Statement of Requirement (SoR)

Reference Number	[REDACTED]
Version Number	**1.0**
Date	**[REDACTED]**

1.	**Requirement**
1.1	**Title**
	Automated Ingest and Processing of Cyber Data
1.2	**Summary**
	Dstl is looking to develop a set of automated ingest pipelines to enable the Cyber Data Framework to mature and move towards an active capability. Phase 1 and Phase 2 of this work have seen the development of a standard Cyber Data Framework (CDF). The CDF is a [REDACTED] owned framework to aid in storing, sharing and managing [REDACTED] cyber event data, and provides a platform for research and development of novel analysis tools. The CDF is heavily influenced by the [REDACTED] data structure and also includes a set of specialist additions recommended by previous work. Most of the work up to this point has been primarily focussed on defining and designing the Cyber Data Framework and its hosting solution. The ingestion pipelines of raw data sources into the CDF data structure were not a focus of Phase 1 or 2 - as such many assumptions were made about these processes and their design was left undefined. In this work, attention should be given to how data could be ingested into the CDF in a way that is as automated as possible. **Research Questions and Objectives** The following questions form the basis of this work:

Question 1: Given a set of **known and well-structured cyber data** formats, how can an ingestion pipeline be constructed to **automate** the re-structuring of these raw data file types into the Cyber Data Format?

Question 2: Given a **new (unseen) cyber data format** and the CDF data specification, how could an ingestion pipeline be constructed to **automate** the re-structuring of the raw data into CDF data, as much as possible?

The two questions have subtle differences, the first is focussed on engineering software to enable standard data types to be ingested into the CDF. The second is more research focussed and is an exploration into what new, machine-aided techniques exist that could help with this problem.

The primary objectives of this SOR are:

- To use the proposed CDF Data Specification v2.0 and CDF Implementation Specification from Phase 2 to develop a set of ingest pipelines which are able to transform [REDACTED] into CDF structured data.
- To explore the art of the possible for combining [REDACTED] and other machine-aided techniques with Cyber Data Processing, for an automated approach to ingesting raw data into the CDF.

Task Structure

The two research questions naturally organise the work into two work packages. It is recommended that the work packages are run sequentially as it is expected insights from WP1 will help in WP2, though this is not a required and may be done in parallel:

1) WP1 – Ingest Pipeline Engineering

The first work package aims to answer Question 1 of the research questions. Many of the data sources the CDF expects to be presented with originate from a set of popular open-sourced cyber security tools, which write their output in well-documented, tabular data formats. In order to more rapidly advance the CDF, there must be an automated way to ingest these data sources, reducing the load on the analyst as much as possible. This work package should be focussed on developing software which is able to automatically extract CDF objects from the raw data and input into the CDF Storage.

2) WP2 – Ingest Pipeline Art of the Possible

	The second work package aims to answer Question 2. The landscape and origins of data sources for the CDF will evolve over time, and the CDF aims to be resilient against changes to the Cyber Data domain. When presented with a new data structure, ideally the CDF will be able to automatically ingest as much of this data source as possible. With recent advances in [REDACTED] and Machine Learning in the Cyber security domain, this work package will include a general scoping of how these techniques could be exploited within the CDF Ingestion process, and where appropriate develop a prototype demonstrating the findings.
1.3	**Background**

This SOR builds upon the following work previously delivered in Phase 1 and 2:

Phase 1

Phase 1 of the CDF task was focussed on exploring the feasibility of a standard Cyber Data Framework (CDF) which enables a range of cyber data sources to be condensed into a single data structure which can be used by analysts in Defensive Cyber operations. The output of Phase 1 was an evaluation of existing open source and commercial Cyber Data structures, followed by a proposed Cyber Data Framework v1.0 based on the [REDACTED] object format, with an additional set of custom [REDACTED] objects.

Phase 2

Phase 2 builds upon the proposed CDF v1.0 delivered in Phase 1, and considers how the CDF may be implemented in practice. The scope of this phase included an evaluation of the suitability of CDF v1.0, identification of implementation considerations and proposed a hosting system design.

The four work packages comprised:

- WP1 – CDF for Analysis: To test, understand and describe how the CDF can be used to store data to support analysis in support of cyber defence goals.
- WP2 – CDF for Knowledge Transfer: To identify and overcome the practical knowledge management challenges and describe a pathway for the CDF's integration with wider [REDACTED] organisational structures and assets.
- WP3 – CDF for Exploration: To investigate and define how data interrogation in support of cyber defence analysis can be managed effectively using the CDF.
- WP4 – CDF for Implementation: To provide clear guidance on the implementation of the CDF within the Dstl Cyber Research Team and a roadmap for extending this across the desired user community.

The work was broken down into 4 sequential sprints:

1) Framework Evaluation: Evaluated the appropriateness of CDF v1.0, and suggested suitable extensions - resulting in the CDF Data Specification v2.0
2) CDF Use Case definition: Designed a set of use cases and worked examples to illustrate how the CDF Data Specification v2.0 is used in practice.
3) System Analysis: Examined the challenges and system requirements of using the CDF Data Specification v2.0 in practice for the proposed Use Cases.

	4) Solution Design: Designed a system architecture which meets the requirements and tackles the challenges identified in the System Analysis
1.4	**Requirement**
	Definition 1.1 (Ingestion Pipeline) A CDF **Ingestion Pipeline** is a set of processes which takes a set of raw cyber data files and re-structures them into the format described in the CDF Data Specification from Phase 2. The Ingestion pipeline contains at least the following processes as proposed in Phase 2: • Raw Data Format Identification – Is this a known log type, or unseen? • Data Cleansing – Are there any malformed entries, or missing information? • Data Reformatting to CDF Data Structure • Raw File metadata capture (including Data Classification/Caveats and other protective markings) • Putting Data into CDF Storage with relevant CDF Auditing **WP1 – Ingest Pipeline Engineering** This work package should be primarily focussed on answering Research Question 1: *'Given a set of known and well-structured cyber data formats, how can an ingestion pipeline be constructed to automate the re-structuring of these raw data file types into the Cyber Data Format?'* This should be done by creating software which enables ingest of 'common' cyber data into the CDF. 'Common' cyber data in this context covers any data format expected by Cyber Security Analysts within [REDACTED] and currently includes (but not limited to) [REDACTED] alerts. 1.1 Must use [REDACTED] alerts as the known data structures, these data sources are relevant to[REDACTED] and the CDF's intended users. 1.2 Must propose a suitable software pipeline, which is compatible with the proposed CDF Data Specification v2.0 and integrate with the proposed CDF implementation architecture. 1.3 Where possible the proposed pipeline should be resilient to software changes with the CDF data solution, changes with the CD Data specification and resilient to expected future upscaling of the CDF.

1.4 The software pipeline must use components licenced suitably for [REDACTED] to distribute with no further cost or imposed restrictions. Specific permitted licences to be agreed between Dstl and the supplier.

1.5 Must develop a set of ingest pipelines for each of the common data formats identified in 1.1, which include each of the steps highlighted **Definition 1.1.** Where any of these steps can't be automated, a reasonable manual process should be identified instead.

1.6 Must capture any design choices and limitations faced during this process and document in the accompanying report.

1.7 Must develop, alongside any software produced, suitably detailed documentation and user guides.

1.8 The ingest pipelines must exceed the processing rates expected through manual input. Moreover, performance improvements that take minimal effort to implement should be made

1.9 Similarly, whilst no specific error rates are required, the number of errors made by the pipelines must not exceed the expected number of errors made during manual processing.

1.10 The pipeline must be able to deliver acceptable performance for injecting one hours' worth of logging (roughly [REDACTED] with an average size of [REDACTED] and a maximum size of [REDACTED] in an hour')

1.11 Must define requirements for deploying such applications, and how the CDF data solution should be adapted to account for these additional processes.

WP2 – Ingest Pipeline Art of the Possible

This work package is focussed on answering Research Question 2 - "*Given a new (unseen) cyber data format and the CDF data specification, how could an ingestion pipeline be constructed to automate the re-structuring of the raw data into CDF data, as much as possible?"*

This work package should explore state of the art Machine Learning techniques that could aid in any of the steps of the ingest pipelines during the processing of unseen cyber data. This could be exploiting [REDACTED] on unseen text-based logs or similar. The required level of automation is not strict, this activity should explore which parts of the ingestion pipeline can be automated. If Machine Learning is not a viable option, then other software solutions should be

	recommended which reduce the load on the operator during the processing of new data structures. 2.1 Must perform a literature review and scoping for relevant [REDACTED] techniques or other automation approaches that may be appropriate for this work. 2.2 For each stage in the pipeline, where appropriate, identify possible methods of automating the process. This should take into consideration any issues that may arise when using these methods (trust, ethics, compute resources, level of human input required). 2.3 Must develop a software prototype which exploits the techniques identified on a new, unseen cyber data source. 2.4 In software where Machine Learning is used then the [REDACTED] framework is preferred and full access to the model and training set is required. 2.5 Must define requirements for deploying such applications, and how the CDF data solution should be adapted to account for these additional processes. **Security Requirements** The outputs of the project shall be classified no higher than [REDACTED]. **Timescales** All work must be delivered in [REDACTED]
1.5	**Options or follow on work**
	Not applicable

1.6	Deliverables & Intellectual Property Rights (IPR)						
Ref.	**Title**	**Due by**	**Format**	**TRL***	**Expected classification (subject to change)**	**What information is required in the deliverable**	**IPR DEFCON/ Condition**
D1	CDF Ingestion Pipelines & Documentation	REDACTED]	Software and Documentation	3	[REDACTED]	Software and associated documentation and instructions for the outputs of WP1: • User Guides • Installation Process Where appropriate, libraries, source code, and other artefacts needed to build and distribute the pipeline will be delivered	DEFCON 703
D2	CDF Ingestion Pipeline Report	[REDACTED]	Report		[REDACTED]	Report describing the design process of the pipelines for WP1, including: • Platform evaluation process • Pipeline design considerations and limitations • Results from testing	DEFCON 703

						• Resilience against future software changes/upscaling	
D3	Automated Cyber Data Processing Report	[REDACTED]	Report		[REDACTED]	Report describing the research performed in WP2, including: • Literature Review • Art of the possible mapping to ingest pipeline stages • design and performance analysis • Wider CDF solution considerations	DEFCON 703
D4	CDF Automated Ingest Pipelines & Documentation	[REDACTED]	Software and Documentation		[REDACTED]	• Software (source code) and associated documentation for the outputs of WP2, (including documentation, and ML models and training set) Where appropriate, libraries, source code, and other artefacts needed to build and distribute the pipeline will be delivered	DEFCON 703

*__Technology Readiness Level required__

1.7	**Standard Deliverable Acceptance Criteria**
	As per Framework T&Cs ll Reports included as Deliverables under the Contract e.g. Progress and/or Final Reports etc. must comply with the **Defence Research Reports Specification (DRRS)** which defines the requirements for the presentation, format and production of scientific and technical reports prepared for MoD. Interim or Progress Reports: The report should detail, document, and summarise the results of work done during the period covered and shall be in sufficient detail to comprehensively explain the results achieved; substantive performance; a description of current substantive performance and any problems encountered and/or which may exist along with proposed corrective action. An explanation of any difference between planned progress and actual progress, why the differences have occurred, and if behind planned progress what corrective steps are planned. Final Reports: shall describe the entire work performed under the Contract in sufficient detail to explain comprehensively the work undertaken and results achieved including all relevant technical details of any hardware, software, process or system developed there under. The technical detail shall be sufficient to permit independent reproduction of any such process or system. All Reports shall be free from spelling and grammatical errors and shall be set out in accordance with the Statement of Requirement (1) above. Failure to comply with the above may result in the Authority rejecting the deliverables and requesting re-work before final acceptance.
1.8	**Specific Deliverable Acceptance Criteria**
	All reports shall be delivered as both a Microsoft Word (.docx) and Portable Document Format (PDF) (.pdf) document unless specified below (or agreed with the Authority prior to delivery).

2.	Quality Control and Assurance
2.1	**Quality Control and Quality Assurance processes and standards that must be met by the contractor**
	☒ **ISO9001** (Quality Management Systems) ☐ **ISO14001** (Environment Management Systems) ☐ **ISO12207** (Systems and software engineering — software life cycle) ☐ **TickITPlus** (Integrated approach to software and IT development) ☐ **Other:** (Please specify below)
2.2	**Safety, Environmental, Social, Ethical, Regulatory or Legislative aspects of the requirement**
	Not Applicable

3.	Security	
3.1	**Highest security classification**	
	Of the work	[REDACTED]
	Of the Deliverables/ Output	[REDACTED]
3.2	**Security Aspects Letter (SAL)**	
	Yes If yes, please see SAL reference- **[REDACTED]**	
3.3	**Cyber Risk Level**	
	Choose an item.[REDACTED]	
3.4	**Cyber Risk Assessment (RA) Reference**	
	[REDACTED] If stated, this must be completed by the contractor before a contract can be awarded. In accordance with the Supplier Cyber Protection Risk Assessment (RA) Workflow please complete the Cyber Risk Assessment available at https://suppliercyberprotection.service.xgov.uk/	

4.	Government Furnished Assets (GFA)				
GFA to be Issued - Yes					
GFA No.	**Unique Identifier/ Serial No**	**Description:**	**Available Date**	**Issued by**	**Return Date or Disposal Date (T0+)**

GFA-1		GFI – the final deliverables from Phases 1 and 2 of the CDF work			[REDACTED]
		Datasets from [REDACTED] and other sources may be required			[REDACTED]

5.	**Proposal Evaluation criteria**
5.1	**Technical Evaluation Criteria**
	The standard EW&C scoring criteria will be used **:**

Technical – to include individual team members experience and expertise	45% Weighting
Value for Money	15% Weighting
Delivery – to include plan and resource management	15% Weighting
Quality	20% Weighting
Commercial	5% Weighting

5.2	**Commercial Evaluation Criteria**
	As per framework evaluation criteria