DSIWG

SCSC Data Safety Initiative – WG Meeting 69

16th June 2022, Bath, UK and Zoom

Minutes

Attendees

Paul McKernan (PMcK) - Dstl, Mark Templeton (MT) – Qinetiq, Michael Green (MG) – Ecomergy, Paul Hampton (PH) – CGI, Mike Parsons (MP) – AAIP, Andy Williams (AW) – Consultant, Dave Murray (DM) – BAE, Nick Hales (NH) – Consultant, Martin Atkins (MA) – MCA, Divya Atkins (DA) – MCA, Roland Rosier (RR) – TomTom, Jennifer Kracht (JK) – TomTom, Bob Oates (RO) – Blackberry, Carl Tipton (CT) – Johnson Matthey, Carl Jackson (CJ) – Johnson Matthey, Arch McKinley (AMcK) – St Louis, Brent Kimberley (BK) – Durham, Dave Banham (DB) - Blackberry

Apologies

Oscar Slotosch (OS) – Validas, Richard Garrett (RG) – SQEP, John Bragg (JB) – MBDA, Mark Nicholson (MN) – University of York

Agenda

1) Welcome
2) Proposal to make an ISO TR document
3) Machine Learning Data
4) Consultation: Data storage and processing infrastructure security and resilience
5) Update on Tooling
6) The Golden Ray Car Carrier Accident
7) Naxos Algorithms
8) Migrating, Porting and Importing Data - ctd
9) SSS'23 Call For Abstacts
10) Further Thoughts on 'Black Swan Data'
11) Data Safety in the News
12) Actions
13) Next meeting

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/69th_DSIWG_Slides_v1-1401.pptx

1. Welcome

MP opened the meeting and welcomed those attending.

2. Proposal to make an ISO TR document

MP explained that David Ward of Horiba-Mira suggested that we propose the Data Safety Guidance document as an ISO Technical Report (RT). He suggested that this would be relatively little work.

There was then a discussion about the pros and cons of this. RR said that the pros included that the work would have more 'weight', would gain more traction in industry and could be referenced from other ISO standards.

DB suggested that our first 'port of call' ought to be BSI who could guide us through the process. We need to decide where it would fit with other ISO standards. AW suggested that we could propose a smaller, cut-down, document to ISO, keeping the larger document under our control. PMcK said that the guidance part could be split out.

It was thought that the DSIWG could control the original document and progress it as thought necessary, separate from the TR.

MP explained that he was concerned with the amount of work required and the resources available to do it.

ACTION 69.1 (CT) – Establish a list of similar / related TRs that we could use as examples.

3. Lifecycles

There was a good discussion about lifecycles for data, and those for software involving data. RR asked if we could introduce a lifecycle for data. CT mentioned the issue of data and software mismatch if they are being developed to different lifecycles and/or schedules.

JK mentioned this work which looks relevant: https://lakefs.io/what-is-data-lifecycle-management/

MP noted that there is something in the guidance already on software and data, (e.g. section 5.3.2.2 of DSG 3.4), but probably more is needed.

ACTION 69.2 (RR) – Explore the issue of data / software compatibility issues and to what extent data can impose requirements on software

There was a general discussion as to the scope of the DSG and where it fits with other standards and guidance.

ACTION 69.3 (PMcK) – Develop a scoping diagram that shows how the DSG fits into the overall lifecycle process and other standards

4. Machine Learning (ML) Data

MP suggested that the ML section in the guidance requires elaboration and updating.

ACTION 69.4 (CJ) – Propose an update for the ML Appendix in the guidance (Appendix J of DSG 3.4)

5. Consultation: Data storage and processing infrastructure security and resilience

This security-focussed consultation looking at cloud storage of data was discussed, https://www.gov.uk/government/publications/data-storage-and-processing-infrastructure-security-and-resilience-call-for-views/data-storage-and-processing-infrastructure-security-and-resilience-call-for-views

DA suggested we should consider a response to this request.

MP suggested that the key takeaway from this work was thinking about how data aggregation can create new risks, possibly through emergent properties.

This led to a fruitful discussion about aggregation risks:

There can be problems introduced by incorrect aggregation, or the wrong conclusions can be drawn from the result of the aggregation, perhaps because of emergence, or because something has been lost or hidden. There is also the issue of compatibility of the data used in the aggregation (one set may be out of date, different format or context, etc.). There is also the issue of too much data being produced, hiding some data values for instance.

MG mentioned that he had seen issues with low-level format incompatibilities (e.g. Excel and Google Sheets).

RO mentioned that aggregation can affect privacy and become a security issue as things can be deduced that may not be apparent from each component data set.

MA mentioned that there may be an issue with 'opportunistic aggregation' e.g. in medical records where an X-Ray image may or may not be attached to set of medical notes.

AMcK commented that *"Aggregation is the same as large acquisition ingestion. We have many examples of basis shift, time shift, etc., that SHOULD be obvious in metadata cleaning of data during ingestion. We have a severe and manual process on ingestion to find, tag, and clean these time shifts but the provider is the root cause."*

In general aggregation may create 'more than the sum of the parts' as gaps or holes in the data can be filled with multiple data sets.

ACTION 69.4 (MA) – Write a short note on the issues of aggregation

ACTION 69.5 (RO) – Look at the government call for information and see if there were any opportunities for the group to provide useful input.

6. The Golden Ray Car Carrier Accident

MP mentioned this accident involving the complete loss of a car-carrier vessel and all of its cargo due to a small data error regarding ballast loading, resulting in an $250,000,000 financial loss

A short video on the accident was presented, https://youtu.be/z3b4Cuot4C4

7. NAXOS Algorithms

These were briefly discussed.

8. Update on Tooling

(See separate slide: https://scsc.uk/file/gd/RADISH-1404.pptx)

MA presented the latest version of the Data Safety Tool and the group thought this looked really promising. This is now available free of charge for evaluation. Feedback and comments are requested on this tool.

The original data safety tooling web site is at: https://data-safety.tech/tooling/

CT mentioned that he really liked it.

MP mentioned that the SCSC can help with promotion via: (i) an SSS paper detailing an example, (ii) links to the website for the tool, (iii) links on the DSIWG group pages on the SCSC website and (iv) a LinkedIn article.

Tool qualification was mentioned as possible future issue, but MA said this was not considered as yet.

ACTION 69.6 (MA/DA) Update the data safety tool to use the latest version of the guidance document

9. Dragon Kings, Black Swans and Data

MP said that he had been considering what Dragon King data might be (PH and MP had covered Black Swan data at the last meeting).

The key thing with Dragon King risks is that they are the result of some escalating, non-linear process and rapidly get out of control. They may have been identified but the rate at which things go wrong is what makes it a Dragon King.

MP said that some security risks such as viruses and ransomware might come into this category. RO mentioned the NotPetya attack affecting Maersk: https://www.industrialcybersecuritypulse.com/throwback-attack-how-notpetya-accidentally-took-down-global-shipping-giant-maersk/

There can be DK data events where too many advisories or warnings are rapidly produced, overwhelming the operator.

Configuration data mistakes, weights used for data or parameter data can also cause rapid failure scenarios.

CT mentioned that his PhD was in non-linear dynamics.

It was noted that the A400M crash might also come into this category,
https://www.bbc.co.uk/news/technology-33078767

The case of Lorenz's small rounding errors leading to large changes was mentioned,
https://en.wikipedia.org/wiki/Edward_Norton_Lorenz
"In 1961, Lorenz was using a simple digital computer, a Royal McBee LGP-30, to simulate weather patterns by modeling 12 variables, representing things like temperature and wind speed. He wanted to see a sequence of data again, and to save time he started the simulation in the middle of its course. He did this by entering a printout of the data that corresponded to conditions in the middle of the original simulation. To his surprise, the weather that the machine began to predict was completely different from the previous calculation. The culprit: a rounded decimal number on the computer printout. The computer worked with 6-digit precision, but the printout rounded variables off to a 3-digit number, so a value like 0.506127 printed as 0.506. This difference is tiny, and the consensus at the time would have been that it should have no practical effect. However, Lorenz discovered that small changes in initial conditions produced large changes in long-term outcome."

[MP subsequently had an email from AMck which referenced a very informative paper: https://www.sciencedirect.com/science/article/pii/S0951832021001678?ref=pdf_download&fr=RR-2&rr=71ca87477a9a76bf He has since suggested to PH that the forthcoming Newsletter article make reference to 'Perfect Storm' data as well.]

10.Migrating, Porting and Importing Data - ctd

MP covered the slides from the last meeting, slightly revised. There was a discussion about general data chains / pipelines and transformations. In general data is now a product of complex chains, branches and paths and so the picture is more difficult. MP thought it best to just cover migration for now.

RR noted that the 3rd case in the slides ought to be bi-directional.

In general there were thought to be many issues in the synchonisation of data when migration occurs (e.g. data may be stale, data may need to backed out, data may be rejected, etc.)

NH mentioned that using emulators for the migration process might help.

11.AOB

MP mentioned that the initial KO meeting for the Systems Approach to Safety of the Environment WG (SASEWG#1) is being held 17/06/22 by Zoom.

12.Actions, etc.

See table at end.

13.Next Meeting

Next meeting will be held 21st July 4:00-5:30pm 2022 by Zoom:
https://us02web.zoom.us/j/84867419359?pwd=QzU3RGp4eFUzeVBLU1NPaGNicUh1Zz09

14.Thanks

PMcK mentioned that he is retiring. The WG thanked PMcK for all the help and support with the DSIWG over the years.
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
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
56.2	DA	Consider impact of Dark Data on the Data Safety Tool	N/A
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
65.1	MP	Contact Davy Pissoort and see if any interest in this funding route for Data Safety	-
65.2	DA/MA	See if access to the tool can be given to TomTom for evaluation	-
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 [Note: suggest also add Perfect Storm data as well.]	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	