SCSC
FOR EVERYONE WORKING IN SYSTEM SAFETY

DSIWG

SCSC Data Safety Initiative – WG Meeting 71

29th September 2022, Bath, UK & Zoom

Minutes

Attendees

Paul Hampton (PH) – CGI, Oscar Slotosch (OS) – Validas, Andy Williams (AW) – Consultant, Dave Banham (DB) – Blackberry, Richard Garrett (RG) – SQEP, Michael Green (MG) – Ecomergy, Mike Parsons (MP) – AAIP, Nick Hales (NH) – Consultant, Martin Atkins (MA) – MCA, Divya Atkins (DA) – MCA, Jennifer Kracht (JK) – TomTom, Brent Kimberley (BK) – Durham, Dale Callicott (DC) – BAE Systems, Roland Rosier (RR) - TomTom, Bob Oates (RO) – Blackberry, Arch McKinlay (AM) – NGA, Minh Pham (MPh) – TomTom, Paul McKernan (PMck) - Consultant

Apologies

Fan Ye (FY) – ESC, Dave Murray (DM) – BAE, Rhiannon Chilton (RC) – Dstl, Mike Standish (MS) – Dstl, Mark Nicholson (MN) – University of York, Ali Hessami (AH) – Vega, Paolo Giuliani (PG) - EDF

Agenda

1. Welcome
2. W3W data issues
3. Data 'Cruft'
4. New Data Risk Types
5. Addition of Security Properties
6. Data Life Cycle
7. How important a bit of data is…
8. Fuzz Testing
9. DSITN (Data Safety in the News)
10. Update on Tooling
11. Proposal to make an ISO TR document
12. Migrating, Porting, Exporting and Importing Data - ctd
13. SCSC Seminar and SSS'23
14. Actions
15. Next meeting
16. AOB

NOTE: All comments or opinions in these notes are attributed only to individual attendees of the meeting, not to their respective organisations.

*[Note that actions are presented in the form **N.Mx** where **N** is the meeting number, **M** a reference number for the action raised in that meeting and **x** is an optional letter that differentiates related actions arising from the same discussion point].*

The meeting slides are available at: https://scsc.uk/file/gd/71st_DSIWG_Slides_v1-1437.pptx

1. Welcome

MP opened the meeting and welcomed those attending.

2. W3W (What3Words) Issues

MP presented some of the information and links found by Andrew Whitehead of the SCSC SAWG on the issues of the W3W (What3Words.com) location service. This service can give you a highly accurate location based on only 3 words (e.g. 'dishes.spins.vocal' is MP's home desk location). The service has been heavily marketed in the UK and is used by emergency services.

However there are problems with words which sound the same (homophones)[1], such as Flower and Flour. If these are read out over a phone there may be confusion. Surprisingly W3W doesn't avoid (all of) these. Also plurals of words are allowed, and of course, people may not pronounce the plural ending of a word, such as 'egg' and 'eggs'. What is worse is that the algorithm used by W3W doesn't necessarily separate the locations of the ambiguous words, so they could be in the same area, and emergency services may mistakenly attend the wrong location. This all leads to possible confusion, especially if someone is calling over a bad phone connection or is in distress.

There was a good discussion of the issues in the meeting. The general consensus was that the choice of words could be improved, or a 4th "check word" added.

Apparently W3W is available in other countries and uses words in the native language. This may also lead to confusion as, for instance, a foreign tourist to the UK may know a location by other language words, and so a UK W3W operator may have to deal with different language sets.

Links:
https://www.bbc.co.uk/news/technology-56901363
https://cybergibbons.com/security-2/why-what3words-is-not-suitable-for-safety-critical-applications/
https://techcrunch.com/2021/04/30/what3words-legal-threat-whatfreewords/

The Data Safety Guidance document would definitely apply to a system such as W3W. It was thought that the 'human in the loop', i.e. human reading out W3W location from a phone and human listening to the call was something we should explicitly consider. This is like the 'Chinese Whispers' issue, amplified by the ambiguities in W3W.

Note this is related to the 'Aliasing' problem already mentioned in the DSG.

Action 71.1 (MP) – Add Homophones/Homonyms explicitly to the guidance.

JK noted that the W3W approach has limitations: e.g. a four year old child might not be able to read, but might be able to use symbols [or indeed emojis!].

[1] A homophone is a word that sounds the same as another word but has a different meaning and/or spelling. "flower" and "flour" are homophones. Homonym is also sometimes used but has several meanings, so to avoid ambiguity, we use homophone. See https://www.vocabulary.com/articles/chooseyourwords/homonym-homophone-homograph/

3. Data Cruft

MP described the concept of Data Cruft based on the existing one from software, https://en.wikipedia.org/wiki/Cruft:
Cruft is a jargon word for anything that is left over, redundant and getting in the way. It is used particularly for defective, superseded, useless, superfluous, or dysfunctional elements [in computer software].

He explained that there could be data analogies for both Dead Code and Deactivated Code, i.e. Dead Data and Deactivated Data. He then went on to explain the types of data that might form part of these categories:

4. New Data Risk Types

Data Risk Category	Data Risk Type	Explanation
'The Dead'	**Dead Data**	Data which cannot be used by system design, and should therefore be removed to avoid possible risk of future usage
	Deactivated Data	Data unused due to configuration of the system or usage, and thereby presents a risk as it is unused, until it is eventually used (e.g. due to a change, planned reactivation or a failure)
	Zombie Data (Undead Data)	Data which should not be used (and is assumed to be Dead or Deactivated Data) but somehow is used, perhaps in an obscure way, or in parts of the system rarely used.
	Resurrected Data	[Added in the meeting] Data which is assumed lost or deleted but in fact can resurface when certain conditions arise.
	Dormant Data	[Added in the meeting] Data which is temporarily unavailable or intermittently unavailable, on a regular or irregular basis.
Legacy-Derived	**Fossil Data**	Data which is unused for now, and as such never gets updated as it should to current values / standards / usage. So when it is finally used (e.g. due to change or re-configuration, or indeed change of staff) may present a problem.
	Hangover Data	Legacy data which has been left over from previous versions of the system and should not be needed or used, but it may be used and you are not sure of use
	Morphed Data	Data which originally had a defined single purpose or set of purposes, but has since been re-used and re-purposed so that it may be influencing many more things and you don't know what it impacts

Data Risk Category	Data Risk Type	Explanation
Change-Related	**'C-3PO'[2] Data Also known as Metathesiophobia Data[3]**	1. Data you are not aware of and therefore have no idea what it does 2. Data you are aware of but have no idea what it does, and therefore are reluctant to modify 3. Data you are aware of and think you know what it does (but actually you don't) – therefore change may cause problems 4. Data you are aware of and you do know what it does 5. [Added in meeting] Data you are aware of but afraid to change its format (e.g. update to latest schema)

There was then some discussion in the meeting about these types. Some of these relate to Dark Data and some to Dazzle Data (see latest Data Safety Guidance document, https://scsc.uk/scsc-127G). There may be another 'Dead' type where data is temporarily unused or unavailable, perhaps 'Dormant Data'. It was thought that data embedded in software executables by the build process inserted by compilers, linkers, etc. (typically version strings, but can also contain information about who made edits, etc.) is a type of Dead Data, possibly Zombie Data. [It can be used by e.g. build configuration checkers.]

RR mentioned an example of Fossil Data (or possibly Hangover Data) being a mobile phone storing the names of the access points that it has previously successfully connected to at least once - allowing a Rogue AP: https://kalitut.com/rogue-ap-fake-access-points/

There is also an additional type of 'Resurrected Data' where it has been assumed deleted (and can't be accessed in its original context) but in fact can resurface later e.g. a bit-wise copy is made of a disk, copying over deleted sectors to the new disk. It was thought that many storage devices have this problem, e.g. flash memory where data is not actually deleted until the space is reused (this is why recovery software works). It was noted there have been cases of people buying second-hand computer equipment to find they can access all sorts of data left by the previous owners (RO mentioned that from a security perspective, complete physical destruction of a device is the only way to make sure the data is really deleted).

It was noted that Dead Code has different definitions, some more formal than others. Can we come up with a formal definition for Dead Data? Also, in the software world there is a subtle distinction in that there is also Unreachable code. Is there an analogy for data [E.g. data past a sentinel marker, or perhaps padding data between records]?

There was also discussion about the concept of 'Code Coverage' when testing software. It was thought there may be a similar case for data: so for instance, 100% data coverage is where every data item has been touched/used in some way. It was noted that SPARK Ada does a degree of data coverage.

[2] From Star Wars™ - C-3PO was hesitant about change. Alternative names/analogies are welcome!

[3] Metathesiophobia, is a phobia that causes people to avoid changing their circumstances due to being afraid of the unknown

For the Change category it was thought there was also an issue where people are afraid to change formats (for instance to latest version of data schema) because they are unsure what will happen.

MA pointed out that cleaning or sanitizing the data for an upgrade presents its own risks.

Some additional data risk types were discussed:

- Abandoned / Disowned / Disavowed [Orphaned?] data – data which no longer has an owner / maintainer
- 'Derelict' data – data which has not been updated / maintained and is no longer fit for purpose, and so is 'worked around' at every use or invocation of the system. Note this may effectively be the same as Fossil Data.
- 'Dr Bob' data (Magic Numbers) – data typically distributed through software where its derivation is not obvious (typically a literal number or character string where a constant definition should have been used). The issue here is that if there are many cases of the same value it is not clear if all should be changed if one is updated.

It was noted that PH and MP have agreed that it would be useful to have an ontology of these data types, and hope to have a SCSC eJournal article on this next year, with a Newsletter summary article.

Action 71.2 (MP/PH) – Consider these additional data risk types (Abandoned/Derelict/Magic Numbers, etc.) in the list of types and in the articles

5. Addition of Security Properties

DA presented some of the work to look at including security properties in the data guidance, e.g. confidentially, authenticity, and non-repudiability. There was some discussion as to whether these apply to data specifically or the whole system.

Action 71.3 (PH/DA/RO) – Develop security properties thinking further for next DSIWG

Action 71.4 (PH/DA/RO) – Present security properties work to next SISWG meeting

6. Data Life Cycle

AM presented some of the work he had been doing on enterprise-level data lifecycles. This work looked very interesting and clearly a lot of thought had gone into it. A discussion followed and there was a suggestion that a structuring may be useful, e.g.
https://www.iso.org/obp/graphics/std/iso_std_iso-sae_21434_ed-1_v1_en/fig_1.png or
https://www.iso.org/obp/graphics/std/iso_std_iso_26262-1_ed-2_v1_en/fig_1.png

Action 71.5 (AM) – (i) Establish if any of this can be published within the DSIWG and (ii) Consider a structuring similar to that used in security standards or ISO26262

7. How important a bit of data is…

MP presented two articles of relevance:

BBC News - NHS Covid pass down, leaving some passengers struggling to board flights: https://www.bbc.co.uk/news/uk-62599489

Self-Driving Vehicles – Data is key: https://eandt.theiet.org/content/articles/2022/09/can-the-road-network-cope-with-self-driving-vehicles

8. Fuzz Testing

MP mentioned the concept of Fuzz Testing or Fuzzing, https://en.wikipedia.org/wiki/Fuzzing which has data safety aspects to consider (e.g. choice and distribution of data used). There was also the issue of deciding if the result of a Fuzz test case was correct or incorrect.

RO said that Fuzzing was often used in security where the choice of distribution is important and reproducibility is key. It also has to be well-documented.

Action 71.6 (RO) – See if an expert can be found to give a presentation on Fuzz Testing to next meeting

9. Data Safety in the News (DSITN)

MA mentioned the following via email:

Fly-by-wire "data" problem (see the second paragraph): https://qr.ae/pNZM0P

Experienced crew struggled with instrument flight after 737 lost autopilots: https://www.flightglobal.com/safety/experienced-crew-struggled-with-instrument-flight-after-737-lost-autopilots/140072.article (reasons not yet known, but data issue at least in that the voice data recorders did not provide voice recordings.)

Dark Data: Is this relevant to the dark data discussions? https://www.pinterest.co.uk/pin/756464068674050138/

10. Update on Tooling

DA said that the Data Safety Tool was available for trial use: https://data-safety.tech/tooling/ The tool has not yet been updated to use the latest version of the guidance (v3.4).

11. Proposal to make an ISO TR document

No update

12. Migrating, Porting and Importing Data - ctd

No update.

13. SCSC Seminar and SSS’23

MP mentioned the upcoming December on Testing: https://scsc.uk/e966 which has a data aspect. Also SSS’23 takes place in February 2023: https://scsc.uk/e898 - bookings are open.

Safety-Critical Systems Club
SSS’23 PROGRAMME
31st Safety-Critical Systems Symposium
The Principal York, UK
7 - 9 FEB
2023
SCSC

RO mentioned that he is presenting a paper at this conference on supplier maturity and safety arguments.

It was thought that there was significant crossover between the SCSC working groups and some cross-representation would be good. There was a suggestion to give each group a short slot for a presentation in another group. Data Safety particularly has shared interests and approaches with security and autonomy working groups.

14. Actions

61.2 will be progressed via meetings
63.1 Ongoing (4 pages written)
68.1 Close
69.1 Further explanation provided in the meeting. Still open.
69.5 Close

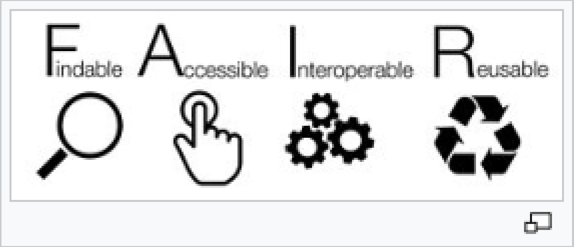
15. AOB

JK asked if there is a safety maturity model for organisations doing data safety. The consensus was that this was not yet in place, although the DSG does the following two items: *Appendix C Data Safety Culture Questionnaire* and also a questionnaire for *Supplier Data Maturity* (Appendix D).

BK mentioned the concept of Steganography, https://en.wikipedia.org/wiki/Steganography which is *“…the practice of concealing a message within another message or a physical object. In computing/electronic contexts, a computer file, message, image, or video is concealed within another y file, message, image, or video.”* This would seem to have relevance to both data safety and security, and possibly counts as Dead data.

There were further references mentioned regarding Dead data:
https://wiki.sei.cmu.edu/confluence/display/cplusplus/EXP62-CPP.+Do+not+access+the+bits+of+an+object+representation+that+are+not+part+of+the+object%27s+value+representation

CT mentioned the concept of FAIR data (https://en.wikipedia.org/wiki/FAIR_data) which is relevant to the group: *"FAIR data are data which meet principles of findability, accessibility, interoperability, and reusability (FAIR). The acronym and principles were defined in a March 2016 paper in the journal Scientific Data by a consortium of scientists and organizations."*



Action 71.7 (MP/CT) – Consider impact of FAIR data on the guidance

One meeting attendee reported the following:

> *"Whilst conducting my weekly checks I noticed the relatively new emergency phone required a software update. These phones were introduced as we moved away from traditional landlines at our desks, (which of course never had software updates) … In the event of an emergency, it would be terrible to have to wait for the update to use the phone. This is an uncontrolled change to the site LOPA (layers of protection analysis). Some further questions: 1) Why does an emergency phone need an update when we are only calling for site emergencies? 2) How do I know this update is real and not a cyber-attack; 3) How do I check whether we should install the update – who has pushed and authorised the update?"*

16. Next Meeting

Next meeting will be held end November in London/Bath/Reading if a suitable venue can be found (also accessible by Zoom). MP to arrange.

17. Thanks

Thanks to all those who sent data safety in the news links.
Thanks to MP for taking the minutes.
Thanks to MP for chairing.

Summary of Open Actions

Actions greyed out are considered closed and will be removed from the list at next issue.

Ref	Owner	Description	Target Guidance Version
42.9	MP	Work out a matrix of data categories (previously 'types') and data properties (as per DB discussion)	N/A
43.4	MP	Write up a data focussed FMEA approach.	4.0

Ref	Owner	Description	Target Guidance Version
44.2	MP	To discuss with AK on how to get the Wikipedia article published	N/A
46.1	MP	Review the application of DSALs to higher level forms of aggregation	N/A
49.6	MT	Review Overleaf briefing material and aim to hold a briefing before end of March 2021 in the use of Overleaf in the production of the guidance.	N/A
53.1	MP	To talk to Kevin King about what we need to do in the guidance for digital twins.	4.0
61.2	AW	Research the relevance of digital currencies and report back to the group (with MA and MT)	4.0
63.1	CT	Look at both Dark Data and Dazzle Data for sensors (e.g. when a sensor is saturated, in noisy environment or when readings are below the detection level floor)	4.0
64.1	MP	Contact Thor and establish the details of the guidance proposals in the paper.	4.0
66.6	MT	Add these three properties [‘Analysability’, ‘Explainability’, ‘Verifiability’] to the user-visible further work section. If time allows then develop into the guidance further.	4.0
68.1	MP/PH	Develop the Black Swan / Dragon King Data work further and consider publishing as a newsletter article	4.0
68.2	MP/MT	Develop the migration work further and present at next meeting	4.0
69.1	CT	Establish a list of similar / related TRs that we could use as examples.	
69.2	RR	Explore the issue of data / software compatibility issues and to what extent data can impose requirements on software	4.0
69.3	PMcK	Develop a scoping diagram that shows how the DSG fits into the overall lifecycle process and other standards	4.0
69.4	MA	Write a short note on the issues of aggregation	4.0
69.5	RO	Look at the government call for information and see if there were any opportunities for the group to provide useful input.	-
69.6	MA/DA	Update the data safety tool to use the latest version of the guidance document	-
70.1	MA/DA	Investigate feasibility of creating searchable web database of data safety-related accidents.	-
71.1	MP	Add Homophones/Homonyms explicitly to the guidance.	4.0
71.2	MP/PH	Consider these additional data risk types (Abandoned/Derelict/Magic Numbers) in the list of types and in the articles	4.0
71.3	PH/DA/RO	Develop security properties thinking further for next DSIWG	4.0
71.4	PH/DA/RO	Present security properties work to next SISWG meeting	-
71.5	AM	(i) Establish if any of this can be published within the DSIWG and (ii) Consider a structuring similar to that used in security standards or ISO26262	-
71.6	RO	See if an expert can be found to give a presentation on Fuzz Testing to next meeting	-
71.7	MP/CT	Consider impact of FAIR data on the guidance	4.0