in turn.

2.4.2 Principle 1

Data safety requirements shall be defined to address the data contribution to system hazards

Data pervades active system operation, as well as the system's specification, realisation, verification, validation, certification, maintenance, and retirement. Moreover, data may be passed from one system to another; sometimes with a significant passage of time. It may be assimilated, and converted from prior uses into new uses, or simply used as is by many systems. It is stored in media whose storage integrity decays.

The system context for data safety requirements may be specific to a particular system's (or [safety] engineering process's) use of the data, or it may be generalised to a class of related systems.

Hence data safety requirements are needed for any safety-related system that interacts with data.

2.4.3 Principle 2

The intent of the data safety requirements shall be maintained throughout requirements decomposition

Data safety requirements establish the system's safety properties for data, for the system's use of data, for the management of data, and for the engineering lifecycle of both the system and its associated data. The system's requirements hierarchy must preserve the intent of the data safety requirements (and hence the system's data safety properties). Moreover, the applied engineering process for both the system's realisation and subsequent lifecycle stages shall demonstrate that the data safety properties are preserved.

2.4.4 Principle 3

Data safety requirements shall be satisfied

Evidence is required that the system satisfies all of the data safety requirements imposed on it for all anticipated operating conditions. Moreover, the data safety requirements that pertain to the data's lifecycle outside of the system shall be evidentially demonstrated prior to the system acting on such data, or else that the system is able to adequately defend against broken data safety requirements. In other words, either the data can be shown to conform to its safety properties prior to being used, or the

system can implement adequate defences and mitigations against data that does not conform to the required safety properties.

2.4.5 Principle 4

Hazardous system behaviour arising from the system's use of data shall be identified and mitigated

This is an intentionally broad statement because data is conceptual and not physical; it is the contextualised use of data that could result in a system hazard. Data Principle 1 deals with system level hazards arising from data, whereas Data Principle 4 is concerned with hazards that arise from the way the system uses its data; that is, whether the system's design and implementation introduce further hazards. An example is a ship navigation system's display of hydrographic chart data, where a wide field display results in small shallow underwater features disappearing due to image scale when it is critical that situational awareness of such hazards is maintained.

2.4.6 Principle 4+1

The confidence established in addressing the data safety principles shall be commensurate to the contribution of the data to system risk

The confidence in the evidence that demonstrates establishment of the first four Data safety Assurance Principles shall be proportionate to the contribution data has with the system hazards.

3 Managing Risks

> *Errors using inadequate data are much less than those using no data at all.*
> ***Charles Babbage***

3.1 Purpose

This section outlines a general approach to risk management which is based on ISO 31000 [3]. The purpose of this section is to demonstrate how managing the risks posed by data within safety critical systems interacts with organisational risk management processes and to introduce the high-level risk-management framework that will guide the structure of the following sections (i.e. Sections 4 to 8, inclusive). This section also outlines the links between data safety and security.

3.2 Definition and Structure

There are a number of competing definitions for "risk". This makes finding a general purpose definition, which would apply across multiple sectors, rather challenging. However, in order to achieve an acceptable level of risk, it is important to define metrics and rules which allow developers to elicit verifiable requirements. Measures to assure the performance of a system against the desired level of risk are also important. From the perspective of this document, risk is an attribute of hazards. It is interpreted as a function of the likelihood[2] of an outcome and the magnitude of the consequences.

Whilst there are a variety of different types of risk management plan, the high-level steps in ISO 31000 are indicative of most processes. As such, they have been used as a structure for the central material within this document:

1. Establishing the Context This step involves describing the system under consideration, identifying the risk appetite and scoping the required assessments. From a data safety perspective a key part of this involves completing the Organisational Data Risk (ODR) assessment and, if appropriate, conducting a Dataware Framework Assessment. This step, which should lead to a set of Data Artefacts, is discussed in more detail in Section 4.

2. Risk Identification As the title suggests, this step is concerned with identifying risks. This may be achieved by a variety of methods, including a Hazard and Operability study (HAZOP). Making reference to a generic set of risks can also be informative. This step is discussed in more detail in Section 5.

3. Risk Analysis This step involves assigning categories, or levels, of risk to the respective sources. The concept of "integrity levels" is widely used and well understood. Hence, this document advances the notion of "Data Integrity Levels". This step is discussed in more detail in Section 6.

4. Risk Evaluation This step compares the risks (categorised in the previous step) with the risk appetite (which was established in the first step). This is briefly considered in Section 7.

5. Risk Treatment This step involves resolving, mitigating, treating, avoiding or accepting risks, as deemed appropriate by the evaluation performed in the previous step. A collection of existing methods and approaches can be used for this purpose. This step is discussed in more detail in Section 8.

2 "Likelihood" is being used here in its colloquial sense, covering qualitative notions such as "common" and "rare" and quantitative notions such as the probability of occurrences per hour/day etc.

ISO 31000 recommends that there are two other parallel activities which: (1) monitor and review the risk assessment process; and (2) communicate and consult with the stakeholders about the risk assessment process. Aspects like "monitor", "review", "communicate" and "consult" are taken to be part of normal project activities; items like the ODR assessment and the Data Safety Management Plan (DSMP) are intended to provide the necessary information to complete these from a data safety perspective.

3.3 Data Safety and Security

When generating high-level processes and techniques to manage the risks posed by data, it is worthwhile understanding the difference between the safety risks posed by accidental failure to preserve data properties and the security risks posed by actors maliciously undermining the properties of data.

The relationship between safety and security, as engineering concepts, can be summarised by their relationships to cultural, developmental and aspirational properties of systems development. Culturally, embedding both safety and security into an organisation is seen as a key strategic goal for creating systems that are both safe and secure. Developmentally, safety and security are quality factors, generating transverse requirements that impact the entire system. Most importantly, at the aspirational level, both safety and security have the common goal of preventing harm from accidental and malicious interventions respectively.

For an organisation aiming to create systems that are both safe and secure, these connections can be both a benefit and a burden. The shared goal of preventing harm means that both quality factors seek to identify routes to harm through analysis of the system being developed. This can result in shared processes and tools, which in turn can save time and money during systems development. However, safety and security interact in a more volatile way at the functional level. Security failings can undermine the safety case for a system and, conversely, safety requirements can prevent the implementation of standard security solutions. For example, the German government government published a report in 2014 into a fire at a steel works caused by a cyber attack that resulted in the control system being placed into an unsafe state and the safety system being unable to intervene (Section 3.3.1 of [4] - in German). In addition, "fail-safe" states can often leave a system with exposed security vulnerabilities.

These links between safety and security infer that there are connections between the sub-categories of data safety and information security: both attempt to take a data-centric view of the system of interest in order to improve the associated quality factor; and both attempt to prevent harm through the preservation of the properties of data within that system.

In the security domain, the three key properties of data considered are confidentiality, integrity, and availability. Confidentiality, (the failure of which is termed "Information Disclosure" in the Microsoft Security Model) is typically not a safety concern as, without malicious intent, information sharing is not inherently unsafe. However, when considering systems where confidentiality is an important property, the interaction between data safety and security cannot be trivially resolved. For example, accidental disclosure of information can form part of a causal chain which leads to harm from a malicious actor.

Data integrity is a critical property for both domains. The Microsoft Security Model describes malicious removal of the property of integrity as "tampering". Whether by accident or through malicious intent, the potential harm from loss of data integrity can be disastrous to a safety critical system, from the values of drug dosages to control system parameters. Data availability is also important to both domains. Loss of availability, or "denial of service" in the Microsoft security model, is another property that can be lost accidentally or through malicious intervention. Loss of availability prevents systems from functioning properly and can result in undefined behaviour if not mitigated by design.

4 Establish Context

> *It is a capital mistake to theorise before one has data.*
> ***Sherlock Holmes - "A Study in Scarlet" (Arthur Conan Doyle)***

4.1 Purpose

This section is concerned with establishing the context within which system development, enhancement, introduction or integration is occurring. This should establish the risk appetite (essentially, how much effort is devoted to making risks as low as practicable). In turn, this will inform the nature and scope of assessments that are conducted during system development and, furthermore, its introduction into operational service.

4.2 Organisational Data Risk Assessment Form

The **Organisational Data Risk (ODR) Assessment Form** was generated to capture a high-level perspective on the risk posed to an organisation by data safety issues within a specific project. How it integrates with an organisation's existing risk management policies is the responsibility of the implementing organisation. To facilitate this integration, the following paragraphs describe the connections between the ODR and the ISO 31000 standard for risk management. The ODR itself is presented in Appendix B.

Establishing the context of a risk assessment ensures that the system being considered and the scope of any assessment is well defined. This helps prevent an overrun of the assessment's boundaries and allows those items that are out of scope to be explicitly communicated to all stakeholders. In addition, it is the role of this activity to produce the risk criteria that a system will be judged on. The ODR assessment links directly to the sub-tasks identified by ISO 31000 for establishing the risk assessment context and introduces aspects to guide the assessor into focussing on data-specific risks.

Questions 2, 3 and 4 of the ODR align directly with establishing the external context of the risk assessment (Activity 5.3.2 from ISO 31000). They guide the assessor into judging the risk tolerance of external stakeholders, the level of risk that is allocated to the organisation and the regulatory environment within the project will operate.

Question 5 is concerned with establishing the internal context of the risk assessment (Activity 5.3.3), inviting the assessor to comment on the maturity of the organisation in terms of their attitude to not simply risk, but specifically data-driven risks.

Question 6 explores data ownership through the use cases of the system. This is related to the legal frameworks explored in Question 4, but also acts to lay the foundations of Activity 5.3.5, "Defining Risk Criteria", which requires an assessor to identify "the nature and types of causes and consequences that can occur and how they will be measured". This is expanded upon by Questions 1, 7 and 8 which go into data-driven specifics about failure consequences and the issues raised by data complexity, boundary complexity and system complexity for the project.

Finally, the scoring system of the ODR provides a heuristic for defining the risk criteria (Activity 5.3.5) which handles how to combine these different aspects of risk into a single, high-level estimate of the data-related risks associated with a given project. This means that the ODR can, for example, provide some guidance on the "4 + 1" Data Safety Principle; that is, it provides some guidance on the amount of effort that should be directed towards explicit consideration of data safety issues.

It is of note that whilst the completion of an ODR fits within the context establishment activity it also augments the ongoing "Communication and Consultation" activity both by providing a standardised format for capturing the relevant information and by providing a standardised reporting template in order to secure endorsement for the plan.

4.3 Data Safety Culture Questionnaire

Part of the ODR assessment relates to assessing the organization's maturity in managing data safety risks; responses are aimed at establishing the depth of awareness of data safety and the associated management processes within the organization. However, measuring the level of awareness of processes and concepts in an organization is not always easy. There may be sufficient high-level knowledge of this for the purposes of the ODR but it is an area that may warrant further exploration.

A separate questionnaire has thus been developed to explore the specific area of measuring the **data safety culture** for a particular activity; whether this be for the organization as a whole or for a particular project, service or activity. However, here the focus is on a personal view rather than a project or company's view so the questionnaire would be completed by all, or a significant subset of, staff. Responses could then be aggregated to give an overall data safety culture value. A key aspect of this approach is that it can be periodically repeated to determine trends – e.g. if overall scores are declining, this may suggest that further training and briefings will be required.

More details on the data safety culture questionnaire are provided in Appendix C.

4.4 Dataware Framework Assessment Form

Where multiple interconnected systems share data across their interfaces a means to describe these relationships is required. This can be achieved by populating the **Dataware Framework** as the basis to combine and represent a system wide risk model.

A **Dataware Framework Assessment** is a more structured and detailed analysis of the safety risks as it applies to the organisation. This assessment approach provides a framework to better identify and document hazards at the respective boundaries (typically interfaces) in the use of data. This safety analysis is then used to inform a better understanding of key areas of concern (e.g. specific data exchanges or processing in the scope of supply or operational system) and to help set the direction for further in-depth analysis to develop and implement appropriate mitigation strategies.

At the core of the framework is a layered model [5], [6] similar to and sharing concepts with the 'Basic Reference Model for Open Systems Interconnection' (ISO OSI) model[3]. In this model the enterprise is partitioned and represented as communications services between 7 layers with defined interfaces, peer protocols that permit the separation of application development from the underlying communication system. Two important elements of the OSI model are that each layer:

- communicates with its peer layer in a different communication unit; and
- provides a service to the layer above and expects a particular kind of service from the layer below.

3 The ISO Model is a collection of documents describing the services provided by the model; The book "Communications Network Protocols", 3rd Edition, Appendix B (Selected ISO Standards), by Brian Marsden provides a comprehensive list of these documents; ISBN 91-44-23043-5; Chartwell-Bratt Ltd 1992.

Structuring in this way allows the interfaces between layers (both vertical and horizontal) to be more clearly defined and therefore the risks associated with data transit across those interfaces to be better understood and articulated in a consistent manner.

Interfaces can then be controlled with a level of rigour commensurate with the assessed risks. For example, a formal interface contract could be established between an external producer of data across an interface. As well as defining the expected field types, cardinality, optionality (and so on) the contract could also define the level of assurance required for particular data properties.

Further details of this assessment method and a structure for capturing the assessment details are given in Appendix D.

4.5 System Specifics

The previous sub-sections have provided a general discussion of the processes used to establish the context within which data safety risks can be identified, analysed, evaluated and treated. The following sub-sections consider specific parts of the context in a bit more detail.

4.5.1 System Definition

The system under consideration should be understood and documented, including interfaces and safety-related data aspects. The process of documenting the system of interest not only furthers the understanding of the reviewer so they can make sensible judgements about the system, but also both formally declares assumptions that the reviewer is making whilst assessing the system and clearly defines the limits of the assessment. In addition, different levels of risk may be associated with composites or groups of safety-related data where it is easier to manage (or where independence cannot be demonstrated or maintained), so the partitioning of datasets should also be considered and modelled at this stage.

4.5.2 Usage Scenarios

If safety-related data is incorrect it can become dangerous when used, either by making a computer or control system perform incorrect actions, or by misleading human users into making incorrect decisions. Because the danger can only be determined when the usage of the data is understood, risk assessment should involve the consumer of the data as well as the producer.

The **consumer** assesses the use and possible mitigations for the safety-related data, and uses this information to define **data integrity requirements** (e.g. how accurate and reliable the safety-related data must be).

The **producer** investigates how the safety-related data is collected and what errors might occur, and guarantees (or provides a specified level of confidence) that the safety-related data meets the data integrity requirements.

In some cases a producer will be providing safety-related data without any knowledge of a specific user (e.g. mapping data or generic databases that are sold to many users). In these cases the producer will need to make some assumptions about possible users, and then clearly state what level of integrity the data has been produced to. It is then up to the users to check whether the declared integrity matches their need.

4.5.3 Data in System Lifecycles

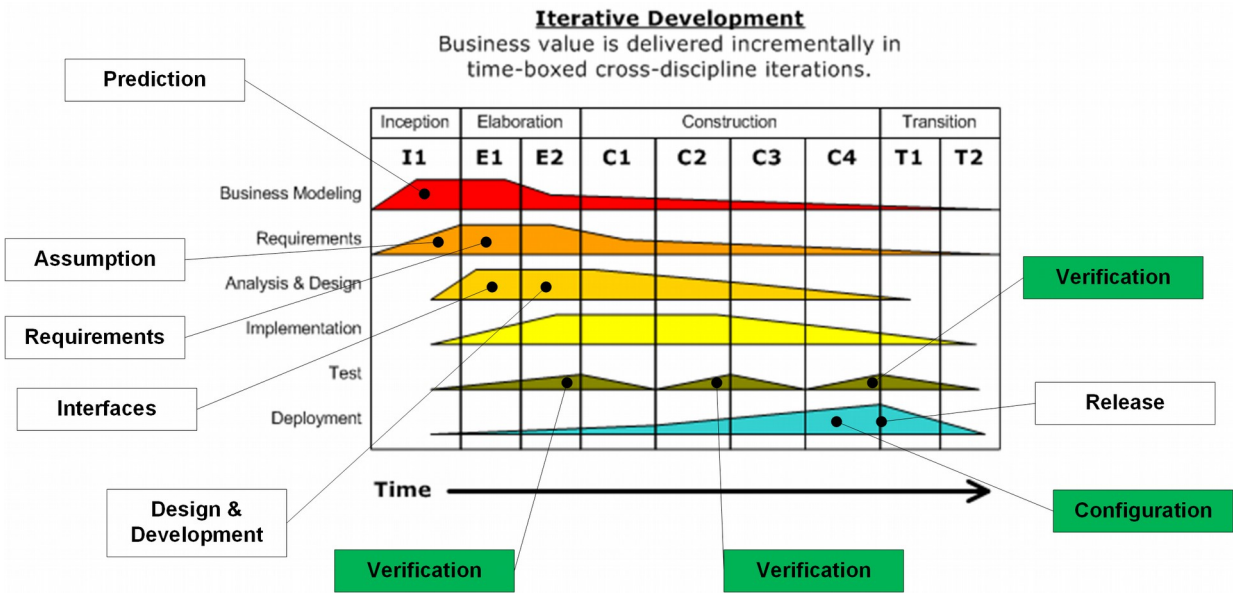
Like other components of a safety-related system, the safety dependency of data is dictated by the context in which it is used and the causal links that become established where loss of one or more of the required properties can contribute to hazardous system states. For example, a given data set (say configuration data) could be used in a number of separate contexts such as:

- prototyping a system to demonstrate solution feasibility of a safety-related system;
- development testing of a safety-related system; or
- live operational use of a safety-related system.

In these cases, the data set is the same but the context of its use changes the safety significance and therefore the level of assurance that it may require. Hence, not only is the type of data under consideration important but also *when* in the organisation's particular process lifecycle the data will be used and depended on. It follows that the assessed integrity level of a data set is also predicated on where and when in the lifecycle the data set will be applied. It is recommended these considerations are addressed in the **Data Safety Management Plan**[4] by modelling the organisation's lifecycles and explicitly documenting where a specific data set will be used and therefore subject to further assurance techniques. To illustrate this concept, a number of generic model lifecycles are discussed below. Note that these are not intended to be prescriptive or mandate the use of any particular model. Instead, they are being used to illustrate how the Data Safety Management Plan could articulate these lifecycle considerations.

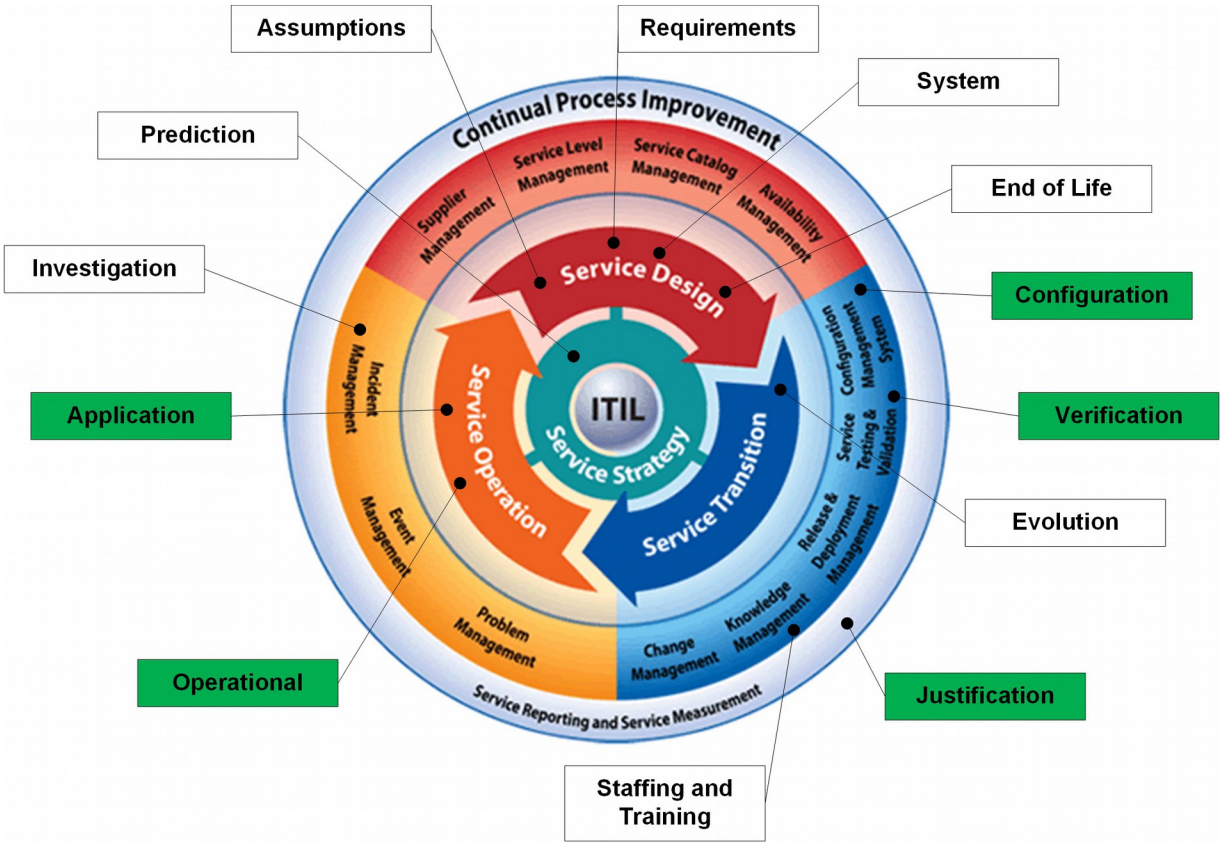
Development The following diagram represents a typical development lifecycle using an iterative development approach. In this model there are key phases as the system transitions from concept through to testable executable code. The process is iterative in that several cycles of functional elaboration, design, development and test may be run and these typically will focus on the areas of the system that bear most technical risk or comprise the key functional use cases so the client gets early visibility of the system. This early awareness allows feedback to be provided into the next iteration to help steer the solution to the client's actual needs. Traditional waterfall implementation can map onto this model on the basis that there is only one iteration in each phase and all activities in one phase need to be completed before progressing to the next.

4 The Data Safety Management Plan is discussed in Appendix E.



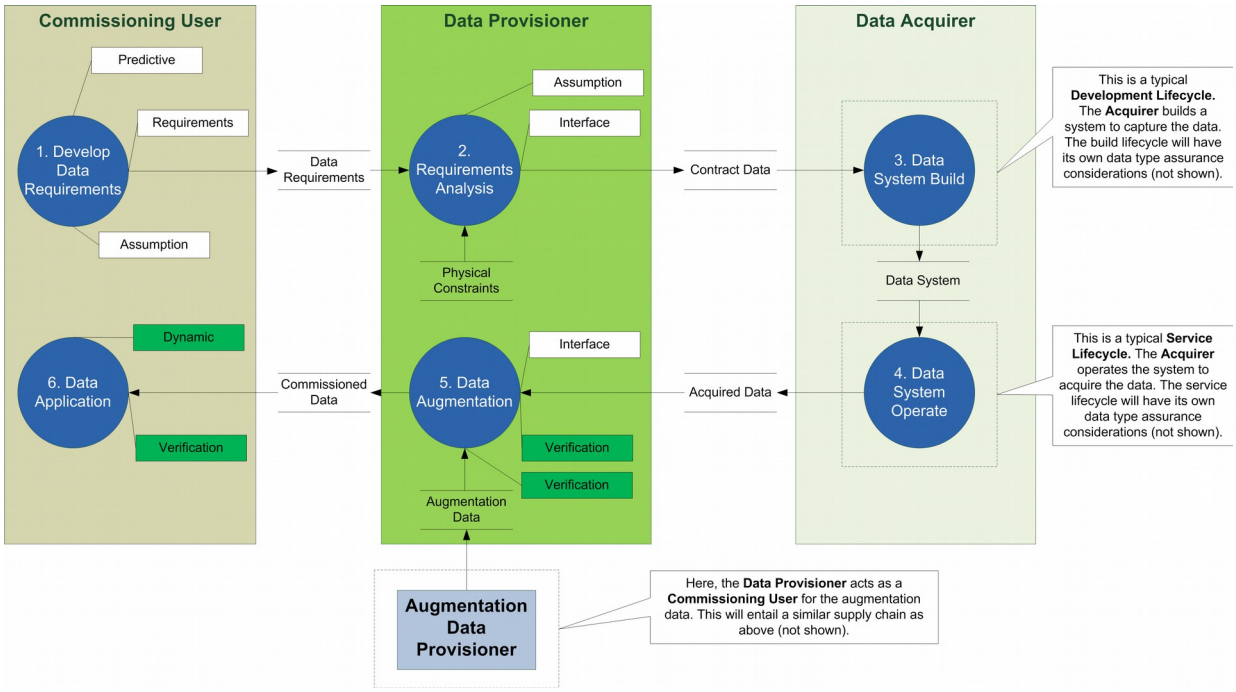
The model itself may vary depending on the specific needs of the project but the diagram illustrates that different data types become significant at different points of the process. It is therefore important that these considerations are explored and documented in the Data Safety Management Plan.

Operational Once a system has been developed it will move into an operational style of lifecycle or indeed, if the data safety has not previously been considered for an enterprise, then the system will already be in operational use. These operational lifecycles tend to be cyclical in nature and the following diagram illustrates a typical model.



Again, specific data types will come into play at different periods in the process. Documenting the relationship between process steps and data types will therefore give clarity as to when a particular assurance technique needs to be applied.

Data supply chains The previous models relate to typical system supply and operate perspectives but there are also other data supply chains where a number of organisations engage in the procurement and use of safety-related data. These processes may include the development and operational lifecycles but a different model is required to fully represent the wider processes that are being employed. The following diagram shows such a model representing a data acquisition lifecycle:



This model represents the interactions between three key organisations:

- The Commissioning User: the organisation that has the need for the data;
- The Data Provisioner: the organisation that will fulfil that need for data;
- The Data Acquirer: the organisation employed by the Data Provisioner to carry out physical collection of data.

In this supply chain, the Commissioning User is a **Consumer** of the data and the Data Acquirer is a **Producer** of data. The Data Provisioner acts both as a **Consumer** (from the Data Acquirer) and **Producer** (to the Commissioning User) of data. Similarly, an organisation that augments data sets is both a **Consumer** and **Producer** of data in the supply chain.

The Commissioning User Requirement Analysis is the key process step where the Commissioning Users expectations for data (including fidelity levels for associated properties) are agreed with the Data Provisioner. The requirements may be adjusted because of physical constraints (e.g. loss of precision because of physical measuring device constraints) and may include additional requirements to augment the captured data with additional information (e.g. airport codes added to a measurement of a given runway length).

The Data Provisioner will employ a Data Acquirer to capture the data, for example, to carry out a physical survey of a site. The acquisition phase may itself require a specialised system to be built to perform the

capture and data refinement to meet the Data Provisioner's specifications. Such systems will then themselves be subject to the Development lifecycle model considerations discussed above. Likewise, the data augmentation phase may require further system development processes or indeed, could trigger an instance of the model again as the Data Provisioner acting as a Commissioning User.

Acquired and augmented data is then fed into the operational system that has been built for providing the service of generating the commissioned data. This system in its service provision role would then typically follows an Operational Lifecycle process model discussed earlier.

Again, it should be stressed that these models are not intended to be prescriptive but rather there to illustrative how authors of the Data Management Plan might address the question of *when* in the process data assurance techniques should be applied for a given data type.

5 Identify Risks

> *I wanted to separate data from programs, because data and instructions are very different.*
> ***Ken Thompson***

5.1 Purpose

This section outlines several methods that can be used to help identify data-related risks. These include an outline top-down / bottom-up approach, a collection of generic data safety issues and a set of guidewords that can be used to support a Hazard and Operability (HAZOP) study.

5.2 Top-Down / Bottom-Up

ISO 31000 uses the term "risk identification". These risks are closely related to the concept of a hazard, which is a generic source of risk of accidental harm. If a hazard (or a risk) is not identified then no action will be taken to manage it, so hazard (or risk) identification must be rigorous and systematic (for example, using a HAZOP approach). Consideration of data safety as an integral part of a holistic system safety assessment is encouraged. Two possible approaches are outlined below:

- Function Based (Top-Down) If the consumer is a system which has clearly identified safety functions, then data can be assessed by considering each function in turn and analysing what data the function depends on. If there are a limited number of safety functions, this is usually the simplest approach.
- Data Based (Bottom-Up) A more general approach is to consider each different category of data and explore the effect of possible data errors. The errors which are most likely to have safety impacts are those which affect the safety-related data properties as described in Section 2. Data-based analysis has the advantage that it may identify safety implications that are currently unknown, but it can require considerable effort for systems with large data sets.

Although it is something of a simplification, the top-down approach has some similarities with the first three Data Safety Principles, whilst the bottom-up approach is related to the fourth Data Safety principle[5].

5.3 Generic Data Safety Issues

There are some issues related to data which are different to any other element in a system (for example, hardware and software); there are others which, because of increasing prevalence, are now presenting a greater risk. Seven specific issues are worth highlighting: the fluidity of data; reuse of data; ageing of data; aggregation of data sets; and archiving / retrieval.

- Fluidity Hardware and software can undergo significant amounts of product assurance and once assured do not change frequently. Where change is required to hardware or software, it can be carefully managed and the impact on the safety case appraised. This is not always the case for data, which is often much more fluid; indeed the ease with which data can be changed is one motivation for the move towards data-driven systems. This fluidity means that it is not always practicable to revisit safety cases when data changes. Instead, the data fluidity, along with any associated safety impacts, needs to be captured in the system safety case.

5 The Data Safety Principles are discussed in Section 2.

- Reuse For the purposes of this discussion, "reuse" is interpreted as use of the same data in a different system or system context. Just because data was valid for use in a particular system, it does not immediately follow that it can be reused again in a similar system. Many considerations associated with data reuse are similar to those of software reuse, for example: similarity of requirements; similarity of role in system; and similarity in required integrity level. One consideration that is different is that of timeliness: data that was valid for use in a particular system at a particular time is not necessarily valid for reuse in exactly the same system at a different time. This aspect is related to data fluidity (discussed above) and data ageing (discussed below).

- Ageing As highlighted above, all safety-related data has a lifetime and this needs to be explicitly managed. This can involve, for example, purging, deletion and alerting. It is also important to note that ageing can occur as a result of changes external to the system (e.g. the positions of other aircraft) or it can result from internal changes (e.g. in sensors monitoring system properties).

- Aggregation Aggregation is interpreted as the combination of data from several disparate sources into a new data set. Data properties (like those discussed in Section 2) are not necessarily preserved throughout the combination process. Furthermore, aggregation relies on a good understanding of data attributes: a data set using kilometres *can* be combined with one using miles, but if the difference in units is not catered for then it may be difficult to detect the resulting error. The main issues are loss of heritage/history/source; data can also appear to become something else. There could also be problems of "false" aggregation, i.e. joining or linking data which appears to be strongly related but in fact is not (e.g. data collected on different dates or by different methods). The issues for safety are that the integrity may become lowered to lowest common denominator, and this needs to be recognised. Additional checks (e.g. validation checks, sanity checks) or assurance measures may need to be put in place to ensure that overall integrity is maintained.

- Archiving and Retrieval Safety-related data needs to be available when required. There is thus a need to think about data accessibility over the complete system lifetime. It is also important to consider what properties of the data need to be preserved and how this affects the choice of storage medium.

- Biasing This is a systemic inaccuracy in data due to the characteristics of the process employed in the creation, collection, manipulation, presentation and interpretation of data. Hence this is usually unintentional distortion in the data set, which may be due to how the set has been selected or originated. Again there is no perfect way of checking for this within the system, although completeness, statistical and validity checks on data sets may help.

- Aliasing This is an effect that causes different data to become indistinguishable when accessed. This could be due to the way the data is filtered, sampled, indexed, stored or retrieved. The data issues are typically related to loss of resolution leading to similar data points appearing to be identical.

5.4 HAZOP Guidewords

A HAZOP provides a common approach for identifying hazards. It involves a multidisciplinary team collaborating to identify potential hazards and operability problems. Structure and completeness are supported through the use of *guideword* prompts for example, considering the implications if software components perform functions early, late or not at all. These prompts are intended to stimulate imaginative thinking, to focus the study and to elicit ideas and discussion.

The following table provides a set of guidewords for a data-focused HAZOP.

Property	Description	HAZOP Data Properties	HAZOP Data Guidewords
Integrity	the data is correct, true and unaltered	Loss, partial loss, incorrect, multiple	Correctness, truth, original, trustworthy, coherency, stability, perfect, unquestionable, faithful, certain, ordered, unadulterated, unmodified, unchanged, clean, uncontaminated, untainted, proper, flawless, organized, exact, undistorted, faultless, guided, connected, linked, traced, unbiased.
Completeness	the data has nothing missing or lost	Loss, partial loss, incorrect, multiple	Whole, complete, entire, finished, done, stable, qualified, certified.
Consistency	the data adheres to a common world view, e.g. units	Loss, partial loss, incorrect, multiple, too early, too late, loss of sequence	Coherent, compatible, congruent, congruous, harmonious, deconflicted, consistent, appropriate, suitable, sound, cleansed.
Continuity	the data is continuous and regular without gaps or breaks	Loss, partial loss, incorrect, late, loss of sequence	Smooth, continuous, regular, gapless, whole, complete, entire, unfragmented.
Format	the data is represented in a way which is readable by those that need to use it	Loss, partial loss, incorrect, multiple	Conformant, suitable, valid, configured, well-formed, setup, composed, well structured, arranged, compliant, organised, exact, unaliased, migrated, transformed.
Accuracy	the data has sufficient detail for its intended use	Loss, partial loss, incorrect, multiple	Accurate, true, correct, undistorted, unbiased, faultless.
Resolution	the smallest difference between two adjacent values that can be represented in a data storage, display or transfer system	Loss, partial loss, incorrect, multiple	Exact, untruncated, retention of detail, clarity, determination, distinguishable, clear, within range, distinct, separated, discernible, discriminatable, unconfused, divisible, unaliased, granularity, precision.
Traceability	the data can be linked back to its source or derivation	Loss, partial loss, incorrect, multiple, too early, too late, loss of sequence	Traceable, verifiable, indexed, linked, connected, justified, proven, evidenced, substantiated, continuous, unfragmented, complete, networked.
Timeliness	the data is as up to date as required	Loss, partial loss	Timely, early, ready, expected, unique, appropriate, opportune, ordered, organised, anticipated, seasonable, converging, settling, on-time, latency, lag, lead time, time slots, real-time, determinism, predictable.
Verifiability	the data can be checked and its properties demonstrated to be correct	Loss, incorrect, partial loss, multiple, too early, too late, loss of sequence	Verifiable, provable, checkable, supportable, demonstrable, sustainable, certifiable, defensible, excusable, justifiable, undisputable, irrefutable, validated.

Property	Description	HAZOP Data Properties	HAZOP Data Guidewords
Availability	the data is accessible and usable when an authorized entity demands access	Loss, partial loss, multiple, too early, too late	Ready, available, obtainable, reachable, accessible, serviceable, operable, functional, usable, capable, released, issued, disseminated, distributed.
Fidelity / Representation	how well the data maps to the real world entity it is trying to model	Loss, incorrect, partial loss, multiple, too early, too late	Representative, accurate, faithful, trustworthy, characteristic, normal, standard, real, expected, natural, typical, regular, fit for purpose, validated, separable, associated, correct units/dimensions, stable, unbiased.
Priority	the data is presented / transmitted / made available in the order required	Loss, incorrect, partial loss, multiple, too early, too late	Current, ordered, included, precedence, hierarchy, pre-eminence, retained, ahead, readiness.
Sequencing	the data is preserved in the order required	Loss, incorrect, partial loss, multiple	Ordered, contiguous, unique, ordered, clear, continuous, successive, uninterrupted, sequential.
Intended Destination / Usage	the data is only sent to those that should have them	Loss, incorrect, partial loss, multiple, too early, too late, loss of sequence	Directed, delivered, copied, sent, transmitted, correct recipient, unintercepted, unseen, integral, received, acknowledged, forwarded, filtered.
Accessibility	the data is visible only to those that should see them	Loss, incorrect, partial loss, multiple, too early, too late	Secure, open, visible, reachable, seen, usable, accessible, obtainable, uncompromised, secure, encrypted, preserved.
Suppression	the data is intended never to be used again	Loss, incorrect, partial, too early, too late, too much, too little	Hidden, encrypted, private, confidential, erased, unlinked, unavailable, unaccessible, redacted.
History	the data has an audit trail of changes	Loss, incorrect, partial loss, multiple	Justifiable, traceable, provable, supportable, demonstrable, sustainable, certifiable, defensible, excusable, justifiable, undisputable, irrefutable.
Lifetime	when does the safety-related data expire	Loss, too early, too late, incorrect, multiple, loss of sequence	Expiry date, age, validity, currency, applicability, durability, duration, lifespan, stretch, tenure, half-life, longevity, span, in-date, best-before, window, established.
Disposability / Deletability	the data can be permanently removed when required	Loss, incorrect, partial, too early, too late	Unavailable, unaccessible, redacted, hidden, filtered, lost, deleted, destroyed, backup, archive, locked, secured, unlinked.

6 Analyse Risks

> *I love data. I think it's very important to get it right, and I think it's good to question it.*
> ***Mary Meeker***

6.1 Purpose

This section presents the concept of Data Integrity Levels. The background to their development is outlined and a possible classification scheme is defined.

6.2 Identifying Integrity Requirements

In order to analyse risks and, more particularly, to align data safety with other risk management processes, there is a need to overcome problems stemming from the use of term "likelihood" in a situations where there may be no failure rates. For this reason the Data Integrity Level (DIL) was developed. The DIL is a heuristic metric which allows developers to classify safety-related data sets in terms of the risk they pose. From here the focus becomes not one a statistical measure of likelihood, but on the way in which an assurance argument is built up for the safety-related data meeting the requirements of the system in question. As such, DILs share a common theoretical basis with concepts like System Integrity Levels and (Item / Function) Development Assurance Levels.

The table below presents an initial classification system, allocating DILs to safety-related data items using a function of likelihood and consequences. Those using the system described below are encouraged to tailor it so that it meets their own needs; the reasons for any such tailoring should, of course, be documented and agreed amongst relevant stakeholders. Furthermore, where this table suggests a low DIL because (for example) a work around is simple to implement it is important to ensure that the work around (or similar) is actually implemented.

To make the analysis applicable to safety-related data, the "consequences" are subdivided further into five categories which impact on the level of risk:

- Proximity: how directly a data failure will lead to an accident;
- Dependency: how dependent the application is on the dataset;
- Detection: the likelihood of being able to detect a data failure prior to an accident;
- Prevention: the ability of the systems architect/developers to guard against errors;
- Correction: the ability of the system to work around or correct errors.

		Likelihood of Data Causing Accident		
	Concern	**High**	**Medium**	**Low**
	Proximity	A known use of the data[6] is highly likely to lead to an accident.	A possible use of the data could lead to an accident.	All currently foreseen uses of the data could lead to harm only via lengthy and indirect routes.
	Dependency	Data is completely relied upon.	Data is indirectly relied upon.	Little reliance on data.
	Detection	Low or no chance of anything else detecting an error.	Some other people/systems are involved in checking the data.	Many other people/systems are involved in checking the data.
	Prevention	Difficult or impossible to guard/barrier against errors.	Possible to guard/barrier against errors.	Easy to guard/barrier against error.
	Correction	Difficult or impossible to correct or workaround errors.	Possible to correct or workaround errors.	Easy to correct or workaround errors.
Severity or Impact of data related accident				
Negligible	Negligible harm. Negligible environmental impact.	DIL0	DIL0	DIL0
Minor	Minor injury or temporary discomfort for 1 or 2 people. Minor environmental impact.	DIL1	DIL0	DIL0
Moderate	An accident resulting in minor injuries affecting several people or one serious injury. Some environmental impact.	DIL2	DIL1	DIL1
Major	A serious accident resulting in serious injuries affecting a number of people, or a single death. Major environmental impact.	DIL3	DIL3	DIL2
Catastrophic	An accident resulting in several deaths. The accident could affect the general public or have wide and catastrophic environmental impact.	DIL4	DIL4	DIL3

6 Where the table refers to "data", this should be read as safety-related data.

There are a number of instances where the system architecture could justify the movement of a dataset from one DIL for another. For example there may be cases where multiple independent data items fulfil the same (or similar) usage and assessors may choose to reduce the DIL requirements on each item to reflect the inherent redundancy (and associated risk reduction) that brings. Additionally, the same dataset may be reused by multiple functions with different levels of risk, in this case it would be recommended to assign the highest required level of integrity to the dataset so that it meets the minimum requirements of the most demanding use case. As discussed, the implementation of a classification system is down to the organisation, but any such manipulation of DILs should be carefully considered and appropriately documented.

7 Evaluate Risks

> *In any collection of data, the figure most obviously correct, beyond all need of checking, is the mistake.*
> ***Finagle's Third Law***

7.1 Purpose

As the title suggests, this section is about evaluating the risks. This involves comparing the risks that have been identified with the risk appetite.

The focus on data that is advocated throughout this document does not introduce any special requirements for this activity; the approach espoused in ISO 31000 is entirely adequate.

8 Treat Risks

> *Data that is loved tends to survive.*
> ***Kurt Bollacker***

8.1 Purpose

This section provides guidance on how data-related risks may be treated. This is achieved by cross-referencing various methods and approaches against different types of data and against different Data Integrity Levels. By virtue of the way they were constructed, these tables are neither complete nor authoritative. Despite that, approaches to data that are widely at odds with their contents are likely to need detailed and compelling justification.

8.2 Methods and Approaches Tables

The following tables provide a wide spectrum of methods and techniques that represent best practice in mitigating the risks associated with safety-related data. The tables indicate where a particular method/approach is applicable to a given lifecycle data type, and for each Data Integrity Level, whether the method/technique is:

- No recommendation for or against being used (-);
- Recommended (R);
- Highly Recommended (HR).

The lifecycle data types are those under consideration for this version of the guidance (see Section 2) are as follows:

- Verification (V);
- Infrastructure (I);
- Dynamic (D);
- Performance (P);
- Justification (J).

The methods/approaches are not intended to be prescriptive, but they should be sufficiently well-defined to allow interpretations to be applied to the given context in which the guidance will apply. Methods/techniques employed are expected to be more rigourously applied as the DIL level increases. For example, the depth, level of coverage and effort/resources employed for analysis techniques must be proportionate to the DIL - sampling may be appropriate for lower DILs where full coverage will likely be expected for higher DILs. Assurance methods and approaches must be considered for each stage of the data life cycle as appropriate for the given DIL. Strategies for dealing with large data sets must be fully justified with respect to the DIL.

The Data Safety Management Plan can be used to document:

- planned compliance with the tables;
- the interpretation for the given method/technique (e.g. depth of checking);

- justification in the case where a technique is not to be adopted.

The overall safety justification for the given project/service/operational context must then provide evidence of compliance against the plan.

It is important to note that the following tables are incomplete. As well as only considering certain data types (as highlighted above) each table only considers one particular aspect of data safety. No claim is made that the collection of tables provides complete coverage either across the system lifecycle or across all possible approaches to one part of the lifecycle. Also, these tables merely identify techniques using a few key words; in almost all cases, further information on the technique should be readily available from a range of sources. Despite these limitations, it is hoped, however, that the guidance they contain will prompt considerations that lead to the use, and justification for the use, of an appropriate set of methods and approaches for any given project (or system).

8.2.1 System Design

Methods and Approaches - System Design	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Built-in-Test / Built-in-Test Equipment (BIT/BITE)	-	-	Y	-	-	-	R	HR	HR	
Cyclic / Continuous BIT	-	-	Y	-	-	-	-	R	HR	Application applies tests to the data it is processing continuously (e.g. for a live data stream) or periodically according to a periodicity strategy (e.g. every nth message, every hour etc).
Backward recovery	-	-	Y	-	-	R	R	HR	HR	If a fault in data has been detected, the system resets to an earlier internal data set, which has been proven consistent.
Parity Checks	-	-	Y	-	-	R	R	HR	HR	Within data, e.g. Hamming codes, Reed-Solomon, also Hagelbarger.
Automatic Error Correction	-	-	Y	-	-	R	R	HR	HR	Detected errors are corrected automatically.
Checksums / Cyclic Redundancy Checks (CRCs) / Hashes	-	-	Y	-	-	-	R	HR	HR	Digests of datasets are produced, included with the dataset and checked to provide confidence that the data is unaltered.
Digital Signatures	-	-	Y	-	-	-	R	HR	HR	For non-repudiation and integrity of data.
Sequence Numbers	-	-	Y	-	-	R	R	HR	HR	Data bears sequence numbers so the integrity of a data stream can be checked (e.g. data items not monotonically increasing, duplicate detection).

Methods and Approaches - System Design	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Automatic Repeat Request	-	-	Y	-	-	R	R	HR	HR	Automatic Repeat-reQuest (ARQ) to repeat transmission of data which has not been received correctly.
Auditing Facilities	-	-	Y	Y	-	-	R	HR	HR	Changes to data properties are audited so the before and after values are recorded and also other relate information such as the author (e.g. person, function or system) and the time of the change.
Logging Facilities	-	-	Y	Y	-	R	R	HR	HR	Data processing events are logged to allow support staff to monitor the health of the system and provide diagnostic information if problems are detected.
Encapsulation	-	-	Y	-	-	R	R	HR	HR	The hiding of data so that it is only accessible through well defined interfaces.
Multiple Stores	-	-	Y	-	-	-	-	R	HR	The same instance of a data set or data items is stored in multiple locations.
Homogeneous Redundancy	-	-	Y	-	-	-	-	R	HR	Data is processed using homogeneous redundant channels; detected faults in data of one channel will cause processing to switch to another channel.
Heterogeneous Redundancy	-	-	Y	-	-	-	-	R	HR	Data is processed using heterogeneous redundant channels (same functionality but different implementations) detected faults in data of one channel will cause processing to switch to another channel.
N-Version Programming	-	-	Y	-	-	-	-	R	R	Data is processed using heterogeneous redundant channels (same functionality but different implementations) with both channels active and a form of voting across channels to determine data output or control behaviour.
Data Integrity Sampling	-	-	Y	-	-	HR	HR	R	R	The integrity of subsets of data is periodically checked, in accordance with a given selection criteria (eg. random, critical records etc), frequency and volume of data to check.

Methods and Approaches - System Design	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Sanity/Reasonability Checks	-	-	Y	-	-	R	R	HR	HR	Dedicated processing implemented to check that data is within reasonable tolerances and/or logically/semantically consistent with what the data represents. For example, range checks, date checks, record counts, record sizes, special values (e.g. NaN) etc.
Data Correlation	-	-	Y	-	-	R	R	HR	HR	Data from a number of sources exists to permit a cross-correlation of the data supplied from one source (the master, or prime source) with other sources.
Data Partitioning	-	-	Y	-	-	R	R	HR	HR	To separate data that is managed differently, creating independence between data so that whole data set do not require validation after a change.
Syntax Checks	Y	Y	Y	-	-	R	R	HR	HR	Semantic checking of data values and sequences based on defined rule sets.
Database Management System (DBMS)	-	-	Y	-	-	HR	HR	R	R	Use of established 3rd Party products for storage and management of data.
Feedback testing	-	-	Y	-	-	HR	HR	R	R	To check output data by comparing it with the input source.
Information Redundancy	-	-	Y	-	-	HR	HR	R	R	Additional redundant information is supplied from diverse sources. The validity of the data coming from the diverse sources can be checked against each other.
Reverse Translation	-	-	Y	-	-	-	R	HR	HR	To verify that the data output of a process is correct, by attempting to create the source data from the output data and comparing this with the source used to create the output data.
Meta-data	-	-	Y	Y	-	-	R	HR	HR	Auditable data are sent with the data that is about the data, e.g. source, issue state, expiry date.

8.2.2 Data Assurance

Methods and Approaches - Data Assurance	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Review / Inspection	Y	Y	Y	Y	Y	HR	HR	HR	HR	Manual review/inspection of data possibly involving data visualisation tools.
Sanity Check	Y	Y	Y	Y	Y	R	R	R	R	Informal check of data, eg. total record counts, filesizes within expectations.
Statistics-Based Sampling	Y	Y	Y	Y	Y	-	R	HR	HR	More appropriate for realtime large and/or volume data, not necessary if inspection gives full coverage but gives extra defence in depth. Could be manual selection, a form of random selection or comparison against statistical norms.
Ground-Truth Check	Y	Y	Y	Y	Y	R	R	HR	HR	Inspection against physical measurements (eg. lengths, positions, heights) taken in the real world.
Auditing	Y	Y	Y	Y	Y	R	R	HR	HR	A period of comprehensive internal and external testing of the data quality process, where Data is verified according to its intended use and definition.
Tracing	Y	Y	Y	Y	Y	-	R	HR	HR	Ability to trace data from source across multiple participants in the data supply chain.
Defined Verification Frequency	Y	Y	Y	Y	Y	-	R	HR	HR	Data should contain an indicator of how often it should be revalidated against other (e.g. real world) source.
Defined Data Lifetime(s)	Y	Y	Y	Y	Y	R	R	HR	HR	When does data validity expire?
Data Quality Measurement	Y	Y	Y	Y	Y	-	R	HR	HR	Uses criteria established to provide an objective measurement of the quality of a given dataset.
Data Quality Trend Analysis	Y	Y	Y	Y	Y	-	-	R	HR	Checking that a dataset is consistent with a model of the expected data behaviour. Eg. vibration data increases over time.
Authorisation	Y	Y	Y	Y	Y	R	R	HR	HR	A security model is established to control who is authorised to create, view, edit, delete the data.
Authentication	Y	Y	Y	Y	Y	R	R	HR	HR	Data is authenticated to validate its provenance.
Defined Confidence / Trust Levels	Y	Y	Y	Y	Y	R	R	HR	HR	Criteria are established to provide an objective measurement of the confidence or trust in a given dataset.

Methods and Approaches - Data Assurance	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Independent Check	Y	Y	Y	Y	Y	-	-	R	HR	A separate person or system is used to check the data independently.

8.2.3 Data Procedures

Methods and Approaches - Data Procedures	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Data Management Plan	Y	Y	Y	Y	Y	R	R	HR	HR	
Governance Model	Y	Y	-	-	Y	R	R	HR	HR	A governance model is established that defines aspects such as data ownership, processing roles & responsibilities (who can do what to the data), processing authorisations and permissions (what can be done to the data) etc.
Data Process Definition	Y	Y	Y	Y	Y	-	R	HR	HR	Documented and agreed process definitions for how data is handled.
Data Flow Diagram	Y	Y	Y	Y	Y	HR	HR	HR	HR	To describe the data flow in a diagrammatic form.
Data Model	Y	Y	Y	Y	Y	HR	HR	HR	HR	To articulate how data is organised.
Data Safety Training	Y	Y	Y	Y	Y	R	R	HR	HR	For individuals.
Data Safety Competence Assessment	Y	Y	-	-	Y	-	R	HR	HR	For individuals.
Client Sign-Off	Y	Y	-	Y	Y	R	R	HR	HR	
Data Quality Correction Mechanisms	-	-	-	Y	-	-	R	HR	HR	A process, strategy and tooling for data that breaches a given data quality criteria.
Configuration Management	Y	Y	Y	Y	Y	HR	HR	HR	HR	The recording of the production of every version of every 'significant' deliverable and of every relationship between versions of the different deliverable.
Data Dictionary	Y	Y	Y	Y	Y	HR	HR	HR	HR	A data dictionary is a collection of descriptions of the data objects or items in a data model for the benefit of data users.
Formal Methods	-	-	Y	-	-	-	R	R	HR	To specify data in a formal, mathematical manner.

Methods and Approaches - Data Procedures	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Update Comparison	Y	Y	Y	Y	Y	-	R	R	HR	Updated data is compared to its previous version. The list of changed elements can be compared with a similar list generated by the supplier.

8.2.4 Data Media Handling (Paper / Physical Storage)

Methods and Approaches - Data Media Handling (Paper / Physical Storage)	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Photographic Copies	Y	Y	Y	Y	Y	R	R	HR	HR	
Scan to Electronic Format	Y	Y	Y	Y	Y	R	R	HR	HR	
Copies Held at Different Locations	Y	Y	Y	Y	Y	-	R	HR	HR	
Limited Access	Y	Y	Y	Y	Y	-	R	HR	HR	
Secure Storage	Y	Y	Y	Y	Y	-	R	HR	HR	
Manual Inspection	Y	Y	Y	Y	Y	-	R	HR	HR	
Suitable Physical Environment	Y	Y	Y	Y	Y	-	R	HR	HR	
Defined Handling Procedures	Y	Y	Y	Y	Y	-	R	HR	HR	
Repair / Restoration Programme	Y	Y	Y	Y	Y	-	-	R	HR	
Indexing / Cataloguing	Y	Y	Y	Y	Y	R	R	HR	HR	
Lifetime Planning	Y	Y	Y	Y	Y	-	-	R	HR	

8.2.5 Data Media Handling (Electronic Storage)

Methods and Approaches - Data Media Handling (Electronic Storage)	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Regular Refresh/Rewrite	Y	Y	Y	Y	Y	R	R	HR	HR	Of magnetic media or flash memory. Life of a hard disk might be <3 years. Life of a properly-stored DVD might be 5 years. Life of a USB might be 10 years. Life of a magnetic tape might be longer if a clean tape drive was used and the tape was stored properly.
Suitable Storage/Handling	Y	Y	Y	Y	Y	R	R	HR	HR	DVDs and tapes should be kept in cases and stored vertically. Media should not be dropped. Avoid fingerprints and dust, keep out of sunlight, do not put near electronics, machinery or other possible sources of magnetic fields.
Suitable Physical Environment	Y	Y	Y	Y	Y	R	R	HR	HR	Store media in a clean, low-humidity environment at a steady temperature, cool but not cold
Copies at Different Locations	Y	Y	Y	Y	Y	R	R	HR	HR	Physically separate to cover natural disasters, accidental or malicious damage.
Backups/Duplication	Y	Y	Y	Y	Y	R	R	HR	HR	Backups are essential. Frequency of backup is dependent on the rate of change of the data. The number of generations of backup to be kept should be commensurate with the impact of data loss.
Sample Restores	Y	Y	Y	Y	Y	R	R	HR	HR	Sample restores should be performed at intervals to ensure that the backups are readable and retrievable.
Multiple Copies	Y	Y	Y	Y	Y	R	R	HR	HR	At least two backups should be kept, preferably in diverse formats.
Copy to Latest Media Format	Y	Y	Y	Y	Y	-	R	HR	HR	Anticipate obsolescence and plan a smooth transition to new technologies.

Methods and Approaches - Data Media Handling (Electronic Storage)	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Media Physically Secured	Y	Y	Y	Y	Y	-	R	HR	HR	Access to, and removal of, media should be controlled by procedures ensuring appropriate authorisation by means of locks, automatic pass recognition, and surveillance to an appropriate level. Access permissions should be reviewed at intervals. Access and removal of media should be logged and edited. Access to local and remote workstations giving access to data should also be controlled. Ensure that media disposal or reuse renders the previous content inaccessible.
Resilient / Redundant Format	Y	Y	Y	Y	Y	-	-	R	HR	This may involve less use of compression, use of error detection and correction protocols, and (at the highest level) two or more redundant data servers.
Long-Lifetime Format	Y	Y	Y	Y	Y	-	-	R	HR	The best formats should be adopted where available (the M-DISC format is said to provide enhanced lifetimes, although this is yet to be demonstrated).
Easily Translatable / Convertible Format	Y	Y	Y	Y	Y	-	-	R	HR	Adopt widely-used, general-purpose formats in preference to specialist proprietary formats. In the event of obsolescence, a widely-used format is more likely to have an easily-accessible means of conversion to a newer format.
Copy to Cloud Storage	Y	Y	Y	Y	Y	-	R	HR	HR	Must specify whether a private cloud or a public cloud shall be used. A view must be taken regarding the required level of integrity and confidentiality of the data and an appropriate solution adopted. Cloud storage may not be suitable for highly confidential data.
Copy to Archiving Organisation	Y	Y	Y	Y	Y	-	R	HR	HR	A view must be taken regarding the required level of data integrity and confidentiality. The integrity and long-term viability of the organisation, and steps to be taken in the event of its ceasing to function, must be considered. Many of the considerations mentioned elsewhere would apply.

8.2.6 Data Usage Confidence

Methods and Approaches - Data Usage Confidence	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Limited / Pre-Operational Deployment	-	Y	Y	Y	-	-	R	HR	HR	
Client Sign-Off of Data	Y	Y	-	Y	Y	-	R	HR	HR	
Non-Critical Trialling	-	-	Y	-	-	-	R	HR	HR	
Beta Testing	Y	-	-	-	-	-	R	HR	HR	
Parallel Running	-	Y	Y	Y	-	-	R	HR	HR	
Widespread Distribution to User Community	-	Y	Y	Y	-	-	R	HR	HR	
Open Source Techniques	-	-	Y	-	-	-	R	HR	HR	

8.2.7 Test Data Generation

Methods and Approaches - Test Data Generation	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Using Informal / ad-hoc means	Y	-	-	-	-	R	R	-	-	This is where data is generated by simple spreadsheets, or by simple scripts or programmes. It may also be legacy data or basic assumptions. There is no formal checking or review of the method of generation.
Using Testbed	Y	-	-	-	-	-	R	HR	HR	A dedicated testbed (i.e. a specific set of hardware and software tools designed for testing) is a good way to produce test data, providing the testbed has the functionality required. It may require configuration and tailoring for the particular application, and this configuration should be managed.
Using Simulator	Y	-	-	-	-	-	R	HR	HR	Simulators (software or hardware) may be able to produce very good test data, obviously depending on how close and detailed a simulation they can achieve. Timing data may be very different using a simulator, and there may be issues with numerical accuracy.

Methods and Approaches - Test Data Generation	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Using Prototype	Y	-	-	-	-	-	R	HR	HR	Prototypes are often a good way of generating test data for the real system, as they are often have much of the required functionality of the final system. However they may not produce data with the appropriate range, accuracy or precision. They may also not cover all the error cases.
Using Manual means	Y	-	-	-	-	R	R	-	-	Simple test data can be produced by manual means, although this may be prone to human error. However manual checking of a sample of test data generated using tools is a useful verification method.
Using Dedicated Platform	Y	-	-	-	-	-	R	HR	HR	For complex and critical systems a dedicated test platform is required which can produce realistic test data for all interfaces and inputs. The dedicated platform can become a large project development in itself.
Using Existing/Established System	Y	-	-	-	-	-	R	HR	HR	Where a new system replaces an old one, then data can often be extracted from the old system to test the new one. This can work well, but data formats may change and translation/reformatting may be required.
Using Initial Runs of New System	Y	-	-	-	-	R	R	R	R	This method is often used where the system is breaking new ground and there is no prototype or legacy system to produce test data. This must be carefully used as initial operations can be very different to eventual usage, and so the test data suite must also evolve.
Derived from Real Data	Y	-	-	-	-	R	R	HR	HR	Where real data is available this is usually a good basis for generating test data (e.g. by modification to increase the test space coverage). However there are potential issues of sampling and coverage, i.e. is the real data a representative sample?
Statistical Profiling Post-Production	Y	-	-	-	-	-	-	R	HR	If a statistical analysis of the data can be produced then greater confidence in the quality of the test data can be obtained. A good example is the generation of pseudo-random numbers, where the distribution is known.

Methods and Approaches - Test Data Generation	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Produced by Client	Y	-	-	-	-	R	R	R	HR	Ideally the client is involved in producing or at least checking the test data. The client will often know the data intimately and can highlight any issues quickly.
Client Sign-Off	Y	-	-	-	-	R	R	HR	HR	Where possible, the client should formally agree and sign-off the test data as appropriate. This gives the system developer some confidence in the data and also some protection of the data is in fact incorrect or not representative.
Error Seeding	Y	-	-	-	-	R	R	HR	HR	This is where errors are deliberately inserted into the dataset to demonstrate the effectiveness of data validation.
Data Re-use	Y	-	-	-	-	R	R	HR	HR	Re-using data for one project that was created and thoroughly assured for another project. This can be effective but the read-across should be established
Feedback testing	Y	-	-	-	-	R	R	R	R	To check output data by comparing it with the input source.

8.2.8 Test Tools

Methods and Approaches - Test Tools	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Informal / Office Tools	Y	-	-	-	-	R	R	R	R[7]	Risk assess if used for calculations or other data transformations.
Specific Tools / Scripts	Y	-	-	-	-	R	HR	HR	R[7]	Validate and verify before use.
Formally Developed Tools / Scripts	Y	-	-	-	-	R	HR	HR	HR	Validate and verify before use.
Risk-assessed Tools	Y	-	-	-	-	R	R	HR	HR	See Tools Assurance Sub-section 8.3.
Qualified Tools	Y	-	-	-	-	-	R	HR	HR	Formally approved according to a 'standard'.

7 Generic commercial tools may not be suitable for high-integrity use - specialist tools may be required.

8.2.9 Test Results Analysis

Methods and Approaches - Test Results Analysis	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Informal / Office Tools	Y	-	-	-	-	R	R	R	R	
Specific Tools / Scripts	Y	-	-	-	-	R	HR	HR	HR	
Formally Developed Tools / Scripts	Y	-	-	-	-	R	HR	HR	HR	
Dedicated Analysis Platform	Y	-	-	-	-	-	R	HR	HR	
Commercial Tools	Y	-	-	-	-	-	R	HR	HR	

8.2.10 Data Migration

Methods and Approaches - Data Migration	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Informal / Office Tools	Y	Y	Y	Y	Y	R	R	-	-	Use of office tools such as spreadsheets, databases, delimited character files to hold and manage the data
Manual Load	-	-	Y	-	-	R	R	-	-	Data is entered into the system manually relying on human validation and verification.
Dedicated Translation and Loading Platform	-	-	Y	-	-	-	R	HR	HR	For example, using mature enterprise migration Commercial Off-The-Shelf (COTS) products.
Existing/Established System Transfer	-	-	Y	-	-	-	R	HR	HR	Use of an existing/established proven transfer mechanism.
Client Supervision	Y	Y	Y	Y	Y	-	R	HR	HR	The client provides independent supervision of activities checking processes, inputs and outputs at agreed points throughout and at the end of the process.
Client Sign-Off	Y	Y	Y	Y	Y	-	R	HR	HR	Formal acceptance of the migrated datasets in the target system.
Incremental switch-over	-	-	Y	-	-	-	R	HR	HR	Users are incrementally switched over to the new system rather than as a 'big bang'.
Parallel Load With Existing System	-	-	Y	-	-	-	R	HR	HR	Parallel running of the new system alongside the existing system with data crosschecks between the two systems.

Methods and Approaches - Data Migration	Lifecycle Data Types					Data Integrity Level				
Technique	**V**	**I**	**D**	**P**	**J**	**DIL 1**	**DIL 2**	**DIL 3**	**DIL 4**	**Notes**
Shadowing	-	-	Y	-	-	-	R	HR	HR	Parallel running of the new system alongside the existing system such that only data from the existing system is used operationally and with an experienced user crosschecking between the two systems so as to validate the new one (or not, as the case may be).
End to End Import-Export Verification	-	-	Y	-	-	-	R	HR	HR	Data is traced and verified at all stages through the entire end to end migration process

8.3 Tool Assurance

Tools in this context are considered anything that automates all or part of a process, e.g. data creation or data transformation; it could also be a test tool, rather than part of a data system.

Tools have different potential for adverse impact on data safety, depending on both their function and how they are to be used. For tools to be considered fit for purpose it is necessary to show that the tool meets its requirements in the context in which it is to be used. The activity to ensure a tool is fit for purpose is usually called "tool qualification".

The first step is, of course, to define the purpose for which the tool is required to be fit. Once that is done, and the tool's requirements are specified, there are three main strategies available for qualification:

1. Use evidence of a previous certification of the tool by a trusted third party (unlikely to be available in most industry sectors);
2. Base tool qualification on the practices used when designing and developing the tool (only practical for tools developed within the organisation);
3. Use one of the available industry-specific guidance documents that admit COTS solutions, e.g. EUROCAE Document ED-215 (RTCA/DO-330) [7].

There is an alternative, risk-based and perhaps more practical, approach. Assess the potential risks presented by use of the tool and provide assurance that they are adequately managed. The method proceeds as follows:

- Draft a procedure for the use of the tool to achieve the stated purpose;
- Identify threats to data safety associated with using the tool;
- Specify adequate mitigations for each identified threat;
- Augment and formally issue the tool requirements and the usage procedure to implement the specified mitigations;
- Demonstrate that the tool and its mitigations perform as expected; and

- Provide a compelling assurance argument based on the previous steps and any other evidence that will improve confidence, e.g. reputation of the supplier, configuration management of the tools and its documentation, competence of the tool user and of those who check the outputs.

9 Partial Worked Example

> *Experts often possess more data than judgement.*
> ***Colin Powell***

9.1 Purpose

This section provides a partially worked example of applying the Data Safety guidance. The example that is used is intended to be sufficiently realistic to allow key features of the guidance to be illustrated. However, aspects of it have been deliberately simplified. In addition, it is entirely imaginary; it is not based on any current or future system. Furthermore, any numerical figures that appear in the example are only included for illustrative purposes; they are not intended to be realistic and should not be construed as being so.

The following typographical conventions are used:

- *Italic text is used for items that explain aspects of the partial worked example.*
- Normal text is used for the actual contents of the partial worked example.

9.2 Establish Context

This worked example begins when a need has been identified but no significant system design activity has been completed. However, it should be noted that the guidance can be applied at any stage of a system's lifecycle. If an iterative approach is being used for system design then several of the activities may need to be repeated at different times during the development lifecycle. The first part of the worked example is concerned with establishing the system-level context in which the data will be used. A relatively simple example is provided below.

Background:

Helicopter undercarriages are currently replaced on a time-based maintenance schedule. In particular, each undercarriage is replaced after a year, regardless of usage.

The system under development, which is known as the Helicopter Undercarriage Replacement System (HURS), is intended to support a maintenance schedule that includes both time-based and usage-based elements. In particular, it is planned that the future maintenance schedule will involve replacing helicopter undercarriages no sooner than one year after fitting and when one of the following two conditions is satisfied: either the undercarriage has been fitted for a period of two years; or the undercarriage has been subjected to two hundred landings. Analysis of previous usage rates suggests that the two hundred landings is typically achieved after between eighteen and thirty months' worth of use.

In order to achieve this new maintenance schedule, the number of landings needs to be recorded, stored and analysed.

(The time since the undercarriage was last changed also needs to be recorded, stored and analysed. For the purposes of this example it is assumed that there will be no changes to the way that this is currently achieved. This is an obvious simplifying assumption that has been adopted for the purposes of this partial worked example.)

The aircraft is equipped with a number of sensors that could be used to provide information relating to the number of landings:

1. Part of the engine management software records the number of times that the engine is started. This count provides a lower-bound on the number of landings: the engine will only be started when the aircraft is on the ground; but some landings may be sufficiently short that they do not entail an engine shut-down.

2. A simple 8-bit counter is applied to the weight on wheels sensors. Two such sensors are available, one on each of the main undercarriage wheels. These are basic pressure switches, which are sufficiently sensitive that a "bouncy" landing can register multiple signals.

3. A number of accelerometers have been fitted to the aircraft. Whilst these were originally fitted for the purposes of vibration monitoring, they can capture signals that are indicative of a landing. However, similar signals can occur if the aircraft is subject to a significant upward gust of wind. Four such accelerometers are on the aircraft, each of which is in a different position on the fuselage.

The first two of these items are included in the standard data that is downloaded from the aircraft. These downloads are currently stored, alongside other aircraft data, in a computer-based system.

The third item is not currently downloaded on a routine basis. It requires a specialist cable for the download; the data is stored on a separate machine to the other two items identified above.

In addition to the three digital items noted above, paper flight records are also maintained. These are currently stored in a filing cabinet in the maintenance hangar, with a duplicate copy being stored at the company's head office.

The above text provides some background to the planned system development. However, establishing the context also involves considering organisational issues as well.

The helicopters are owned and maintained by the same organisation. Engine maintenance is conducted by the engine manufacturer. The values of two-years and two hundred landings suggested for the new maintenance schedule have been provided by the Original Equipment Manufacturer (OEM).

Note that, for the purposes of this worked example, the values provided by the OEM are considered to be fault-proof. However, in general, there is a need to have a good understanding of Data Safety issues related to any externally-provided data items (like "two-years" and "200 landings"); the depth of this understanding should be related to both the criticality of the system being developed and the criticality of the externally-provided data items within that system.

The computer-based system used to store the engine and weight on wheels data was developed by a sub-contractor. A different sub-contractor now provides support and maintenance for the system. Data input, analysis and retrieval is conducted by employees of the helicopter operator.

The system used to record and analyse the accelerometer data was developed by a third sub-contractor. It is not routinely used by the helicopter operator.

Having provided a textual description of the context, the next step involves assessing organisational risk. This is achieved using the Organisational Data Risk (ODR) Assessment Form, which is contained in Appendix B. *For ease of presentation, the textual description has been provided before the ODR assessment. However, in practice, completing the assessment often highlights additional aspects that should be included in the description.*

Organisational Data Risk (ODR) Assessment:

The following list identifies the questions in the ODR, along with the assessments for HURS.

Q1: How severe could an accident be that is related to the data? Could it be caused directly by the data?

Failure of landing gear could lead to a significant accident, which could result in significant injuries to a number of people; several deaths are a possible, but unlikely, outcome. Some environment impact is possible, but a major environmental impact is considered highly likely.

On balance, this is question assessed as: 1d; Score 8.

Q2: What would be the impact on the organisation, client or public if an accident occurred related to the data?

The public would be concerned about any accident, rather than being alarmed. Regional press interest is likely, as is a formal investigation. An accident traced to the proposed new maintenance schedule could be seen as a result of "poor practice".

On balance, this is question assessed as: 2d; Score 8.

Q3: How much responsibility does this organisation have for data safety?

The organisation (ie, the helicopter operator and maintainer) has prime control over the data safety responsibility. Some aspects of this may be delegated to the developers of the systems used to store, analyse and retrieve the data, but these considerations are not yet explicitly identified in contracts.

On balance, this is question assessed as: 3d; Score 7.

Q4: What legal and regulatory environment will this work be subject to?

Helicopter flying is subject to a single, well-understood legal framework. It is a highly-regulated sector, with established guidelines and standards, supported by a formal certification process.

On balance, this is question assessed as: 4a; Score 1.

Q5: How mature is this organisation regarding data safety?

The organisation has a good understanding of data as a source of safety risk. In some areas (e.g. flying hours) formal processes are in place. Other areas (e.g. chart management) rely on informal processes. There is good support and funding for the identification and resolution of data-related risks.

On balance, this is question assessed as: 5b; Score 2.

Q6: How widely used is the data and who by?

There is a single user for this data, which is an organisation with independently audited quality management systems. Staff will be provided with appropriate training to operate the HURS system.

On balance, this is question assessed as: 6a; Score 1.

Q7: What is the scale, sophistication and complexity of the data and its manipulation?

Simple data structures are associated with the engine and weight on wheels data. The number of interfaces is small (sensor to bus to transfer disk to storage system) and no transformations are required. The data is relatively easy to verify (e.g. by comparing to paper flight records) and easy to trace.

The accelerometer data is structured, but complicated, and transformations are required to identify landing signatures in the data. Three interfaces (sensor to on board store to transfer disk to storage system) are involved in extracting the data from the aircraft.

On balance, this is question assessed as: 7b; Score 2.

Q8: How well defined and understood are the boundaries and interfaces for this data scenario?

The boundaries and interfaces are simple, well-understood and well-documented. Those associated with the engine and weight on wheels data are regularly used; the ones associated with the accelerometer data are used less frequently.

On balance, this is question assessed as: 8a; Score 1.

This ODR illustrates that an answer very rarely falls into a single category. Judgement is required to establish the appropriate level. In many cases it may be appropriate to adopt a pessimistic / worst-case approach, but care needs to be taken to ensure the overall assessment is not pessimistic to an incredible extent. Also note that this ODR has been filled in at the planning stage of the HURS project. Hence, it makes some assertions (e.g. appropriate training will be provided): critical assertions should be captured in the DSMP and tracked through to closure.

The final score is 30, which corresponds to ODR2. This score is driven by the potential consequences of inappropriate data usage in this scenario, rather than, for example, the complexity of the planned system.

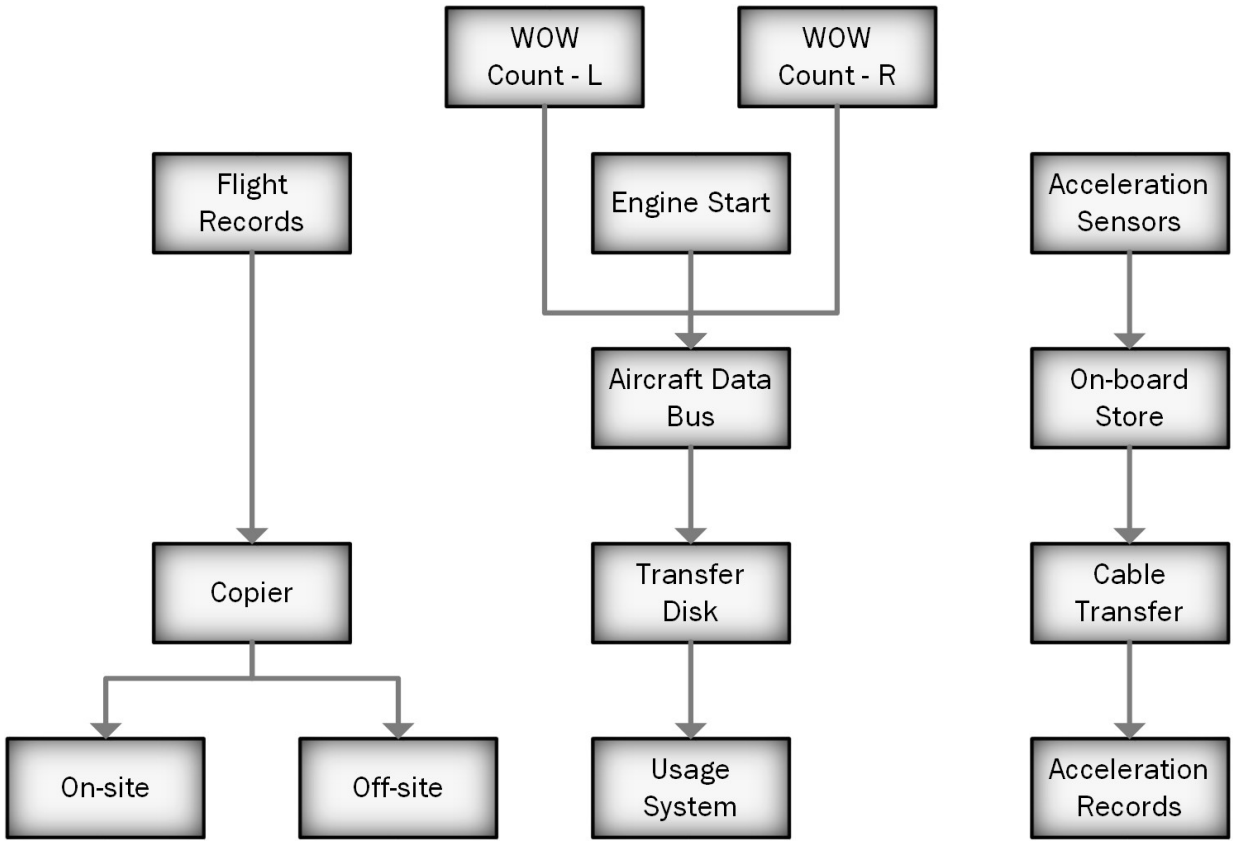
ODR2 is a mid-level result. This, combined with the nature of the organisation (which operates in a formally regulated environment and which has considerable experience of system safety considerations), means there is judged to be little value in completing a Data Safety Culture Questionnaire.

The ODR2 result, along with the simple interfaces associated with the planned HURS, also means there is judged to be little value in completing a Dataware Framework Assessment.

From the perspective of this worked example, the final activity in the "Establish Context" phase involves producing an initial system design. This should allow the relevant Data Artefacts to be identified.

<u>System Design - Outline:</u>

The following schematic outlines the system design that has been adopted.



The primary part of the system is shown in the central column above. This involves taking data from the two Weight On Wheels (WOW) sensors, as well as the number of engine starts, and storing this in the Usage System. This is a computer-based system that is currently used to provide Business Intelligence (BI) information (rather than to directly support safety-related or safety-critical activities).

Paper flight records will be used as a secondary source of data. These are intended to be used to conduct spot checks on the data going into the Usage System. They may also be used to resolve any discrepancies between the various data items going into the Usage System.

Data from the Acceleration Sensors will be captured on an irregular basis. This will be used as another means of spot-checking the data that is being stored in the Usage System.

System Design - Data Artefacts:

The Data Artefacts included in the system design are:

- DA-1 Paper-based flight records, which detail the planned number of landings and take-offs.
- DA-2 The value of the 8-bit counter associated with the left-hand WOW sensor.
- DA-3 The value of the 8-bit counter associated with the right-hand WOW sensor.
- DA-4 The value of the engine start counter.
- DA-5 The transfer disk that contains the WOW sensor and engine start counters.
- DA-6 The on-board acceleration records, which are a compressed, time-stamped history of the values recorded by the acceleration sensors.

- DA-7 The storage system containing WOW sensor counts (both left-hand and right-hand) and engine start counts.
- DA-8 The off-board acceleration records, which are the results of processing the on-board records (processing is conducted before the data is stored).

Note that the Data Artefacts have been identified at a range of levels. For example, the WOW sensor counts are identified as Data Artefacts, as is the transfer disk that is used to move the data from the aircraft to the relevant storage system. This allows the specific nature of these items to be considered in subsequent analysis, albeit at the cost of duplicating some parts of that analysis.

9.3 Identify Risks

Three complementary methods are provided for identifying risks. These range from a simple top-down / bottom-up approach, through consideration of generic data safety issues to a full Hazard and Operability (HAZOP) study using appropriate guidewords. The method(s) that is (are) chosen should be commensurate with the nature of the system and the reasons for the choice(s) should be documented.

To recap, the preceding discussion determined that this activity had an ODR2 and that this value was driven by the potential consequences of the data's use rather than (for example) the complexity of the data or the associated system. Against that background, there is judged to be no need to conduct a full Hazard and Operability (HAZOP) study.

Initially, risks are identified using the top-down / bottom-up approach, as outlined below:

- A paraphrase of the safety function is: “Land safely”.
- The overall top-level risk is: “Accident on landing due to fatigued undercarriage”.
- From the perspective of the system under consideration, the key contributing factor is: “Failure to replace undercarriage when necessary”.
- Lower-level factors are:
  - Incorrect replacement schedule. *(This captures the possibility that the stated number of landings that the undercarriage can tolerate is incorrect. As stated earlier, this possibility is not considered further in this worked example.)*
  - Incorrect counting of number of landings.
- The lowest-level risks are:
  - LLR-A. Incorrect data on paper flight records.
  - LLR-B. Incorrect data on aircraft - left-hand WOW sensor.
  - LLR-C. Incorrect data on aircraft - right-hand WOW sensor.
  - LLR-D. Incorrect data on aircraft - engine start counter.
  - LLR-E. Incorrect data on transfer disk - corruption on write.
  - LLR-F. Incorrect data in usage system from single flight - misread.
  - LLR-G. Incorrect data in usage system - storage and retrieval.

- LLR-H. Incorrect data on aircraft - acceleration sensors.
- LLR-I. Incorrect data transfer - acceleration data via cable.
- LLR-J. Incorrect data in acceleration records.

A comparison of these risks with the outline system design indicates that this top-down approach has also captured all relevant items from a bottom-up perspective.

This outcome, where the bottom-up approach adds little extra information is a consequence of the simple nature of the system. However, as indicated by the fourth Data Principle, it is always important to consider things from a bottom-up perspective.

These risks are now considered from the perspective of the generic data safety issues. For reasons of brevity, those generic issues that are not relevant to this system are not mentioned below.

- Aggregation is used when combining data from the WOW and engine start sensors onto the transfer disk. This raises the possibility that one set of data may overwrite another. It also raises the possibility that there is a mismatch between the locations in which data is stored and the locations from which it is read.
- Archiving and Retrieval is part of the usage system and the acceleration records.
- Biasing applies to all of the data obtained from the aircraft:
  - The WOW sensors can record multiple signals from a bumpy landing. Hence, they are biased towards over-counting.
  - Conversely, the WOW sensors are 8-bit counters. Hence, the maximum value they can store is 255. This limit may inadvertently introduce another bias into the system.
  - The engine may not be turned off at every landing. Hence, the engine start counter is biased towards under-counting.
  - Upward gusts of wind can cause signals in the accelerometers that mirror those associated with landing. Hence, the accelerometer data is biased towards over-counting.
- Biasing is also an important issue at the system level. If the landings are over-counted then undercarriage will be replaced sooner than would be necessary, which is a safe situation. Conversely, under-counting landings would delay undercarriage replacement, leading to a potentially dangerous situation.
- Aliasing is important in the usage system, which will store records from multiple aircraft. If the records from one aircraft get aliased to a different aircraft (ie, a different tail number) a potentially dangerous situation could occur.

9.4 Analyse Risks

The next phase of the process described in the Data Safety document involves analysing the risks. This involves comparing the severity of impact of a data-related accident with the likelihood of data-related issues (or, more specifically, the loss of a required data property) causing an accident.

As noted previously, only a single accident is being considered in this analysis. In particular, we are concerned with an "accident on landing due to fatigued undercarriage". The considerations made when

completing the ODR suggest that the Severity of this accident is Major. This applies to all of the previously identified Data Artefacts.

The following list considers the Likelihood of a specific Data Artefact causing an accident.

For reasons of brevity, this list only considers "DA-2: The value of the 8-bit counter associated with the left-hand WOW sensor" and "DA-5: The transfer disk that contains the WOW sensor and engine start counters".

1. DA-2: The value of the 8-bit counter associated with the left-hand WOW sensor.

This data item is used alongside a number of "sibling" data items. (e.g. DA 3: the right-hand WOW sensor, DA-4: the engine start counter). These sibling items reduce the Proximity and Dependency of this Data Artefact; they also enable Detection, Prevention and Correction of data-related errors. Hence, this Data Artefact is assessed as having a Low likelihood of causing an accident.

The Major severity, combined with a Low likelihood, leads to a Data Integrity Level (DIL) of 2.

2. DA-5: The transfer disk that contains the WOW sensor and engine start counters.

With regards to Proximity, the nature of the transfer disk means its data is closer to a potential accident than was the case for the WOW sensor. From the perspective of that concern, a failure in this Data Artefact is assessed as having a Medium / High likelihood (of causing an accident).

Likewise, whilst there are other sources of data (e.g. DA 1: paper-based records; DA 6: acceleration records) these are not expected to be used to cross-check every single data transfer. Hence, with regards to Dependency, a failure in this Data Artefact is also assessed as having a Medium / High likelihood.

With regards to Detection, Prevention and Correction, DA 1 and DA 6 (noted in the previous paragraph) provide an opportunity to detect, prevent and correct errors; checksums will be used to detect errors prompting prevention and correction activities as necessary. Similarly, even though it is primarily conducted for BI and planning purposes, analysis conducted by the storage system (e.g. to show landings by tail number) should help in detecting significant data errors. Overall, a failure in this Data Artefact is also assessed as having a Medium likelihood for the three concerns of Detection, Prevention and Correction.

The possibility of using the storage system to, for example, determine the "fleet leader" in terms of numbers of landings on current undercarriage was not previously identified in the discussion of this worked example. The fact that it has become apparent during the current phase indicates the value of considering each Data Artefact from the perspective of the five likelihood concerns (ie, Proximity, Dependency, Detection, Prevention and Correction).

Based on the above considerations, this Data Artefact is assigned a DIL of 3.

It is apparent from the preceding discussions that a Data Artefact rarely falls into a single severity or a single likelihood category. As with other aspects of the Data Safety processes, judgement is required. Whilst a conservative approach is often sensible, care needs to be taken that the overall assessment is not conservative to an incredible extent. As noted earlier, only DA 2 and DA 5 are considered in this partial worked example. However, for completeness, the full list of DIL allocations is provided below.

The following table summarises the DIL allocations for all the identified Data Artefacts.

Data Artefact	DIL
DA-1 : Paper-based flight records	DIL 2
DA-2 : Left-hand WOW sensor counter	DIL 2
DA-3 : Right-hand WOW sensor counter	DIL 2
DA-4 : Engine start counter	DIL 2
DA-5 : Transfer disk	DIL 3
DA-6 : Acceleration records	DIL 2
DA-7 : Storage system (counters)	DIL 3
DA-8 : Storage system (acceleration records)	DIL 2

It is apparent from the above that the most critical Data Artefacts are the transfer disk (DA-5) and the system that stores the WOW and engine start counter information (DA-7). This makes intuitive sense, since there is some redundancy between the on-board Data Artefacts (DA-2, DA-3, DA-4 and DA-6) and both the paper-based records (DA-1) and the acceleration records storage system (DA-8) will be used to provide spot-checks for the counters storage system.

9.5 Evaluate Risks

This phase involves comparing the risks that have been identified with the risk appetite. It also provides an opportunity to review (and, if necessary, update): the ODR value; the identified Data Artefacts and associated DILs; and the proposed system design. Since this is a partial worked example, no details of these documents are provided.

9.6 Treat Risks

This phase involves determining appropriate treatments for the identified risks. For reasons of brevity, only those Data Artefacts that were assessed at DIL 3 are considered in this partial worked example.

DA-5: Transfer disk

- Checksums used to confirm data is unaltered on load from transfer disk.
- "Out of bounds" data trapped before disk is written.
- Data includes a date / time stamp. (So that we can check we're not simply adding the last flight's data to the storage system.)

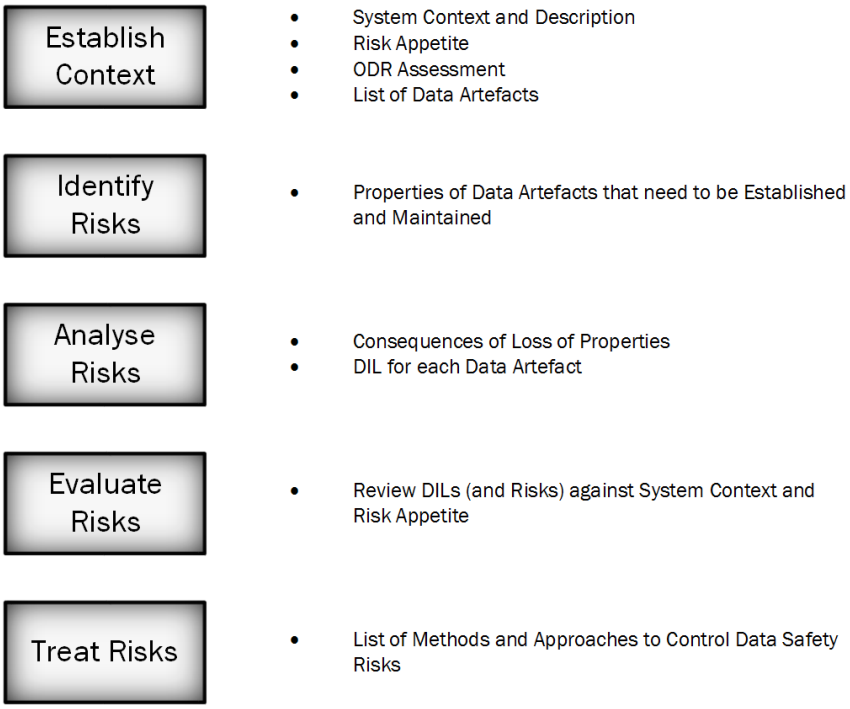
DA-7: Storage system (counters)

- Appropriate contractual mechanisms established with the software developer.
- Appropriate development process used for software and suitable development artefacts produced (generated retrospectively, where necessary and possible).
- Tail numbers entered (and stored) in a manner that reduces the risk of accidental aliasing.
- Appropriate backup / archive / off-sire storage process in place.
- Appropriate staff training conducted.
- Suitable logging is in place, so that (as a minimum) all changes to data are recorded.

- A statistically-meaningful number of cross-checks conducted with both the paper records and the acceleration data.
- Explicit consideration of whether the monthly BI reporting from the tool includes any apparent anomalies in the number of landings (by tail number).
- Explicitly account for overflowing of either the left-hand WOW counter or the right hand WOW counter.
- Store each of: left-hand WOW; right-hand WOW; engine start counter. Use the highest value for the number of landings that have been conducted. (This should bias the system in a safe direction.)

9.7 Process Summary

The following figure summarises the main process steps and associated outputs.



10 Conclusions

> *The world is one big data problem.*
> ***Andrew McAfee***

The nature of systems is changing, with the role of data becoming ever more prominent. This means that data and, more specifically, the properties it is required to exhibit have a direct effect on system safety. A number of accidents and incidents have already occurred in which "inappropriate data" was amongst the causal factors.

Raising data to a level where it is considered as a "first class citizen" (alongside software and hardware) in system safety analyses can help protect against such occurrences. This guidance document has outlined a collection of activities that facilitate this, including: a set of data types, and associated properties; a way of establishing the appropriate context (e.g. for system assessments); and methods for identifying, analysing, evaluating and treating risks. In addition, the "4+1" software safety principles have been re-interpreted in the context of data safety and, furthermore, the close link between data safety and security has been elucidated.

Although it is not a panacea, and it is still under active development, this guidance document provides a means of managing and mitigating the risks associated with the use of data in safety-related systems. As such, the authors' hope it will provide a valuable contribution to the development and safe operation of such systems.

Appendix A Incidents and Accidents

> *Two men were examining the output of the new computer system in their department. After an hour or so of analysing the data, one of them remarked: 'Do you realise it would take 400 men at least 250 years to make a mistake this big?'"*
> ***Anon***

A.1 General

The following 'War Stories' describe incidents and accidents in which data is considered to have been a contributory factor. A data perspective has been taken to demonstrate the need for data to be given equal footing alongside software, hardware and human factors.

Note: The analysis presented here has no legal standing whatsoever. The purpose of this section is not to discredit, contradict or undermine any existing accident analysis; the aim is simply to view these incidents from a data perspective. Where possible accident reports have been referenced with the role of data highlighted. All references have been taken at face value and not independently verified.

A.2 Lake Peigneur Drilling Accident

Summary Lake Peigneur is located in Louisiana, United States of America. It was a ten-foot deep freshwater lake popular with sportsmen. On 20th November 1980 an exploration rig drilling for oil in the lakebed was evacuated as it began to sink; this was perceived by the crew as a structural collapse. Meanwhile, the nearby Jefferson Island salt mine was being evacuated due to the sudden onset of flooding.

The rig crew had been drilling a test well into deposits alongside a salt dome under Lake Peigneur. By some miscalculation, the assembly drilled into the third level of the nearby Diamond Crystal Salt Mine. Fresh water from the lake soon began trickling into the salt mine. Over the course of the morning, the fresh lake water began dissolving the salt and enlarging the hole until water was literally flooding into the mine.

The whirlpool created as the lake drained into the mine sucked in the drilling platform, eleven barges, trees and soil. The Delcambre Canal, which usually drains from the lake into a bay on the Gulf of Mexico, had its flow reversed. This resulted in Lake Peigneur becoming a saltwater lake. Fortunately, no injuries or loss of human life were reported.

Role of Data Federal experts from the Mine Safety and Health Administration were not able to determine the cause of the accident due to confusion over whether the rig was drilling in the wrong place or whether the mine's maps were inaccurate. However, the incident demonstrates the potentially significant effects of either a data error, or the misinterpretation of data.

Sources

- Wikipedia, http://en.wikipedia.org/wiki/Lake_Peigneur, last accessed 11/11/2015.
- Oil Rig Disasters, http://home.versatel.nl/the_sims/rig/lakepeigneur.htm, last accessed 11/11/2015.

A.3 Comair Flight 5191

Summary On 27th August 2006 Comair flight 5191 crashed during take-off from Blue Grass Airport, Lexington, Kentucky. The flight crew was instructed to take off from runway 22, but instead lined up on runway 26 and began the take-off roll. The airplane ran off the end of the runway and impacted the airport perimeter fence, trees, and terrain. The captain, flight attendant and 47 passengers were killed.

The National Transportation Safety Board determined that the probable cause of the accident was the flight crewmembers' failure to use available cues and aids to identify the airplane's location on the airport surface during taxi and their failure to cross-check and verify that the airplane was on the correct runway before take-off.

Role of Data The Airport Charts used by the crew were inaccurate. The airport was under construction, and the charts were not kept current with the rapid changes that were taking place during the construction work. The chart did not accurately reflect either the taxiway identifiers and or the taxiway that was closed on the day of the accident.

Due to a previously unrecognised software glitch, any information the chart provider received after normal work hours on Fridays was not included in their regular updates. Furthermore, the chart provider modified the Blue Grass Airport chart after the accident to include a note that Runway 8/26 is "daytime VMC use only", even though this information had been published since 2001.

Additionally there was a local Notice to Airmen (NOTAM) issued advising of taxiway closures due to construction work. However the crew was not provided with this information in their dispatch paperwork.

Sources

- Attempted Takeoff from Wrong Runway - Comair Flight 5191 - Accident Report, National Transportation Safety Board, NTSB/AAR-07/05.
- Wikipedia, http://en.wikipedia.org/wiki/Comair_Flight_5191, last accessed 11/11/2015.

A.4 Mars Climate Orbiter

Summary The Mars Climate Orbiter was a spacecraft launched aboard a Delta II rocket by NASA from Cape Canaveral on 11th December 1998. Its intended mission was to study the Martian atmosphere and climate, whilst acting as a communications relay for other spacecraft on or near Mars.

The plan was that the rocket would place the spacecraft into a transfer orbit to Mars, which would be optimised along the way by a series of four trajectory correction manoeuvres. Insertion into Mars orbit was to take place at an altitude of 226km, but during the week after the final correction manoeuvre, calculations predicted that it would be between 150km and 170km; revised to 110km the day before insertion. The orbiter was able to survive atmospheric stresses down to about 80km.

On 23rd December 1999, the spacecraft passed behind Mars, and so out of radio contact, earlier than expected; communications were never regained.

Final calculations placed the spacecraft in a trajectory that would have taken it within 57km of the Martian surface, but it is likely to have disintegrated before getting to that point.

Role of Data It transpires that the orbiter's Flight Management System (FMS) software was designed to work with metric Newton seconds, whereas a FMS data-file generated by ground system software used

pound-force seconds. A Newton is about 22.5% of a pound-force or a factor of 4.45. (See Section 4 of reference [8].

The cost of the mission was stated by NASA to have been $327.6 million in total ($193.1 million to develop the spacecraft, $91.7 million for launch and $42.8 million for mission operations).

This incident shows the importance of tracking data properties (including units) throughout the entire system.

Sources

- Wikipedia, http://en.wikipedia.org/wiki/Mars_Climate_Orbiter, last accessed 11/11/2015.
- See also [9].

A.5 MS Oliva

Summary At about 0510 (UTC) on 16 March 2011, OLIVA, a Maltese registered bulk carrier ran aground on the north-west coast of Nightingale Island in the Tristan Da Cunha Group. OLIVA was on a loaded passage from Santos, Brazil to China. The vessel sustained severe bottom damage to almost all of her water ballast tanks that resulted in the vessel developing a 12 degree list to port.

On 18 March, the vessel broke up in two sections; the forward section drifted away and the aft section capsized and sank. All this resulted in widespread pollution around the islands of Nightingale and Inaccessible because of the diesel and fuel oil that escaped from the vessel's fuel tanks.

Role of Data Both the second mate and chief mate were not aware that OLIVA was heading towards Nightingale Island. This was because there was apparently no indication on the plotting chart to alert them of the dangers ahead. It appeared that the bridge team was focused on following the GPS track (red course line) superimposed on the radar screen instead of monitoring the vessel's position in relation to surrounding hazards.

'No Go' areas were not marked on the chart. It appeared that the vessel did not have BA Chart 1769, which was the appropriate large scale chart covering the Tristan Islands.

This incident highlights the importance of data resolution and availability.

Sources

- 'Safety Investigation into the grounding of the bulk carrier OLIVA On Nightingale Island, Tristan Da Cunha on 16 March 2011', Transport Malta - Marine Safety Investigation Unit, Marine Safety Investigation Report No. 14/2012.

A.6 Sichem Osprey

Summary On 10 February 2010 at 0436 (local), the chemical tanker SICHEM OSPREY, on her way from Panama to Ulsan (South Korea) stranded at more than 16 knots on the north-easterly part of Clipperton Island, although an Officer Of the Watch and a lookout were on the bridge and no damage was reported prior to the accident. A 100 metre fore part of the vessel had been grounded. No pollution was observed.

Role of Data Anti-collision radar alarm thresholds were apparently not set according to the Captain's instructions. The adjustments were not reappraised by any of the Officers or the Captain. There were sizeable discrepancies between the fixes plotted on the chart and those displayed on the radar.

This incident highlights the role of adaptation type data, used to set radar alarm thresholds. It also relates to the data properties of accuracy and traceability (e.g. for the adaptation data).

Sources

- Stranding of the chemical tanker vessel SICHEM OSPREY on 10 February 2010 on Clipperton Island, Bureau d'enquêtes sur les évènements de mer.

A.7 The Pride of Canterbury

Summary On 31 January 2008, the roll-on roll-off Passenger ferry, PRIDE OF CANTEBURY grounded on a charted wreck while sheltering from heavy weather in an area known as 'The Downs' off Deal, Kent. The vessel suffered severe damage to her port propeller system but was able to proceed unaided to Dover, where she berthed with the assistance of two tugs.

The vessel had been in the area for over 4 hours when, while approaching a turn at the northern extremity, the bridge team became distracted by a fire alarm and a number of telephone calls for information of a non-navigational nature. The vessel overshot the northern limit of the identified safe area before the turn was started. The Officer Of the Watch (OOW) became aware that the vessel was passing close to a charted shoal, but he was unaware that there was a charted wreck on the shoal. The officer was navigating by eye and with reference to an electronic chart system which was sited prominently at the front of the bridge, but he was untrained in the use and limitations of the system. The wreck would not have been displayed on the electronic chart due to the user settings in use at the time. A paper chart was available, but positions had only been plotted on it sporadically and it was not referred to at the crucial time.

Role of Data Although the Voyage Management System (VMS) was loaded with Electronic Navigational Charts (ENC) for the vessel's area of operation, the system had not been approved by the Maritime and Coastguard Agency (MCA) as the owner's policy was for the VMS to be used as an aid to navigation only, with PRIDE OF CANTEBURY's paper charts being utilised as the primary means for navigation. Relevant admiralty charts were supplied to the vessel for this purpose.

Although the VMS was not approved for use as the primary means of navigation, the officers on PRIDE OF CANTEBURY were apparently using it as if it was, despite the fact that many of them, including the Chief Officer, who was in charge at the time of the accident, were not fully trained in its use.

Among other things, this incident reflects shortcomings in the completeness (data property) of the justification and instructional data types.

Sources

- Report on the investigation into the grounding of Pride of Canterbury 'The Downs'- off Deal, Kent 31 January 2008, Marine Accident Investigation Branch, Report No 2/2009, January 2009.

A.8 LOT Flight 282

Summary Just after take-off from Runway 09R at London Heathrow Airport (LHR), the pilots noticed that most of the information on both of the Electronic Attitude Director Indicators (EADI) and Electronic

Horizontal Situation Indicators (EHSI) had disappeared. The aircraft entered Instrument Meteorological Conditions (IMC) at about 1,500 feet Above Aerodrome Level (AAL), and the co-pilot had no option but to fly using the standby attitude indicator and standby compass. He experienced difficulty in following radar headings. The aircraft returned to land at LHR after a flight of 27 minutes.

Role of Data The single error made by the co-pilot during the pre-flight preparation initiated the subsequent problems. This was the use of 'E' instead of 'W' when the longitude co-ordinates were entered into the FMS.

The airports around London, because of their proximity to the Prime Meridian, can lead flight crews to make such co-ordinate entry errors of this nature. It is of note that the operator's route network is such that there are few destinations to the west of the Prime Meridian and hence the majority of longitude co-ordinates that need to be entered would be eastings. Inertial Reference System (IRS) alignment warnings should have alerted the crew but may have been dismissed.

This incident relates to the accuracy of the data entered into the FMS.

Sources

- Air Accidents Investigation Branch, Bulletin 6/2008.

A.9 Cedars Sinai Medical Centre - CT Scanner

Summary A software misconfiguration in a Computed Tomography (CT) scanner used for brain perfusion scanning at Cedar Sinai Medical Center in Los Angeles, California, resulted in 206 patients receiving radiation doses approximately 8 times higher than intended. This error persisted for an 18 month period, starting in February 2008. Some patients reported temporary hair loss and erythema.

Role of Data The problem reportedly resulted from an error made by the hospital in resetting the CT machine after it began using a new protocol for the procedure in February 2008, but it wasn't detected until one of the patients reported patchy hair loss in August 2009.

“There was a misunderstanding about an embedded default setting applied by the machine,” according to a statement from Cedars-Sinai. “As a result, the use of this protocol resulted in a higher than expected amount of radiation.”

This incident reflects the importance of data verifiability, especially with regards to default (and adaptation) data.

Sources

- Los Angeles Times, http://articles.latimes.com/2009/oct/10/local/me-cedars-sinai10, last accessed 11/11/2015.
- HealthImaging, http://www.healthimaging.com/topics/diagnostic-imaging/update-cedars-sinai-explains-ct-perfusion-radiation-overexposure, last accessed 11/11/2015.

A.10 Fort Drum Artillery Incident

Summary Two artillery shells were fired more than a mile off target during an Army firing exercise at Fort Drum in Northern New York in March 2002. The shells landed near a mess tent where a Battalion were having breakfast. Two soldiers were killed, 13 were injured.

Role of Data The initial artillery site was unsuitable so the unit had to move nearly a mile from the initial site. The unit then had trouble setting up its digital and wire communications. The movement of the unit was not taken into account when programming the firing coordinates. Also, in what was termed a 'software behavioural shortfall' the system was designed to reset the gun elevation to zero. The correct altitude for the new site was not entered into the safety calculations, and the mistakes were not captured by the data review process.

This incident relates to the integrity (and possibly the accuracy) of the firing coordinates and gun elevation data.

Sources

- The New York Times, http://www.nytimes.com/2003/07/02/nyregion/officer-found-negligent-in-deaths-of-2-at-fort-drum.html, last accessed 11/11/2015.
- AP News Archive, http://www.apnewsarchive.com/2002/Army-Reports-on-Ft-Drum-Accident/id-539bf2ea24b8dd66009c6efee2be926c, last accessed 11/11/2015.

A.11 Quantas Boeing 737 Loading Incident

Summary On 9 May 2014 a Quantas Boeing 737 was preparing for departure from Canberra to Perth. There were 150 passengers, 87 of which were primary school children. These children were all seated together at the rear of the cabin. All had been assigned an 'adult weight' of 87 kg.

During take-off the aircraft appeared nose heavy. Significant back pressure was required to rotate the aircraft and lift off from the runway. The aircraft exceeded the calculated take-off safety speed by about 25 kt. The aircraft rose at a higher initial climb speed than usual, but the crew did not receive any warnings. No further issues were experienced during the flight.

Role of Data A 'name template' was completed by a travel agent on behalf of the school group. This group was travelling from Perth to Canberra and returning back to Perth. Despite being marked as mandatory, the “Gender Description” field in this template was left blank; options for this field were “Adult”, “Child” and “Infant”.

As per company procedures, two days before the Perth-Canberra leg of their journey this group was 'advance accepted' into the booking system. Since the fields recording the number of children and young passengers in the group were blank, the Customer Service Agent assumed all of the group were adults. No loading-related issues were experienced during this flight.

Two days before the return flight the group was again “advance accepted” as all adults. They were checked in at Canberra Airport and assigned seats at the rear of the aircraft. The load discrepancy caused the issues noted above.

Fortunately, there were no serious consequences. However, this incident demonstrates, once again, the importance of data. This includes checking mandatory fields are completed (i.e. the completeness data property, as well as taking opportunities to verify data as it progresses through a system.

Sources

- Loading Issue Involving a Boeing 737, VH-VZO, http://www.atsb.gov.au/publications/investigation_reports/2014/aair/ao-2014-088.aspx, last accessed 11/11/2015.

A.12 Dallas Hospital Ebola Incident

Summary On 26th September 2014, a Dallas hospital mistakenly sent home a man who had the Ebola virus having missed what would have appeared to be an obvious potential case: a Liberian citizen with fever and abdominal pain who said he had recently travelled from Liberia. He returned to the hospital, was eventually diagnosed with the illness, but subsequently died. Two nurses that had treated the man also contracted the virus but later recovered.

Role of Data There have been mixed reports on the cause of the problem, but what is clear is that external social phenomena such as the Ebola outbreak, which are outside the hospital's electronic health record (EHR) system and processes, can change the safety significance of data held in the EHR. If the importance of the data is not recognised and elevated appropriately in the support tools and processes, then the risk of unintended harm can increase.

This conclusion is reinforced by system vendors who are now updating their systems to reflect the Ebola crisis in light of the Dallas incident.

Sources

- NBC News, http://www.nbcnews.com/storyline/ebola-virus-outbreak/texas-hospital-makes-changes-after-ebola-patient-turned-away-n217296, last accessed 11/11/2015.

A.13 Interception of Communications

Summary In July 2015 it was reported that a public authority was undertaking an investigation into the uploading of indecent images of children and requested details of the account connected to the IP address used to upload the images. Issues with a new upgrade of the communication provider's system resulted in the incorrect data being disclosed. Investigations revealed that a further five requests had resulted in the incorrect data being disclosed. Data was acquired in six cases that related to individuals unconnected with the investigations. In one of these cases a welfare check was delayed on a child believed to be in crisis.

Role of Data Under the Regulation of Investigatory Powers Act 2000, Internet Service Providers and indeed other communication service providers (e.g. mobile phone network providers) are required to provide data to investigatory bodies such as the Police. This data can be used to support criminal investigation and prosecutions and in the protection of vulnerable children and adults. The data clearly has the potential to be safety related, but there is no obligation for data providers to treat it as such. In this case the data errors (i.e. loss of integrity) led to a child being exposed to additional risk of harm.

Sources

- I O C C O , http://www.iocco-uk.info/docs/2015%20Half-yearly%20report%20(web%20version).pdf, last accessed 26/10/2015.

A.14 Other Incidents

The following list highlights a number of other incidents where inappropriate data may have played a role:

- Xerox copier re-writing bug due to encoding/compression, http://www.dkriesel.com/en/blog/2013/0802_xerox-workcentres_are_switching_written_numbers_when_scanning, last accessed 11/11/2015;

- American Airlines Flight 965 at Buga, December 1995 ~ An accident due to inconsistent data naming conventions? - reference [10];
- The Provide Comfort Blackhawk Shootdown Incident - reference [11];
- Patriot Missile versus Tornado 2003 - reference [11].
- Enduring Freedom GPS Friendly Fire Incident - reference [11].

Appendix B Organisational Data Risk

> *Data is becoming the new raw material of business.*
> ***Craig Mundie***

This form is used to determine the safety risk related to data for a particular organisation and usage.

This form must be completed from the perspective of one of the organisations involved; typically this will be the organisation using the safety-related data or the contractor supplying the system that handles the safety-related data. This form needs to be completed for each instance / application / scope / risk profile and should consider a defined boundary for the analysis, e.g. the scope of supply for the contractor or the limit of the safety-related data user's operational responsibility. It may be useful for both contracting parties to complete the form from their respective positions to check the safety-related data risk responsibilities and apportionment.

It is anticipated that this form will be used during early phases of a procurement or supply and also for changes to existing supplies. It can also be used to assess existing legacy scenarios.

Answer the questions as they apply in the context of the scope of supply. Mark the response with the best fit for the given scenario. Note that not all elements have to be satisfied. For each response also add a brief justification for that particular selection as opposed to any other choice.

If the answer to a question is completely unknown at this stage; it is suggested that the middle value or higher is chosen and an explanation added to the justification.

When all the relevant questions have been answered and justified, add the scores together to give a final total and record the value in the appropriate field. Use this total to determine the final ODR level based on the stated ranges.

The ODR level determined may be used to determine the management regime required to mitigate the risk associated with the safety-related data.

Organisational Data Risk (ODR) Assessment Form

This form is used to determine the safety risk related to data for a particular organisation and usage.

*This form must be completed from the perspective of **one** of the organisations involved; typically this will be the organisation using the data or the contractor supplying the system that handles the data. This form needs to be completed for each instance / application / scope / risk profile and should consider a defined boundary for the analysis, eg. the scope of supply for the contractor or the limit of the data user's operational responsibility. It may be useful for both contracting parties to complete the form from their respective positions to check the data risk responsibilities and apportionment.*

It is anticipated that this form will be used during early phases of a procurement or supply and also for changes to existing supplies. It can also be used to assess existing legacy scenarios.

Answer the questions as they apply in the context of the scope of supply. Mark the response with the "best" fit for the given scenario. Note that not all elements have to be satisfied. For each response also add a brief justification for that particular selection as opposed to any other choice.

If the answer to a question is completely unknown at this stage; it is suggested that the middle value or higher is chosen and an explanation added to the justification.

When all the relevant questions have been answered and justified, add the scores together to give a final total and record the value in the appropriate field. Use this total to determine the final ODR level based on the stated ranges.

The ODR level determined may be used to determine the management regime required to mitigate the risk associated with the data.

Data Scenario/Context Name:			
Data Scenario/Context Description:			
Scope/Data Boundary and Perspective:			
Applicable Data Sets:			
Completed By:		Date Completed:	

Answer each question using the response that forms the best match for the particular scenario. Not all statements have to be satisfied and some judgment is required; it is expected that the majority of statements in the selected response can be satisfied with some interpretation. The use of multiple criteria in each question enables a smaller and manageable set of questions to be posed to provide a holistic view of the overall risk.

QUESTION 1 – SEVERITY AND PROXIMITY

How severe could an accident be that is related to the data? Could it be caused directly by the data?

This question considers the safety consequence, the proximity and contribution of the data to the accident sequence.

1a	All currently foreseen uses of the data could not contribute to an accident. The data is not relied upon for safe operation. Negligible environmental impact.	1	☐
1b	A possible use of data could contribute to a minor accident, but only via lengthy and indirect routes. Could lead to minor injury or temporary discomfort for 1 or 2 people. Many other people/systems are involved in checking the data Some aspects of safe operation rely very indirectly on the data. Minor environmental impact only via indirect routes.	2	☐
1c	A use of the data could lead to a significant accident resulting in minor injuries affecting several people or one serious injury. Several other people/systems are involved in checking the data. There is a dependency on the data for safe operation. Environmental impact is possible.	4	☐
1d	A likely use of the data could directly lead to a serious accident resulting in serious injuries affecting a number of people, or a single death. One human or independent check is involved for all data. There is major dependency on the data for safe operation. Major environmental impact possible.	8	☐
1e	An intended use of the data could easily lead to an accident resulting in death for several people. The accident could be caused by the data with little chance of anything else detecting and mitigating data issues. The accident could affect the general public or cause catastrophic environmental impact.	16	☐

Justification:

QUESTION 2 – ORGANISATIONAL AND SOCIETAL IMPACT

What would be the impact on the organisation, client or public if an accident occurred related to the data?

This question considers the tolerability within this industry sector and the general public. How much would it affect the organisation or society? Would a claim be likely? Would it generate press interest? Would a formal investigation ensue?

2a	Little interest, accidents happen all the time in this sector; very high societal tolerability. Negligible chance of claims or investigations. No adverse publicity likely.	1	☐
2b	Some concern from the client, but accidents happen occasionally; high societal tolerability. Small chance of claim against the organisation. Local or specialist press interest. Minor investigation or audit.	2	☐
2c	Public would be concerned, accidents are rare in this sector; some societal tolerability. Significant chance of claim against the organisation. Regional press interest. Client inquiry or investigation likely.	4	☐

2d	Public would be alarmed and consider the accident a result of poor practice; little societal tolerability. Claims very likely. National press or media coverage a possibility. Legal or independent inquiry may follow.	8	☐
2e	Public would be outraged and consider such an accident unacceptable; almost no societal tolerability. Multiple claims/fines from regulators or courts are likely. International press or media coverage. Official and/or public enquiry possible.	16	☐
Justification:			

QUESTION 3 – RESPONSIBILITY

How much responsibility does this organisation have for data safety?

This question considers how much legal and other responsibility and ownership the organisation has for data safety aspects within this scenario. What liabilities for consequential losses / 3rd party claims does the organisation have via the contract or other means? What is the scale of the organisation's contribution to the overall scope?

3a	The organisation is not responsible for any data safety aspects. No liabilities for accident claims related to the data lie with the organisation. Client or other party has accepted full data safety responsibility. The organisation is fully covered and indemnified by the client or a 3rd party.	1	☐
3b	The organisation is a small part of a large consortium. It has minimal liability for data safety via the contract. It is partly covered by explicit client or 3rd party protections. All safety data is managed by subcontractors, the organisation only reviews and monitors.	2	☐
3c	The organisation is a significant part of the consortium or team. It has some share of the data safety responsibility. Specific data safety liabilities to the client via the contract are mentioned. There are no indemnities in the organisation's favour. All key safety data obligations are explicitly flowed down to subcontractors.	4	☐
3d	The organisation is prime for a small programme or has the bulk of the data safety responsibility within a team. Specific accident-related liabilities in the contract are significant. The organisation provides some indemnities to others via the contract. Some significant data safety obligations are not flowed down to subcontractors.	7	☐
3e	The organisation is priming a major programme or has total data safety responsibility. Specific accident-related liabilities in the contract are large (or unlimited). The organisation provides explicit indemnities in favour of the client / 3rd parties for accidents. Safety data obligations have not been discussed or are not flowed down to subcontractors.	12	☐
Justification:			

QUESTION 4 – LEGAL AND REGULATORY FRAMEWORK

What legal and regulatory environment will this work be subject to?

This question considers the legal and regulatory obligations that this work will have to conform to. How well is the legal framework defined and understood? Is there an established standards culture? Is there a regulator and certification process?

4a	Well understood and tested legal framework, one jurisdiction. Highly regulated sector with one overseeing body. Well established industry guidelines and standards for safety data. Formal certification processes.	1	☐
4b	Understood and established legal framework, a few related jurisdictions. Regulated sector, more than one overseeing body. Industry guidelines and standards for safety data. Some formal certification processes.	2	☐
4c	Some understanding of legal position, several jurisdictions. Partially regulated sector, several possible overseeing bodies. Some industry guidelines and standards that refer to data. Informal certification processes.	4	☐
4d	Complex, poorly defined legal position, multiple different jurisdictions. Largely unregulated sector with no established overseeing body. Some industry guidelines and standards that mention data. Some informal certification processes.	6	☐
4e	Very complex, untested and unclear legal position, many diverse jurisdictions. Unregulated sector with no overseeing body. No industry guidelines or standards for data. No certification processes.	10	☐
Justification:			

QUESTION 5 – ORGANISATIONAL MATURITY

How mature is this organisation regarding data safety?

This question considers the maturity of the organisation in relation to awareness and management of the risks associated with safety data. Are staff trained, managed and resourced to enable proper handling of data safety risk?

5a	Explicit recognition of data as a source of safety risk. Formal and established processes and procedures in place for the identification and control of safety data. Staff trained and fully aware of safety data risks. Senior management fully aware and supportive of data safety management activities. Management of safety data risks fully supported and funded.	1	☐
5b	Awareness of data as a source of safety risk. Informal processes and procedures in place for the identification and control of safety data. Staff awareness of safety data risks. Senior management awareness of data safety management issues. Good support and funding for management of safety data risks.	2	☐
5c	Some awareness of data as a source of safety risk. Some ad-hoc processes and procedures in place for the identification and control of safety data. Some staff awareness of safety data risks. Some senior management awareness of data safety management issues. Some support or partial funding for management of safety data risks.	4	☐
5d	Little awareness of data as a source of safety risk. Minimal processes or procedures in place for the identification and control of safety data. Little staff awareness of safety data risks. Little senior management awareness of data safety management issues. Little support or minimal funding for management of safety data risks.	7	☐

5e	No recognition of data as a source of safety risk. No processes or procedures in place for the identification or control of safety data. No staff training or awareness of safety data risks. Senior management not aware or in denial of safety data risks. No support or funding for management of safety data risks.	10	☐
Justification:			

QUESTION 6 – OWNERSHIP AND USAGE

How widely is the data used and who by?

This question considers how much usage and what type of users there are likely to be of the data. How complex is the data supply chain? In what geographies is it used? How many owners and interfaces are there?

6a	Minimal or infrequent usage. One data owner, a specialist highly trained user group. Single organisation or recipient usage only.	1	☐
6b	A number of operational data users. Simple linear supply chain. More than one data owner. Specialist user or limited public access. Small scale operation. No general web access. Few user organisations or recipients.	2	☐
6c	Regional usage. Some public or mainstream usage. A few supply chains. A few data owners. Some web access. Several user organisations or recipients.	4	☐
6d	National usage. Public or mainstream usage. Several supply chains. Several data owners. Web access. Some or varied user organisations or recipients	7	☐
6e	International usage. Extensive public or mainstream usage. Extensive web access. Many complex supply chains. Many and diverse data owners. Many and diverse user organisations or recipients.	12	☐
Justification:			

QUESTION 7 – SIZE, COMPLEXITY AND NOVELTY

What is the scale, sophistication and complexity of the data and its manipulation?

This question considers the nature of the data, its lifecycle and how easy it is to detect errors in the data.

7a	Simple data structures. Mature and established data storage and manipulation techniques and technologies. One or two interfaces. No timeliness aspects. No transformations. Data is easily verifiable. Data is easily traceable to original source.	1	☐
7b	Varied data structures. Mainstream data storage and manipulation techniques and technologies. Several interfaces. Few timeliness aspects. Few data transformations. Data is verifiable. Data is traceable to original source.	2	☐
7c	Complex with some unstructured data. Current data storage and manipulation techniques and technologies. Multiple interfaces. Some timeliness aspects. Some data transformations. Data is difficult to verify. Data is difficult to trace back to original source.	4	☐
7d	Complex, varied or partially unstructured data. Novel storage and manipulation techniques and technologies. Multiple complex interfaces. Time critical. Complex data transformations. Data is very difficult to verify. Data is very difficult to trace back to original source.	7	☐
7e	Highly complex, varied or unstructured data. Highly novel storage and manipulation techniques and technologies. Many and complex, ill-defined or dynamic interfaces. Highly time critical. Many and complex data transformations. Data is infeasible to verify. Data is impossible to trace back to original source.	10	☐
Justification:			

QUESTION 8 – BOUNDARIES AND INTERFACES

How well defined and understood are the boundaries and interfaces for this data scenario?

This question considers the number, complexity and definition status of the boundaries and interfaces where data is exchanged. How well understood are the boundaries and interfaces? Are standard formats and protocols used? Is data exchange time critical? Are all assumptions and ambiguities relating to the data exchange resolved?

8a	One well-understood boundary and few, well-defined interfaces. Standard interface formats and protocols. No timeliness aspects to data exchange. No remaining ambiguities, TBCs or TBDs. No assumptions.	1	☐
8b	A few, understood boundaries and several defined interfaces. Mainly standard interface formats and protocols. Few timeliness aspects to data exchange. Few areas of ambiguity, few TBCs and TBDs. Few assumptions.	2	☐
8c	Several, established boundaries, some defined, some undefined and some ambiguous interfaces. Mixture of standard and non-standard interface formats and protocols. Some timely data exchanges. Some areas of ambiguity, some TBCs and TBDs. Some assumptions.	4	☐
8d	Many, poorly understood boundaries, many undefined or ambiguous interfaces. Mostly non-standard interface formats and protocols. Time sensitive data exchange. Many areas of ambiguity, many TBCs and TBDs. Many assumptions.	6	☐
8e	A large number of unclear boundaries; a large number of unknown and undefined interfaces. Completely non-standard, complex interface formats and protocols. Real-time data exchange. Large areas of ambiguity, a large number of TBCs and TBDs. A large number of assumptions.	10	☐
Justification:			

ORGANISATIONAL DATA RISK LEVEL

Record the total score and use it to determine the ODR level based on the ranges given below. If the first 3 question's scores sum up to 6 or less then disregard the scores for the remaining questions.	
Score 14 or less	ODR0
Score 15 to 21	ODR1
Score 22 to 37	ODR2
Score 38 to 47	ODR3
Score 48 and above	ODR4
Total Score for this scenario/context:	
ODR Level for this scenario/context:	

Warning: this form only gives an initial organisational data risk level assessment. Further work is required to establish the safety data risks in detail such as determining a Data Integrity Level (DIL) for relevant data sets.

Appendix C Data Safety Culture Questionnaire

> *Data is the fabric of the modern world: just like we walk down pavements, so we trace routes through data and build knowledge and products out of it.*
> ***Ben Goldacre***

This form helps an organisation appreciate the data safety culture. It can be applied at various levels, including at the project level and at the organisational level.

SCSC DSIWG

Data Safety Culture Questionnaire Form

This form is used to assess the safety culture related to data for a particular programme (the DSC value).

You play a key role in protecting the organisation from data safety risks and your views are important.
This self-assessment survey is designed to assess our current level of data safety culture within the programme.
The output can help us to improve our safety position.

Please tick the box which reflects your view and answer as honestly as possible. Space is provided for explanatory comments. Your response will only be of value if it reflects what you actually believe is the case, rather than what you believe should happen.

If you would like to remain anonymous please print and send this form by post.

The survey should take no longer than 10 minutes. It is anticipated that this form will be used on a regular basis (e.g. annually).

Programme Name:			
Completed By:		Date:	

Answer each question as you see it – there is no right answer!

QUESTION 1 – MY VIEW OF OUR SUPPLY

		Don't Know	Strongly Disagree	Disagree	Maybe	Agree	Strongly Agree
1a	I see data as an important factor in the safety of my programme.	☐	☐	☐	☐	☐	☐
1b	I am familiar with the safety aspects of our data.	☐	☐	☐	☐	☐	☐
1c	I think that data in our solution could contribute to an accident.	☐	☐	☐	☐	☐	☐
1d	I think we could be blamed if there were an accident due to our data.	☐	☐	☐	☐	☐	☐

Comments:

QUESTION 2 – WHAT WE'RE DOING

		Don't Know	Strongly Disagree	Disagree	Maybe	Agree	Strongly Agree
2a	I think that the programme is aware of data safety risks.	☐	☐	☐	☐	☐	☐
2b	I believe we need to implement measures to manage data safety risks.	☐	☐	☐	☐	☐	☐
2c	I think that the programme meets its obligations (e.g. has a Data Management Plan in place and a role with specific responsibilities in this area)	☐	☐	☐	☐	☐	☐

Comments:

QUESTION 3 – MY ROLE

		Don't Know	Strongly Disagree	Disagree	Maybe	Agree	Strongly Agree
3a	I know how my role relates to the management of data and associated safety risks.	☐	☐	☐	☐	☐	☐
3b	If I had a safety concern about our data I would report it.	☐	☐	☐	☐	☐	☐
3c	I know who the data safety representative is on my programme.	☐	☐	☐	☐	☐	☐
3d	I have received adequate training regarding data safety for my role	☐	☐	☐	☐	☐	☐
3e	I feel supported in dealing with data safety risks.	☐	☐	☐	☐	☐	☐
3f	I have adequate time to address any data safety issues.	☐	☐	☐	☐	☐	☐

Comments:

V0.1

Appendix D Dataware Framework Assessment

> *In God we trust; all others must bring data.*
> ***W. Edwards Deming***

D.1 Introduction

The use of the Dataware Framework is not restricted to information systems. It is intended to allow the expression of the respective safety integrity requirements of elements of the framework. A key aspect is recognition that there are insufficient resources for all components of the system to attain the highest integrity of any individual component. Therefore, the system and its respective safety functions are identified in terms of their scope (between layers) and height within the hierarchy.

It is common to identify high integrity functions that require fast acting rule based actuation deployed as close as practicable to the (safety) risk being managed. These high integrity functions should be clearly defined and bounded within the system so that we can have a high degree of confidence in their effective operation. As we rise up through the framework combinations of lower integrity safety functions are collected to form control and administrative activities (such as SCADA) are interfaces (boundaries) with wider corporate, enterprise or third party systems. Where these systems are data-centric, data errors may propagate across the system hierarchy and give rise to hazards within the hierarchy at boundaries between peer, supervisory or subordinate system components.

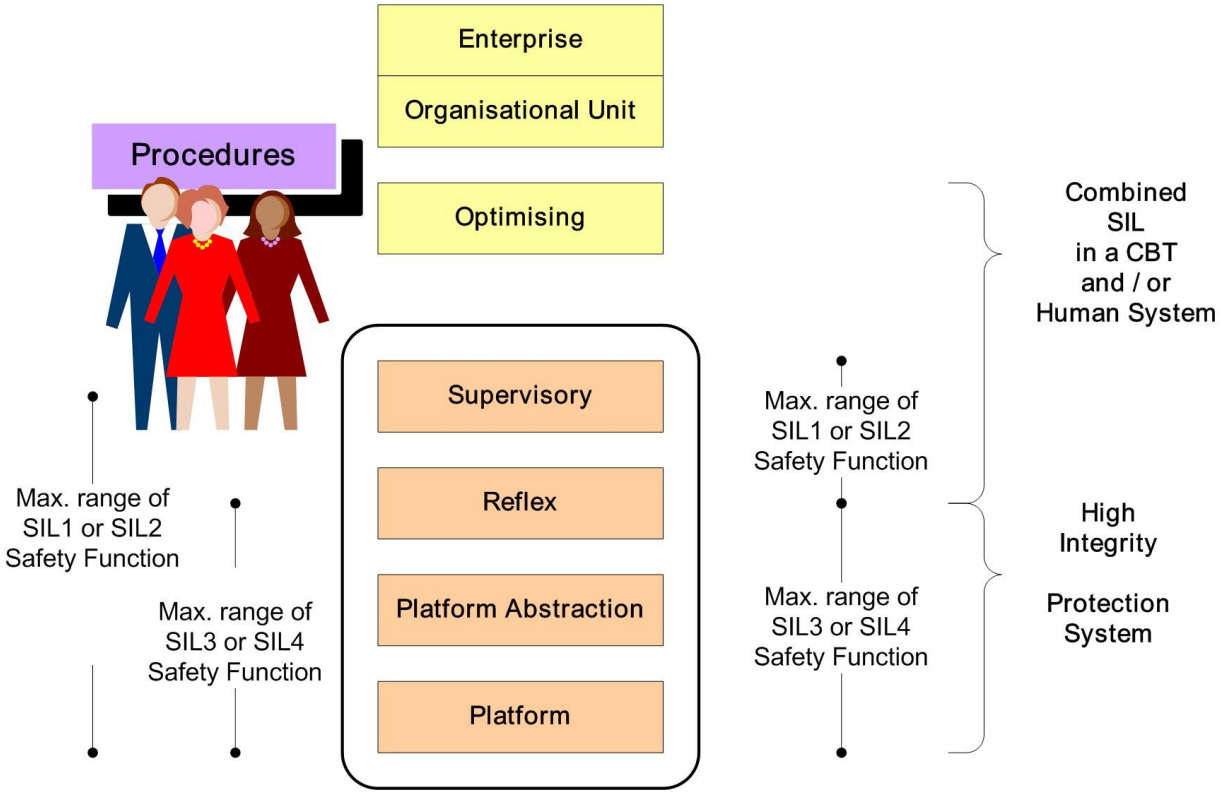
The data assessment framework is intended to allow the expression of Computer Based Technologies (CBT) from many domains to facilitate comparison – and ideally to enable the identification of a set of common techniques and measures. The framework will not address all systems or provide complete coverage of a single system. The rigour of the application of the framework is intended to follow a risk based approach of the greater the risk, the greater the rigour.

D.2 The Layered Model

At the framework's core is a layered model representing an Organisational Hierarchy with elements drawn from concepts contained within the 'Basic Reference Model for Open Systems Interconnection' (ISO OSI) model. The OSI model partitions communications services between 7 layers with defined interfaces, peer protocols that permit the separation of application development from the underlying communication system. Two important elements of the OSI model are that each layer:

- communicates with its peer layer in a different communication unit; and
- provides a service to the layer above and expects a particular kind of service from the layer below.

This abstraction into layers allows the development and replacement of the underlying layers based on respect for the services each layer provides, and preservation of interfaces between them. It also allows control of the access points of data and control coming into the system.



The above figure identifies a number of layers within the system hierarchy and also implies a functional hierarchy within the system.

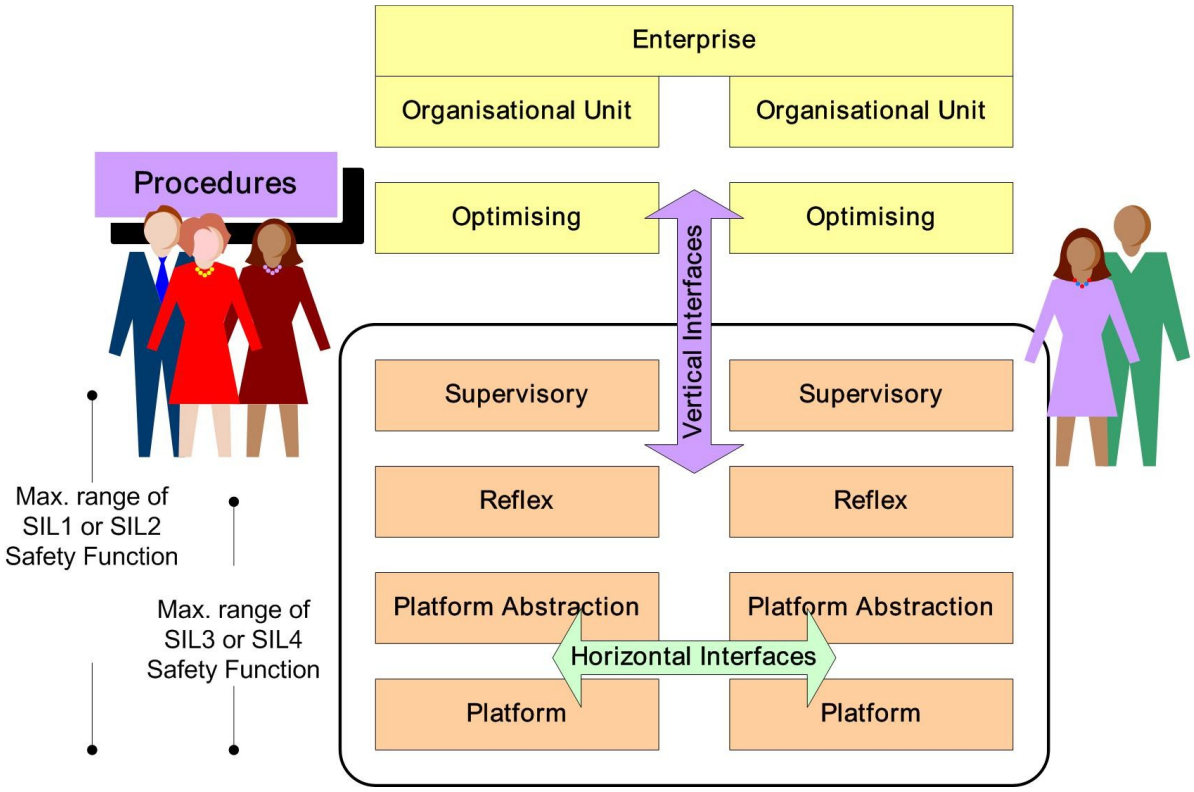
- The "**platform**" layer represents single instances of elements of the platform infrastructure, the physical equipment;
- The "**platform abstraction**" layer represents the interface to platform infrastructure elements. In essence this layer converts physical phenomenon from sensors (including feedback from actuators) into abstract representations such as electrical signals or data. This layer also provides the control (information) interfaces to actuators (human operators);
- The "**reflex**" layer is the lowest layer at which the measured status is interpreted and control (or protection) actions are carried out. These actions may be based upon data (which may include stored information), any demands upon the system and some set of rules. In this reflex layer the rules and information completely determine the control action. In principle all activities in the reflex layer can be automated with the highest levels of autonomy. Safety-critical functions commonly require a fast response and therefore often make use of reflex actions;
- The "**supervisory**" layer represents a more complex level of control. This complexity may be a result of large-scale operation, integrating a number of dissimilar functions, or interpreting complex (or ambiguous) data (or some combination of these). The distinction between the reflex and supervisory layers is the judgement or knowledge that must be applied, particularly in degraded or emergency situations. Supervisory systems are characterised by the need to support the judgement of the operator doing the supervision. Predominantly the supervisory layer is downward looking, viewing the performance of the lower levels;
- The "**optimisation**" layer represents the most sophisticated control layer. At its most developed, the optimisation layer should maximise the use of resources from the delivery of the service. The optimisation layer should respect the performance and safety constraints of the underlying

(transportation) system. The information demands on the optimisation layer are high, requiring a full understanding of the underlying system, the planned service and contingency plans;

- The "**organisational unit**" layer represents the organisational responsibility of the delivery of the planned service. This layer normally plays little part in real-time operations of the system being more concerned with the medium term maintenance (including competencies) and development of the infrastructure, and the subsequent future delivery of the planned service. The organisational unit will become involved in the short-term operation of the system in response to a serious incidents that cause substantial impact on the delivery of the service. Organisational unit is used here in order to provide a generic model; and
- The "**enterprise**" layer represents the corporate entity; responsible for the planning and execution of large-scale changes to the infrastructure; responding to changes in legislation; setting and maintaining standards, procedures and competency requirements.

The supervisory layer is typically the highest layer at which a safety function should be implemented. This boundary is depicted in by the box surrounding the platform, platform interface, reflex and supervisory layers.

Interfaces (both horizontal and vertical) need to be controlled, as they provide a means to control the upgrade and replacement of elements within the hierarchy. Relationships, and therefore interfaces, may also exist between enterprises. Data may need to chain across these interfaces.



D.3 Vertical Interfaces

The OSI model uses the concept of Service Delivery Unit (SDU), and Service Access Point (SAP). The interface between each layer of the OSI 7-layer model is realised through one or more SAP's. The delivery of the (functional) service is via process modules, SDU's. Systems require data of differing levels of abstraction for each of the layers in the framework. The lower levels of the framework require less

abstract, more physical data, whilst the higher levels require more complex abstract representations (of the same physical infrastructure components). The parallel drawn is that in each layer of the layered framework, data is presented and consumed by the adjacent layers. Data is transformed within each layer. Furthermore, each interface (Data) Access Point (DAP) and Data Transformation Unit (DTU) provide a means to specify the data framework, and form the basis for an independence argument.

D.4 Horizontal Interfaces

The definition of the system boundary is an essential step in the definition of the system. The boundary provides a demarcation between those components, which are within the system, and those, which are external. Communication across the system boundary requires the identification and definition of an interface description including but not limited to, the data passed across the interface. Each communication is one step in a data chain.

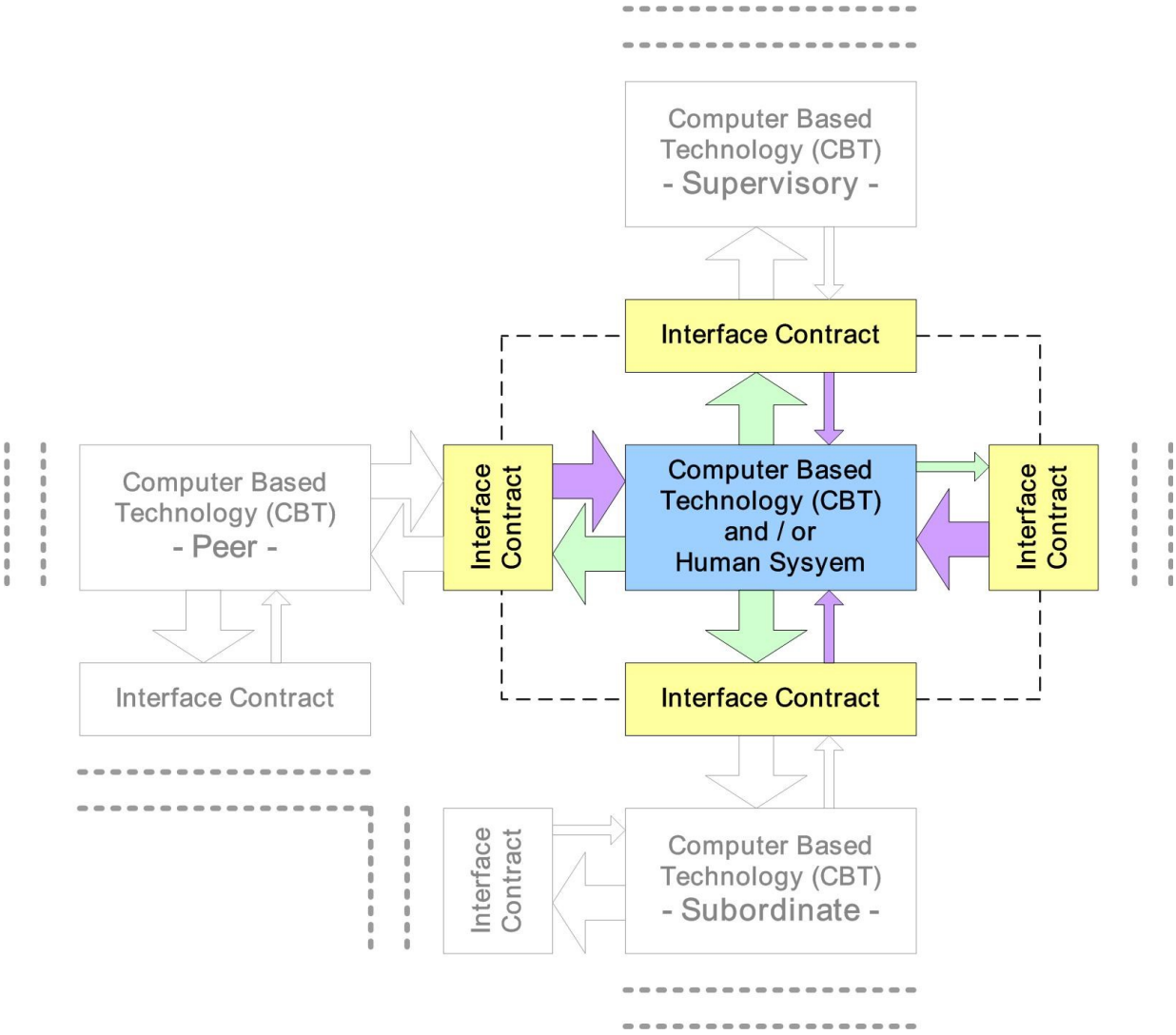
External information systems may provide a range of data including status and schedule data. Data presented at this system boundary may not be error free. Within a systems context these interfaces may contain implicit transformations between the external and internal use of data. In some cases the internal and external meanings of this data may also be different.

Data presented at the system boundary by an external system will be transformed or adapted from the external representation to the internal representation of the system. This transformation or adaptation may occur in some automated function or may require manual intervention. This data will also require verification. Analysis of the system design is required to establish the consequence of faults in this data as this data passes across the boundary of the control system. Further analysis should also establish the sensitivity to changes in this data. Such analysis will facilitate the definition of 'properties' or rules by which faults in the data may be detected at the boundary of the system.

The layered architecture may be partitioned into a number of independent applications sharing a network. In this case each partition can be thought of as a "module" in a larger system that allows appropriate flow of data via data chains and "big data" issues to be addressed. This partitioning holds throughout the vertical hierarchy of the framework, including at the organisational unit and where necessary the enterprise levels.

D.5 Interface Contracts

Whether interfaces are between layers (vertical), within layers (horizontal), as data flows within and between partitions, or to external sources of data Standardisation, Organisation and Control (SOC) will engender improved risk management of data. The SOC takes the form of a safety interface agreement. This can be implemented in a relatively informal manner or via more formal approaches using rely-guarantee interface contract formalisms. These contracts could, for example, define the level of data assurance the consumer demands for data passed to it across the interface.



D.6 Dataware Assessment Report Form

To promote consistency in the application of this assessment framework, a report template has been produced that contains the key topics that should be considered in the assessment. The report has the following structure:

1. Introduction
   - 1.1 Background
   - 1.2 Purpose
   - 1.3 Scope
2. System Description
   - 2.1 Context
   - 2.2 User Profile
   - 2.3 System Hierarchy

a) Supervisory Systems

b) Peer Systems

c) Subordinate Systems

2.4 Identified Interfaces

3. Data Used by the System

3.1 System Data

3.2 Interface A

3.3 Interface B

3.4 ...

4. Safety Assessment

4.1 Hazard Identification

4.2 Causality

4.3 Consequence

5. Integrity Requirements

A more comprehensive description of the Data Assessment Framework can be found in Faulkner and Nicholson [6]. The report template and examples of its use can be found on the DSIWG website.

D.7 History

The **Data Assessment Framework** is distilled from a number of models from a number of industry sectors. Primary amongst these is the standardisation of communications protocols into the ISO Model. The definition of communications into a defined set of layers, each with defined functionality and behaviours, has allowed the replacement on individual layers whilst retaining the overall functionality of communications between equipments. This feature is probably the single enabling factor that has driven the growth of mobile telephony.

The **Data Assessment Framework** provides a basis for abstraction; the addition of guidance and best practice provides a basis for the balance resources between function, timescale, responsiveness and crucially safety. Safety may seem odd bedfellow in a world driven by cost and time. Safety is not availability, nor is it reliability, although it does share some features and measures. Primarily safety is concerned with the management of risk of harm to the individual, assets and the environments. These concepts can also be expanded to operational issues, and although not primarily concerned with efficiency, a common feature of well run organisations is their 'good' safety record.

The need to manage interfaces between high integrity systems most marked in the aviation sector where a number of computational platforms share their resources across a number of applications (and systems). This is Integrated Modular Avionics (IMA) approach is resilient to hardware failures on individual CPU platforms. The formalisation of the definition of IMA has required the definition and management of 'interface contracts' as the formal specification of functionality, timing, recovery mechanisms and – data' between two or more parties.

The are but two areas where the use of layers, and interfaces has been documented to provide the basis for analysis – in this specific case 'safety analysis'. It is clear that a large scale system will consist of many thousands of interconnected systems. These systems of systems will use data as their primary means share information as information produces and or consumers. The Data Assessment Framework is intended as a means to document what data is exchanged. Analysis of these definitions will allow measures and techniques to be developed to identify, control and manage the propagation of data errors across these interconnected systems. Errors produced (at the Data Origin), may be transported across a number of systems (a Data Supply Chain) before presentation at the interface of a consuming system. Without the Data Assessment Framework it is difficult to demonstrate that the architecture, design, or its (safety) analysis is complete.

Appendix E Data Safety Management Plan

> *Things get done only if the data we gather can inform and inspire those in a position to make a difference.*
> ***Mike Schmoker***

This section gives a suggested Data Safety Management Plan (DSMP) table of contents. It is expected that this will be needed only for aspects not already covered in a Safety Management Plan (SMP), or similar. It can be merged with an SMP, if appropriate. However it may be useful to consider the distinct data perspective by using a DSMP as well as an SMP. Regardless, a close connection should be maintained between the SMP and the DSMP.

Data Safety Management Plan suggested contents:

1. Introduction:
   - Scope & Context (Sets the scene, describes the project, scenario, concept of operations, etc.);
   - Boundaries & Interfaces (Describes the main interfaces and exchanges, with a scope boundary diagram.);
   - Owners (Who owns the data under consideration as it progresses through the system?);
   - Producers / Consumers (Who are the producers and consumers of the data the system inputs and outputs?);
   - Assumptions;
   - References;
   - Abbreviations and Acronyms.
2. Analysis of Assigned DIL & ODR Level (Implications of the data analyses. Note this assumes the DIL and ODR analyses have already been performed.):
   - System Integrity Level (SIL), etc., Implications (What impact does the DIL have on the required SIL - or similar measure?);
   - Development Implications (Are there any special development considerations? Derived from the SIL if there is one, otherwise what is deemed necessary for this system.);
   - Verification Implications (Derived from the SIL if there is one, otherwise what is deemed necessary for this system.);
   - Assurance Implications (Derived from the SIL if there is one, otherwise what is deemed necessary for this system.);
   - Process / Procedure Implications (Derived from the SIL if there is one, otherwise what is deemed necessary for this system.).
3. Types of Safety Data in Scope (A list of all the types to be considered in the system context.).
4. Data Requirements Analysis:
   - Lifecycles (What data lifecycles are to be used?);

- Specific Targets (Are there any qualitative or quantitative targets for the data?);
- Security Considerations (How will security be managed in this context? Are there any security/safety conflicts? Are there any security-related causes of data hazards?).

5. Management Approach (How will the organisation manage the data safety risks?):
   - Organisation;
   - Responsibilities;
   - Authorisations;
   - Approvals and Signoffs.
6. Justification Approach (How will the safe usage of the data be justified, e.g. as part of the Safety Case Report?).
7. Analyses/Verifications to be Performed (What analyses or checks are to be performed on the data?).
8. Documents to be Produced (The list of documents to be produced related to data aspects.).
9. Appendix: DIL Guidelines Response (Tailored version of the tables from this document. What is considered applicable/useful and what is not?).

Appendix F Definitions, Acronyms & Glossary

> *The plural of anecdote is not data.*
> ***Mark Bekoff***

	Definition	Source
A		
Accuracy	Closeness of agreement between a test result and the accepted reference value. NOTE A test result can be observations or measurements.	ISO 19113:2005 [12]
	A degree of conformance between the estimated or measured value and the true value	(EU) No 73/2010 [13]
Accuracy (temporal)	Correctness of the temporal references of an item (reporting of error in time measurement). Correctness of ordered events or sequences, if reported. Validity of data with respect to time.	ISO 19138:2006 [14]
Active (data)	Data that changes system functionality	DSIWG
(data) Assurance Level	The required assurance level for the aeronautical data process is identified, based on the overall system architecture through allocation of risk determined using a preliminary system safety assessment.	RTCA/DO-200A [15]
	An indication of how much assurance is required (commensurate to risk) before deploying software into an operational system	J Spriggs, based on (EC) No 482/2008 [16]
Adaptation Data	Data used to customise elements of the Air Traffic Management System for their designated purpose. Adaptation data is utilised to customize elements of the CNS/ATM system for its designated purpose at a specific location. These systems are often configured to accommodate site-specific characteristics. These site dependencies are developed into sets of adaptation data. Adaptation data includes data that configures the software for a given geographical site, and data that configures a workstation to the preferences and/or functions of an operator. Examples include, but are not limited to: a) Geographical Data - latitude and longitude of a radar site. b) Environmental Data - operator selectable data to provide their specific preferences. c) Airspace Data - sector-specific data. d) Procedures - operational customization to provide the desired operational role. Adaptation data may take the form of changes to either database parameters or take the form of pre-programmed options. In some cases, adaptation data involves re-linking the code to include different libraries. Note that this should not be confused with recompilation in which a completely new version of the code is generated.	ED-153 [17]
Aeronautical Data	A representation of aeronautical facts, concepts or instructions in a formalised manner suitable for communication, interpretation or processing	(EU) No 73/2010 [13]
	Data used for aeronautical applications such as navigation, flight planning, flight simulators, terrain awareness, and other purposes.	RTCA/DO-178C [18]

	Definition	Source
Application Data	Data used in the system during operations: this is the data processed or produced by the system which has end-user meaning. It may be displayed and used within the system or may be for transfer or distribution to other systems or downstream users. It is data that has some real â€œapplicationâ€ meaning, i.e. is not to do with the system internals.	SCSC Data Safety Initiative Working Group
Assumption Data	Data used to frame the development, operations or provide context: restrictions, risk criteria, usage scenarios, etc. explaining how the system will be used and any limitations of use	SCSC Data Safety Initiative Working Group
Availability	The property of being accessible and usable upon demand by an authorized entity	ISO27001:2005 [19]
B		
C		
Completeness	Completeness of the data provided	RTCA/DO-200A [15]
Configuration Data	Data that configures a generic software system to a particular instance of its use.	(EC) No 482/2008 [16]
Configuration Data	Data used to configure, tailor or instantiate the system: data used to set up and configure the system to perform a particular function, for a particular installation, product configuration, behaviour or specific usage	SCSC Data Safety Initiative Working Group
Configuration Data	Data that configures a generic software system to a particular instance of its use (for example, data for flight data processing system for a particular airspace, by setting the positions of airways, reporting points, navigation aids, airports and other elements important to air navigation)	ED-153 [17]
Confidentiality	The property that information is not made available or disclosed to unauthorized individuals, entities, or processes	ISO27001:2005 [19]
(data) correctness	Completeness, self consistency, protection against alteration or corruption and consistency with the functional requirements of the data driven system	IEC 61508 Part 3 [20]
(data) coupling	The dependence of a software component on data not exclusively under the control of that software component	RTCA/DO-178C [18]
(data) Criticality	Classification of data by the potential effect of erroneous data on the expected operation that is supported by that data.	RTCA/DO-200A [15]
Critical Data	Data with an integrity level as defined in Chapter 3, Section 3.2 point 3.2.8(a) of Annex 15 to the Chicago Convention, i.e. integrity level one in one hundred million: there is a high probability when using corrupted critical data that the continued safe flight and landing of an aircraft would be severely at risk with the potential for catastrophe.	(EU) No 73/2010 [13]
Customisation (data)	Data used to configure a system or component	Def(Aust)5679 [21]
D		
Data	A thing given or granted; something known or assumed as fact, and made the basis of reasoning or calculation; an assumption or premiss from which inferences are drawn.	Oxford English Dictionary (OED)
	A reinterpretable representation of information in a formalized manner suitable for communication, interpretation or processing	ISO/IEC 2382 [22]

	Definition	Source
Data (Aeronautical)	A representation of aeronautical facts, concepts or instructions in a formalised manner suitable for communication, interpretation or processing	(EU) No 73/2010 [13]
	Data used for aeronautical applications such as navigation, flight planning, flight simulators, terrain awareness, and other purposes.	RTCA/DO-178C [18]
Database	A set of data, part or the whole of another set of data, consisting of at least one file that is sufficient for a given purpose or for a given data processing system.	RTCA/DO-178C [18]
Data Chain	An 'Aeronautical Data Chain' is a conceptual representation of the path that a set, or element of aeronautical data takes from its creation to its end use. An aeronautical data chain is a series of interrelated links wherein each link provides a function that facilitates the origination, transmission and use of aeronautical data for a specific purpose.	RTCA/DO-200A [15]
	A collection of organisational data processing functions, where data is transferred from one chain participant to another between data origination and end use.	P. Ensor [23]
	Any combination of two or more data elements, data items, data codes, and data abbreviations in a prescribed sequence to yield meaningful information; for example, 'date' consists of data elements year, month, and day.	McGraw-Hill Dictionary [24]
Data Chain Participant	A key organisational/functional element within a data supply chain.	P. Ensor [23]
(data) dictionary	The detailed description of data, parameters, variables, and constants used by the system.	RTCA/DO-178C [18]
Data Driven Systems	System which relies upon configuration data or lookup tables to define the functionality of the system.	IEC 61508 Part 4 [25]
Data Intensive System	Systems which make extensive use of large amounts of data	N. Storey [26]
Design & Development Data	Data produced during development and implementation: this is data encompassing the design & development process artefacts: everything from design models and schemas to document review records. It also includes test documents (specification and results) but not the test data itself	SCSC Data Safety Initiative Working Group
E		
End of Life Data	Data about how to stop, remove, replace or dispose of the system: this is data covering all activities related to taking the system out of service or mothballing / storage / dormant phases	SCSC Data Safety Initiative Working Group
(data) Error	Discrepancy with the universe of discourse	ISO 19138:2006 [14]
	Discrepancy between a data value and the true, specified or theoretically correct value or condition.	P. Ensor [23]
Essential Data	Data with an integrity level as defined in Chapter 3, Section 3.2 point 3.2.8(b) of Annex 15 to the Chicago Convention, i.e. integrity level one in one hundred thousand: there is a low probability when using corrupted essential data that the continued safe flight and landing of an aircraft would be severely at risk with the potential for catastrophe.	(EU) No 73/2010 [13]

	Definition	Source
Evolution Data	Data about changes after deployment, i.e. data that cover enhancements, formal changes, workarounds, and maintenance issues. It also covers data produced by configuration management activities, such as baselines or branch data	SCSC Data Safety Initiative Working Group
F		
G		
H		
(data) Hazard	Use of data in the context of a system that could lead to an accident	DSIWG
I		
Information	Knowledge communicated concerning some particular fact, subject, or event; that of which one is apprised or told - intelligence, news - as contrasted with data.	Oxford English Dictionary (OED)
	Knowledge that has a contextual meaning	ISO/IEC 2382 [22]
Information (aeronautical)	information resulting from the assembly, analysis and formatting of aeronautical data	(EU) No 73/2010 [13]
(data) Integrity	The assurance that a data element retrieved from a storage system has not been corrupted or altered in any ways since the original data entry or latest authorised amendment	RTCA/DO-200A [15]
	The degree of assurance that a data item and its value have not been lost or altered since the data origination or authorised amendment	(EU) No 73/2010 [13]
	The degree of undetected (at system level) non-conformity of the input value of the data item with its output value	(EU) No 1207/2011 [27]
	The property of protecting the accuracy and completeness of assets, i.e. that which has value to the organisation	ISO27001:2005 [19]
Instructional Data	Data used to warn, train or instruct users about the system: this is data that explains to users the risks of the systems and gives any mitigations that may be required to be implemented by users, e.g. by process, procedure, workarounds, limitations of use	SCSC Data Safety Initiative Working Group
Intent (data)	Data describing how a system will behave	J. Inge [1]
(data) item	Single attribute of a complete data set, which is allocated a value that defines its current status	(EU) No 73/2010 [13]
Interface Data	Data used to enable interfaces between systems: for operations, initialisation or export from the system: data that exists to enable exchange between systems. Covers start-of-life operations (data import or migration), end-of-life operations and ongoing operational exchange of data between systems.	SCSC Data Safety Initiative Working Group
Investigation Data	Data to support accident or incident investigations (i.e. potential evidence: this is data collected or produced during an accident investigation which may be used in investigation reports, lessons learnt or prosecutions. This can be source data (e.g. photographs of crash site) or may be derived (accident simulations, analyses, etc)	SCSC Data Safety Initiative Working Group
J		
Justification Data	Data used to justify the safety position of the system: data used to justify, explain and make the case for starting or continuing live operations and why they are safe enough. Often passed to external bodies (regulators, HSE, ISAs) for their review.	SCSC Data Safety Initiative Working Group

	Definition	Source
K		
L		
M		
Meta-data	Data that represents information about data itself. Note: One should distinguish between "Structural Meta-data", which is data about the design and specification of data structures (and is more properly called "data about the containers of data") and "Descriptive Meta-data", which is about individual instances of application data, the data content.	J. Inge [1]
N		
O		
Objective (data)	Data describing a system's environment	J. Inge [1]
(data) Originator	Entity responsible for data origination	(EU) No 73/2010 [13]
Operational Data	Data collected or produced about the system during trials, pre-operational phases and live operations: data produced by and about the system during introduction to service and live service itself. Includes fault data and diagnostic data. This may be the results of various phases of introduction and may include trend analysis to look for long-term problems.	SCSC Data Safety Initiative Working Group
(data) Origination	Creation of a new data item with its associated value, the modification of the value of an existing data item or the deletion of an existing data item	(EU) No 73/2010 [13]
P		
Passive (data)	Data acquired from records collected for some other purpose.	online medical dictionary
(data) Product	Dataset or dataset series that conforms to a data product specification.	BS EN ISO 19131:2008 [28]
Prediction Data	Data used to model or predict behaviours and performance: Data for studies, models, prototypes, initial risk assessments, etc. This is the data produced during the initial concept phase which subsequently flows into further development phases	SCSC Data Safety Initiative Working Group
Q		
(data) Quality	A degree or level of confidence that the data provided meet the requirements of the user. These requirements include levels of accuracy, resolution, assurance level, traceability, timeliness, completeness, and format	RTCA/DO-200A [15]
	Process by which the Electronic Chart Systems (ECS) Database is produced, the source materials, the resolution and reproduction accuracy of chart features, and the correctness and completeness of data.	ISO 19379:2003 [29]
	A degree or level of confidence that the data provided meets the requirements of the data user in terms of accuracy, resolution and integrity	(EU) No 73/2010 [13]
(data) Quality Attributes	Accuracy, resolution, assurance level, traceability, timeliness, completeness and format	RTCA/DO-200A [15]
R		

	Definition	Source
Release Data	Data used to ensure safe operations per release instance: explanation of particular features or limitations of a release or instance. May include specific time-limited workarounds and caveats for a release.	SCSC Data Safety Initiative Working Group
Requirements Data	Data used to specify what the system has to do: data encompassing requirements, specifications, internal interface or control definitions, data formats, etc.	SCSC Data Safety Initiative Working Group
Resolution	The smallest difference between two adjacent values that can be represented in a data storage, display or transfer system	RTCA/DO-200A [15]
	A number of units or digits to which a measured or calculated value is expressed and used	(EU) No 73/2010 [13]
Routine Data	Data with an integrity level as defined in Chapter 3, Section 3.2 point 3.2.8(b) of Annex 15 to the Chicago Convention, i.e. integrity level one in one thousand: there is a very low probability when using corrupted routine data that the continued safe flight and landing of an aircraft would be severely at risk with the potential for catastrophe.	(EU) No 73/2010 [13]
S		
(Data) Set	Identifiable collection of data. NOTE A dataset may be a smaller grouping of data which, though limited by some constraint such as spatial extent or feature type, is located physically within a larger dataset. Theoretically, a dataset may be as small as a single feature or feature attribute contained within a larger dataset. A hardcopy map or chart may be considered a dataset.	BS EN ISO 19131:2008 [28]
Software Lifecycle Data	Data that is produced during the software lifecycle to plan, direct, explain, define, record, or provide evidence of activities (including the software product itself). This data enables the software lifecycle processes, system or equipment approval and post-approval modification of the software product.	ED-153 [17]
Staffing and Training Data	Data related to staff training, competency, certification and permits: data which allows staff to perform a function within the wider context of the safety-related system. This may include training records, competency assessments, permits to work, etc.	SCSC Data Safety Initiative Working Group
Standards and Regulatory Data	Data that governs the approaches, processes and procedures used to develop safety-related systems: this is data predominantly in the form of documents that describe and dictate the activities, processes, competencies etc. to be used for a particular development in a particular sector.	SCSC Data Safety Initiative Working Group
System Data	Data about the installed or deployed system and its parts, including maintenance data : data related to location, condition and maintenance requirements of the system under consideration. This may cover hardware, software and data.	SCSC Data Safety Initiative Working Group
T		
Third Party (data)	Data of no direct relevance to a system	J. Inge [1]
Timeliness	A measure of the time delay between a change in the real world and the associated database update being available to the user.	P. Ensor [23]
	The difference between the time of output of a data item and the time of applicability of that data item	(EU) No 1207/2011 [27]
Traceability	Ability to determine the origin of the data	RTCA/DO-200A [15]

	Definition	Source
Trace (data)	Data providing evidence of traceability of development and verification processes software life cycle data without implying the production of any particular artifact.Trace data may show linkages, for example, through the use of naming conventions or through the use of references or pointers either embedded in or external to the software life cycle data.	RTCA/DO-178C [18]
U		
V		
(data) validation	The activity whereby a data element is checked as having a value that is fully applicable to the identity given to the data element, or a set of data elements that is checked as being acceptable for their purpose	RTCA/DO-200A [15]
	Process of ensuring that data meets the requirements for the specified application or intended use	(EU) No 73/2010 [13]
validity (period of)	Period between the date and time on which aeronautical information is published and the date and time on which the information ceases to be effective	(EU) No 73/2010 [13]
(data) verification	Evaluation of the output of an aeronautical data process to ensure correctness and consistency with respect to the inputs and applicable data standards, rules and conventions used in that process	(EU) No 73/2010 [13]
Verification Data	Data used to test and analyse the system: this is data comprising the test values and test data sets used to verify the system. It may include real data, modified real data or synthetic data. It includes data used to drive stubs, and any data files used by simulators or emulators.	SCSC Data Safety Initiative Working Group
W		
X		
Y		
Z		

Appendix G References

> *The goal is to turn data into information, and information into insight.*
> ***Carly Fiorina***

[1] Improving the Analysis of Data in Safety-Related Systems, James Inge, 12 September 2008, http://www.safety.inge.org.uk/20080912-Inge2008b_Improving_the_Analysis_of_Data_in_Safety_Related_Systems-U.pdf

[2] The Principles of Software Safety Assurance, R. Hawkins, I. Habli, T. Kelly, 31st International System Safety Conference, Boston, Massachusetts USA, 2013

[3] ISO31000:2009, Risk Management - Principles and Guidelines. First Edition, 2009-11-15

[4] Die Lage der IT-Sicherheit in Deutschland 2014, Bundesamt fuer Sicherheit in der Informationstechnik, December 2014, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Lageberichte/Lagebericht2014.pdf

[5] Data Integrity - an often-ignored aspect of safety systems, Alastair Faulkner, 2004, (EngD thesis), http://wrap.warwick.ac.uk/1212/

[6] An Assessment Framework for Data-Centric Systems, A. Faulkner, M. Nicholson. In Addressing Systems Safety Challenges, Proceedings of the Twenty-second Safety-Critical Systems Symposium, Bristol, UK. Edited by Chris Dale and Tom Anderson, ISBN 978-1491263648

[7] RTCA/DO-330, EUROCAE Document ED-215, Software Tool Qualification Considerations, January 2012

[8] Mars Climate Orbiter Mishap Investigation Board Phase I Report November 10, 1999, http://sunnyday.mit.edu/accidents/MCO_report.pdf

[9] Disastercast Episode 15 Quantitative Nonsense [including Mars Climate Orbiter], Drew Rae, http://disastercast.co.uk/transcripts/episode-15-transcript/

[10] American Airlines Flight 965, http://en.wikipedia.org/wiki/American_Airlines_Flight_965

[11] Disastercast Episode 18 Friendly Fire [including Data Safety], Drew Rae, http://disastercast.co.uk/transcripts/episode-18-transcript/

[12] BS EN ISO 19113:2005, Geographic Information. Quality Principles

[13] Commission Regulation (EU) No 73/2010 of 26 January 2010 laying down requirements on the quality of aeronautical data and aeronautical information for the single European sky, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:023:0006:0027:EN:PDF

[14] ISO/TS 19138:2006, Geographic Information. Data Quality Measures

[15] RTCA/DO-200A, EUROCAE Document ED-76, Standards for Processing Aeronautical Data, September 1998

[16] Commission Regulation (EC) No 482/2008 of 30 May 2008 establishing a software safety assurance system, as amended, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:141:0005:0010:EN:PDF

[17] EUROCAE Document ED-153, Guidelines for ANS Software Safety Assurance - use for definitions only

[18] RTCA/DO-178C, EUROCAE Document ED-12C, Software Considerations in Airborne Systems and Equipment Certification, January 2012

[19] BS ISO/IEC 27001:2005, Information Technology. Security Techniques. Information Security Management Systems. Requirements

[20] BS EN 61508-3:2010, Functional safety of electrical/electronic/ programmable electronic safety-related systems. Software Requirements, June 2010

[21] DEF(AUST)5679, Issue 2, Safety Engineering for Defence Systems - Standard, October 2008

[22] ISO/IEC 2382-1:1993, Information Technology. Vocabulary. Part 1: Fundamental Terms

[23] Safety Analysis of Navigational Data, Paul Ensor, September 2009

[24] McGraw-Hill Dictionary of Scientific and Technical Terms, 6th Edition, November 2002. ISBN-10: 007042313X

[25] BS EN 61508-4:2010, Functional safety of electrical/electronic/ programmable electronic safety related systems. Definitions and Abbreviations, June 2010

[26] The Characteristics of Data in Data-intensive Safety-related Systems, Neil Storey & Alastair Faulkner. Lecture Notes in Computer Science, Volume 2788, 396-409, 2003

[27] Commission Implementing Regulation (EU) No 1207/2011 of 22 November 2011 laying down requirements for the performance and the interoperability of surveillance for the single European sky, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:305:0035:0052:EN:PDF

[28] BS EN ISO 19131:2008, Geographic Information. Data Product Specifications

[29] BS ISO 19379:2003, Ships and Marine Technology. ECS databases. Content, Quality, Updating and Testing

Appendix H DSIWG History

> *If we have data, let's look at data. If all we have are opinions, let's go with mine.*
> ***Jim Barksdale***

The task of developing generally applicable, pan-sector guidance for data safety issues was taken on by the Data Safety Initiative Working Group (DSIWG) of the UK Safety Critical Systems Club.

The DSIWG's work started with a seminar "*How to Stop Data Causing Harm*", which was held in December 2012. The first meeting of the group agreed the following vision:

To have clear guidance on how data (as distinct from software and hardware) should be managed in a safety-related context, which will reflect emerging best practice.

The group, comprising industry, academic, government and independent consultants, produced an initial guidance document in January 2014. A subsequent version of the document was released in January 2015, with a second seminar "*How to Stop Data Causing Harm: What You Need to Know*" being held in December 2015.

This new version of the guidance document, issued February 2016, has incorporated the latest thinking in this developing area.

Appendix I Contributors

> *Without data, you're just another person with an opinion.*
> ***W. Edwards Deming***

This document has has the benefit of contributions from a large number of people, who work for a variety of organisations, which collectively span a range of different sectors. Note that contributions have been made on an individual basis and, in particular, the inclusion of an organisation in the following list does **not** necessarily mean that organisation agrees with the entire contents of the document.

Significant contributors to the document include:

- Mike Ainsworth, Ricardo
- Rob Ashmore, Dstl
- Janette Baldwin, Thales UK
- Dave Banham, Rolls-Royce plc
- Ian Bingham, CGI UK
- John Bragg, MBDA UK Ltd
- Eric Bridgstock, Raytheon UK
- Simon Brown, QinetiQ
- Dale Callicott, DKCSC Ltd
- John Carter, General Dynamics
- Martyn Clarke, RPS
- Duncan Dowling, DARD
- Andrew Eaton, CAA
- Paul Ensor, Boeing Defence UK Ltd
- Alastair Faulkner, Abbeymeade
- Derek Fowler, JDF Consultancy
- Ken Frazer, KAF
- Ian Glazebrook, Atkins
- Rob Green, NATS
- Nick Hales, MOD
- Amira Hamilton, CGI UK and Cranfield University
- Paul Hampton, CGI UK
- Ali Hessami, Vega Systems

- David Higgins, Dstl
- Pete Hutchison, RPS
- Gavin Jones, Raytheon Systems Ltd
- Tim Kelly, University of York
- Andrew Kent, Thales UK
- Brent Kimberley, Regional Municipality of Durham, Canada
- Julian Lockett, Frazer-Nash Consultancy Ltd
- David Lund, David Lund Consultants
- Dave Lunn, Thales UK
- Nasser Al Malki, University of York
- Mark Nicholson, University of York
- Robert Oates, Rolls-Royce plc
- Mike Parsons, NATS
- David Perrin, Virtual PV
- Andrew Rankine, NATS
- Felix Redmill, SCSC
- Sam Robinson, EDF Energy
- Tim Rowe, EC Harris
- Alan Simpson, Ebeni
- Dave Smith, Frazer-Nash Consultancy Ltd
- John Spriggs, NATS
- Mark Templeton, QinetiQ
- Lesley Winsborrow, EDF Energy
- Fan Ye, ESC

Appendix J Acknowledgements

> *Our ability to do great things with data will make a real difference in every aspect of our lives.*
> ***Jennifer Pahlka***

The document contributors would like to thank:

1. The Safety Critical Systems Club for support and encouragement.
2. Brian Jepson of the SCSC for web hosting support and technical help with the SCSC web site.
3. Chris Tapp (Keylevel Consultants Limited) for his assistance in the production of this document.
4. All the organisations that have hosted (or will be) hosting working meetings: Boeing, CAA, CGI, Ebeni, EDF, Frazer-Nash Consultancy, MOD, NATS, QinetiQ, Rolls-Royce plc, Thales, UKHO, University of Nottingham, University of York.
5. All the organisations that have provided support to the document's contributors.
6. Those that have been unable to attend meetings but have made supporting contributions.

Data as opposed to software or hardware, has been a contributing factor in many accidents and incidents. However, data in safety related systems is not sufficiently addressed by current safety management practices and standards.

There are clear business and societal benefits in terms of reduced harm, reduced commercial liabilities and improved business efficiencies, in investigating and addressing outstanding challenges related to the safety of data.

This book provides clear guidance on how data should be managed in a safety-related context. This work is by the Data Safety Initiative Working Group (DSIWG) - a cross-sector industry group set up to address the challenges of data safety.