NCIA/ACQ/2023/06668
08/March/2023

REQUEST FOR INFORMATION

PROJECT “Deduplication of data items in preparation for archiving”

NCI Agency Reference: RFI-CO- 115878-DATA

NCI Agency is seeking information from Nations and their Industry regarding the availability of solutions among all NATO Nations.

NCI Agency Point of Contact

Contracting Officer: Graham Hindle

E-mail:JGraham.Hindle@ncia.nato.int copy to :
Gracja.Jablonska@ncia.nato.int

NATO Communication and Information Agency
Boulevard Leopold III
110 Brussels
Belgium
www.ncia.nato.int

To: Distribution List (Annex A)

Subject: **NCI Agency Request for Information RFI-CO- 115878-DATA**

1. NCI Agency requests the assistance of the Nations and their Industry to identify Service that can meet or exceed NATO requirements under the scope of project this project.
2. A summary of the requirements is set forth in the Annex B attached hereto. Respondents are requested to reply via the questionnaire at Annex C. Other supporting information and documentation (technical data sheets, descriptions of existing installations, etc.) are also desired.
3. The NCI Agency reference for this Request for Information is **RFI-CO- 115878-DATA** and all correspondence and submissions concerning this matter should reference this number.
4. Responses may be issued to NCI Agency directly from Nations or from their Industry (to the staff indicated at Paragraph 8 of this Request for Information). Respondents are invited to carefully review the requirements in Annex B.
5. Responses shall in all cases include the name of the firm, telephone number, e-mail address, designated Point of Contact, and a NATO UNCLASSIFIED description of the capability available and its functionalities. This shall include any restrictions (e.g. export controls) for direct procurement of the various capabilities by NCI Agency. Non-binding pricing information is also requested as called out in Annex C.
6. Responses are due back to NCI Agency no later than 17:00 Brussels time on 5 April 2023.
7. Please send all responses via email to the following NCI Agency Point of Contact:

   For the attention of : Mr Graham Hindle email: Graham.Hindle@ncia.nato.int;
   With a copy to : Ms Gracja Jablonska Gracja.Jablonska@ncia.nato.int;

8. Service demonstrations are not foreseen during this initial stage. At this stage, clarification requests or any further questions are not accepted in return. NCI Agency reserves the right to invite respondents to explain both their pricing information and the approach with an online session to be arranged in when responses have been received and analysed expected in April 2023.
9. Respondents are requested to await further instructions after their submissions and are requested not to contact directly any NCI Agency staff other than the POC identified above in Paragraph 7.
10. Any response to this request shall be provided on a voluntary basis.
11. Not responding will not prejudice or cause the exclusion of companies from any future procurement that may arise from this Request for Information.
12. Responses to this request, and any information provided within the context of this survey, including but not limited to pricing, quantities, capabilities, functionalities and

requirements will be considered as information only and will not be construed as binding on NATO for any future acquisition.

13. The NCI Agency is not liable for any expenses incurred by firms in conjunction with their responses to this Request for Information and this shall not be regarded as a commitment of any kind concerning future procurement of the items described.

14. Your assistance in this Request for Information request is greatly appreciated.

FOR THE CHIEF OF ACQUISITION:

Graham Hindle

Digitally signed by Graham Hindle
Date: 2023.03.09 10:09:28 +01'00'

Graham Hindle
Senior Contracting Officer

Enclosures:
Annex A (Distribution List)
Annex B (Request for Information - Summary of Requirements)
Annex C (Request for Information - Questionnaire)

NATO Communication and Information Agency
Boulevard Leopold III
110 Brussels
Belgium
www.ncia.nato.int

ANNEX A

Distribution List for Request for Information
RFI-CO- 115878-DATA

All NATO Delegations (Attn: Investment Adviser)

NATO Members Embassies in Brussels (Attn: Commercial Attaché)

NCI Agency – All NATEXs

NCI Agency – (reserved)

ANNEX B

Summary of Requirements

1. Project Scope

1.1 The NCIA is exploring the potential for industry to perform a complex data preparation and analysis tasks on a several hundred terabytes (TB) of data appearing in diverse formats. The main objective of this task will be to identify and mark the duplicated items for given data set. Two options are being considered for this activity.

- Under option A, all work is performed by the contractor at their premises with facility clearance, on their CIS equipment with appropriate security accreditation, by staff holding a Personal Security Clearance (PSC).
- Under option B, all work is performed by the contractor at NCIA's premises, on NCIA provided CIS equipment, by staff holding a PSC.

1.2 Both options require industry to:

- Receive and ingest data;
- Reconstruct the data sets;
- Process the data sets;
- Extract the metadata of the items in data sets;
- Migrate/export to the formats suitable for digital preservation
- Provide the tool for end users that:
    - Report the summary statistics of ingested dataset
    - Report on results of identified duplicates, redundant libraries and non-record items within the datasets;
    - Enables the action to "keep", "merge" or "delete" data items

- Create an archival record set, consisting of records and metadata of permanent value, in accordance with NATO directives

2. High level requirements – Option A

2.1 All work to be carried out at the contractor premises, on the contractors CIS.
2.2 The contractor must have an operating environment with necessary accreditations and approvals to handle NS data.
2.3 The contactor must develop workflows or analytical components that perform items de-duplication, extraction of metadata and extraction of data sets from NATO Functional Area Services
2.4 The work should prepare the records and related metadata of permanent value in accordance with NATO directives
2.5 The work should be coordinated and in accordance with the direction of JFCBS Archivists to whom the **remote access** to the environment should be provided to the

processed data and, if viable, workflows/analytical components. JFCBS will validate results of analytics and data sets, assess progress, and answers RFIs if needed.

2.6 Contractor should provide an IT operating environment, with necessary accreditations and certifications regulated by NATO Security policy that will support following:

- Data transportation - from NATO to contractor facilities and vice versa

- Physical facility with required security approvals and accreditation up to including NS:
    - Storage with capacity to store more than 200TB of data
    - Computing hardware that can process more than 200TB of data

- Operating environment that host the tools should enable following activities:
    1. Extract the content and metadata from the following items:
        - Virtual disks (vmware)
        - File system
        - Zip Archives (zip, tar…)
        - SharePoint database – backup files and running files
        - FASs databases –to define the records (*please see a mandatory FASs in Table 1)
    2. Extract the metadata from the files (for example properties of an office file like a document classification)
    3. Perform deduplication based on defined rules and outputs
    4. Visualize the deduplication process output:
        - Overall summary statistics (for example % of overlaps; number of duplicates; unique file size vs all file size):
        - URIs for binary same items; duplicated count; metadata
        - URIs for high similarity items in content (for example 70% of similarity score and above); metadata
        - Folder level (Library) for duplicated items; percent of binary same or similar items; folder metadata
    5. Enables following actions to the end users:
        - Remote access to the environment
        - Setting up the rules to label and show the items lists (for example – select the rows with items/folders with a % overlap above 70% and file types in the list (pdf….))
        - Applying the rules on the sample versus full set (label the rows according to the rules)
        - Ability to modify rules and re-execute
        - Ability to export the decision sheet (pdf) – list of the items that need to be kept, deleted of merged
        - Executing the decision to “keep”, “merge” or “delete” items
        - Saving/exporting the items labelled with “keep” and “merge” to (save to different medium)

    (*merge – keep a version of the item content and merge metadata from identified duplicated items; keep – keep the items and metadata; delete – delete the items and metadata)

    - Prepare data sets for export to preservice that consist of:
        - Original file structure (library structure) or defined by JFCBS Archivists Content files in “archive” format (configuration provided by archivist)

- Metadata files (alongside content files)
- Potential conversion to long term preservation format like pdf-a, etc.

3. High level requirements – Option B

3.1 All work to be carried out at NCIA premises, on NCIA CIS.
3.2 The contactor must develop workflows or analytical components that perform items de-duplication, extraction of metadata and extraction of data sets from NATO Functional Area Services.
3.3 The work should prepare the records and related metadata of permanent value in accordance with NATO directives.
3.4 The work should be coordinated and in accordance with the direction of JFCBS Archivists to whom the remote access to the environment should be provided to the processed data and, if viable, workflows/analytical components. JFCBS will validate results of analytics and data sets, assess progress, and answers RFIs if needed.
3.5 Contractor will be provided with access to an IT operating environment, with necessary accreditation, certification and an initial software toolkit. Contractor is permitted to install additional software in accordance with NATO security policy.

- The contractor should execute following activities:
  1. Extract the content and metadata from following item containers:
     - Virtual disks (vmware)
     - File system
     - Zip Archives (zip, tar…)
     - SharePoint database – backup files and running files
     - FASs databases – NATO support required to define the records (*please see a mandatory FASs in Table 1)
  2. Extract the metadata from the files (for example properties of an office file like a document classification)
  3. Perform deduplication based on defined rules and outputs
  4. Visualize the deduplication process output:
     - Overall summary statistics (for example % of overlaps; number of duplicates; unique file size vs all file size):
     - URIs for binary same items; duplicated count; metadata
     - URIs for high similarity items in content (for example 70% of similarity score and above); metadata
     - Folder level (Library) for duplicated items; percent of binary same or similar items; folder metadata
  5. Enables following actions to the end users:
     - Remote access to the environment
     - Setting up the rules to label and show the items lists (for example – select the rows with items/folders with a % overlap above 70% and file types in the list (pdf….))
     - Applying the rules on the sample versus full set (label the rows according to the rules)
     - Ability to modify rules and re-execute
     - Ability to export the decision sheet (pdf) – list of the items that need to be kept, deleted of merged
     - Executing the decision to “keep”, “merge” or “delete” items
     - Saving/exporting the items labelled with “keep” and “merge” to (save to different medium)

(*merge – keep a version of the item content and merge metadata from identified duplicated items; keep – keep the items and metadata; delete – delete the items and metadata)

6. Prepare data sets for export to preservice that consist of:
   - Original file structure (library structure) or defined by JFCBS Archivists Content files in “archive” format (configuration provided by archivist)
   - Metadata files (alongside content files)
   - Potential conversion to long term preservation format like pdf-a, etc.
   -

4. Functional Area Systems (FASs)

FAS	Full Name
JOCWatch	Joint Operations Centre Watch
ICC	Integrated Command & Control
JChat	Joint Chat
AMN-EP	Afghan Mission Network Enterprise Portal (including DHS)
DHS	Document Handling System
ANET	Advisor NETwork
JADOCS	Joint Automated Deep Operations Coordinated System
INTEL FS	Intelligence FAS
NITB	NATO Intel ToolBox
SDA	SIGINT Database Application
IJC Portal	IJC Portal

Table 1

ANNEX C
Questionnaire

Organisation name:

Contact name & details within organisation:

Notes

- Please **DO NOT** alter the formatting. If you need additional space to complete your text then please use the 'Continuation Sheet' at the end of this Annex and reference the question to which the text relates to.
- Please feel free to make assumptions, *HOWEVER* you must list your assumptions in the spaces provided.
- Please **DO NOT** enter any company marketing or sales material as part of your answers within this Request for Information. But please submit such material as enclosures with the appropriate references within your replies. If you need additional space, please use the sheet at the end of this Annex.
- Please **DO** try and answer the relevant questions as comprehensively as possible.
- All questions within this document should be answered in conjunction with the summary of requirements in Annex B.
- Cost details required in the questions refer to Rough Order of Magnitude (ROM) Procurement & Life Cycle cost, including all assumptions the estimate is based upon:
    - Advantages & disadvantages of your service/solution/organisation,
    - Any other supporting information you may deem necessary including any assumptions relied upon.

1. **(Applies for option A only)** Are you able to provide a physical facility that possess security accreditation up to including NS (described as option A)?

    a. If not, what will be the time line and the cost to establish the facility?

    b. Is the environment capable of working with datasets of 200 TB+?

2. **(Applies for option A only)** Do you have established processes for handling data up to and including NS considering following:

    a. Transportation, from NATO premises and back

    b. Storage

    c. Compute/process in in secure environment, ensure that hardening of the hardware has been done and tempest tested

    d. People possess current NS clearance

3. (**Applies for option A only)** Are you able to establish a secured and accredited connection from the NS environment on your physical facility (defined above) to NS AIS in order to provide a remote access to JFCBS archivist during the development phase?

4. Do you have experience in processing a big data sets?

5. Do you have previous experience in developing complex data pre-processing and analytics components? (Please provide the examples)

6. Do you have experience in working with disparate data formats (unstructured, semi-structured and structured data formats)?

7. Do you have people who are trained to develop and apply complex data workflows, including the use of Machine Learning when needed?

8. Do you have processes to limit access and use of data sets for agreed purposes only?

9. What methods and tools would be used to identify duplicates and close-duplicates?

10. What methods and processes would you use to include feedback from users during the development process?

11. Please state the Rough Order of Magnitude (ROM) for delivery of the above activities. ROM costs and timelines can be given for option A, option B or both.

Continuation Sheet Please feel free to add any information you may think that may be of value to NCI Agency in the space provided below. Should you need additional space, please copy this page and continue with the appropriate page numbers.	Page __ Of __