

AMENDMENT 4

QUESTIONS AND ANSWERS

Q15. Does the diagram shown in Annex 1 represent the *current* workflow or the *expected* outcome? Is the Data Lookup connection to the OCR/ICR intended to depict and what the three databases connected to the Data & Image Expert component are?

A15. Annex 1 represents *current* imaging workflow, and includes potential locations for accessibility remediation processes. Data lookup can be completed to support OCR, ICR and any type of form or barcode scanning. The 3 databases connected to Data & Image Export are repositories used to show options such as storing raw images for documentary evidence originals or storing images on a secured server accessible to clients.

Q16. Can you please provide sample documents and the resulting output of the current workflow shown in Annex 1 for current workflow and technologies?

A16. Current output usually consists of PDF or PDF/A documents (depending on clients' needs) as well as accompanying XML with a variety of metadata (including scanning information for documentary evidence, as well as any information extracted from scanned documents). Sample may be provided for the pilot project phase.

Q17. Can you please elaborate on the purpose and interaction of the OCR/ICR and the “data lookup” components shown in Annex 1 for current workflow and technologies?

A17. The data lookup usually consists in accessing a database or a reference file to decrypt barcodes, validate data lookup based on file numbers, or validate the precision of extracted information.

Q18. Can you please elaborate on the purpose and interaction of the OCR/ICR and the “Lookup against client system” components shown in Annex 1 for current workflow and technologies?

A18. Please see answers provided for Q15 and Q17.

Q19. Can you please specify which sections in European Union Standard EN 301-549 (2018) are required for the solution?

A19. Specific sections relevant to this challenge are Sections 9 (Web) and 10 (Non-Web documents).

Q20. How do you currently verify if an HTML5 or ePub3 document are compliant?

A20. A validation tool should be provided as part of the quality assurance process.

Q21. What do you use as a remediation console and revision tools in your current workflow?

A21. We are currently doing no accessibility remediation console, thus the need to integrate this tool in the proposed solution.

Q22. Is the Annex 1 diagram found in the RFP purely a conceptual diagram of how DISC sees this solution being designed or is this the required implementation?

What processes in this diagram are currently in place?

A22. All processes indicated are currently in place and are representative of our ongoing operational workflow, except for the two blue boxes (AI assisted enhancement and accessibility conversion) that indicated potential accessibility implementation. These should not be seen as the only option for implementation; we are open to any possibility that satisfy our requirements.

Q23. In the import phase, can you please describe the incoming mail, forms, archive documents, fax documents and electronic documents that are the document inputs to your existing system.

A23. Scanning material is quite varied. It goes from very clean, highly-organised folder and bankers boxes of letter or legal-size documents, to more damaged, bound or stapled, unorganised documents. It also includes incoming mail (including correspondence, forms, cheques, etc.) and often require initial filing and classification steps before imaging.

Q24. In the import phase, please provide Individual descriptions and indicative samples of each document type. This would be very helpful in our analysis.

A24. Publications: Bound and unbounded publications. Bound publications can be scanned with book scanners or table top scanners, or be unbounded and scanned on high-speed IntelliScan scanners

Government Forms: Many models of forms coming from employees our citizens. Forms may comprise text boxes or other information that need be extracted and inserted into XML file.

Program correspondence: Incoming mail includes correspondence that is usually scanned and processed through OCR/ICR.

Archived collections: Most of our volume comes from archived collection that usually include large amount of similar documents (RH files, security screening files, archived dossiers, meteorological or scientific data, etc.). Information extraction is dependent on client's needs and use of imaged material. Collections may be scanned for archival purposes or to be integrated to a database.

Faxes: May include forms, correspondence and general business communications.

Q25. In the import phase, please describe if there are any consistency to these incoming documents so that they may be automatically identified as a specific document type for specific down-stream processing. For example, there may be a significant percentage of incoming mail that are completed forms of a specific type and as such could be identified and processed in a specific way. Is this what the automatic classification step does? If so are specific OCR/ICR processing configurations for each of these identified document types in place?

A25. We generally have several ongoing imaging processes happening simultaneously through our imaging process. Each collection will possess a certain degree of consistency that can range to very specific (a single form received by fax from which we must extract the same information repeatedly) all the way to highly variable (ex: HR archives comprising a variety of documents grouped into folders and boxes, for which only OCR is used to provide search capacity but no data extraction is completed. Projects are usually identified throughout the process using headers and separators that provide information to the imaging system (scanners and support systems) as to which profiles to be used. Initial identification document's properties and origin is usually done manually, either based on its origin (PO box for incoming mail, bankers box identification or

Q26. In the import phase, please provide as many relevant document type samples as you can and the approx. percentage of overall volumes that they represent.

A26. Samples will be provided to support development of pilot project in Phase 1.

Q27. Are there any consistency of images within or across document types that could possibly automate the addition of alt-text in certain situations? If so, what percentage of image alt-text could be identified and applied automatically today? What is your goal for AI assisted alt-text application within such documents for the future.

A27. Consistency between documents is only found inside of specific collections, and may never be applicable to other, future collections. Furthermore, since we deal specifically with governmental publications, most the images we process consist of graphs, bars and charts. Thus we are looking at solutions that make use of AI to properly decode these information-rich images and provide relevant data in an accessible format.

Q28. What are the average annual and monthly page volumes today and what are the average annual and monthly page volumes expected in the future? Are there any peak volume periods that we should be aware of?

A28. We process up to 40 million images per year. Our volume is dependent on major collections being imaged, as well as variations in incoming mails from our clients. It is difficult to evaluate volume of documents that will be going through accessibility remediation at this point, but expect smaller collections and material that has operational value, rather than large archival collections.

Q29. What are the average required amount of person hours intervention per document is required today after scanning? What is your target expected required future amount of person hours intervention per document when this project is completed?

A29. Depending on the degree of quality assurance and keying required by client, current process may take anytime between 5 seconds to 5 minutes per page. We do not have a target for accessibility remediation, but given our high-volume imaging model our intention is to minimize human intervention to reduce cost per page and accelerate remediation processes.

Q30. Can you specify what types of documents (images) will be subject to these treatments?

- Management documents or recent communications: invoice, statement, certificate, contract, mail, etc.?
- Forms filled in by hand and typed?
- Heritage documents: letters, registers, plan, books, newspapers, etc.?

A30. Please see answers provided for Q23, Q25 and Q31.

Q31. Can you specify which physical media the images came from and how they were digitized (scanning material, resolution, depth: gray level, color)?

A31. Scanning material is quite varied. It goes from very clean, highly-organised folder and bankers boxes of letter or legal-size documents, to more damaged, bound or stapled, unorganised documents. It also includes incoming mail (including correspondence, forms, cheques, etc.) and often require initial filing and classification steps before imaging. Digitization specs may vary depending on clients' need and the types of documents to be imaged, but are usually 8-bit bitonal or greyscale for typed documents, and 24-bit color for documents that include images, graphs or maps, with resolution varying between 200-300dpi.

Q32. Are examples available to illustrate these types of documents?

A32. Samples will be provided to support development of pilot project in Phase 1.