- In the tender document you refer to 'OCR in the language of the source document' and elsewhere to 'OCR of all the printed English language text'. Is there any expectation around OCR output of Arabic or other non-English language text?

  There are a few examples of images that contain both English and Arabic text. For the purposes of the response, OCR of non-English material is out of scope, however if we do decide that non-English content is to be processed, this will be a separate discussion and classed as a contract variation (so negotiated separately from the main agreement).

- Do you have estimates as to total percentages of materials in different languages?

  We have so far noted three types of documents in the total selection: a) Purely in English b) Purely in Arabic, c) A mixture of both. We also have a mix of printed and handwritten text. We have not been able to do a reliable estimate of percentages however, and digitisation is still underway, so any estimates would not be complete at this point. At this stage we are assuming all content provided will be in English and responses should be based against that.

- What is your preferred treatment for any dual language printed documents, and any dual language manuscript items?

  If there are instances of images containing a mix of English and non-English content, we only expect the English elements to be OCR'd. Any non-English content will be subject to a further discussion.

- Do you have details as to relative percentage of printed and manuscript documents?

  As we are still in the process of digitising the material, we do not have any accurate estimates of the proportions that can be relied on. Our working assumption has been a 70/30 split (70% typed) although anecdotally the scanning teams have suggested that based on what they've seen so far it could be closer to 90/10.

- How do you wish to treat handwritten annotation of printed documents?

  We do not expect handwritten annotation to be processed. Any images identified for HTR will be primarily handwritten text, not typed text with annotations.

- Would you wish to see HTR of non-English language text?

  There are a few examples of images that contain both English and Arabic text. For the purposes of the response, OCR of non-English material is out of scope, however if we do decide that non-English content is to be processed, this will be a separate discussion and classed as a contract variation (so negotiated separately from the main agreement).

- Please note that the HTR technology we can bring to this project does not allow for transcription. As such, results cannot be unified with OCR outputs. This doesn't comply with the optional requirements you outline. However, it does mean that all manuscript documents will be full text searchable. Can you confirm that consideration will still be given to AMD's HTR provision if this is the case?

Please submit your proposal based on what you can deliver. It will then be evaluated appropriately although the HTR component is an optional requirement and will not be part of the formal evaluation.

- If we were to be unsuccessful in bidding to deliver both Core Requirements and Optional Requirements would you consider a separate bid for Optional Requirements alone?

  The contract will be awarded to the supplier that can best deliver the core requirements. If the contracted supplier cannot deliver the optional requirements to meet our needs, we reserve the right to award the optional requirements under a separate contract to a different supplier.

- Can you confirm whether or not you expect accuracy levels will take into account blank spaces when determining accuracy at page level?

  We will normally award a higher score to the option that produces output that is truest to the source material. This includes the retention of blank spaces. This will be consistently marked across all submissions so if components of the test cannot be processed by all responders, this will be reflected equally within that scoring.

- How do you aim to treat accuracy levels where image legibility is compromised by damage or poor scanning?

  As a general principle, we will be comparing output from different respondents against each other. So, if no respondent is able to recognise such text then everyone will get similar marks on this aspect, but if one or more respondents is able to more accurately capture such material then their scores will reflect this.

- For how many images do you consider scanning of bound volumes results in text being lost in the gutter?

  Imaging is still ongoing so we cannot be definitive. However, most of the material is not in bound volumes so we do not expect any issues with text being lost in the gutter.

- Would you consider recommendations re. your file naming conventions?

FCO_8 (folder)
Content (folder)
1056_Accuracy_spreadsheet.xlsx
1056 (piece folder)
PDF (folder)
1056_0001.pdf
1056_0002.pdf
RTF (folder)
1056_0001.rtf
1056_0002.rtf
txt (folder)
1056_0001.txt
1056_0002.txt
XML (folder)
1056_0001.txt
1056_0002.txt

The ITT specifies that the final output structure for the project will be defined at the setup stage, however the above is indicative of how we have asked for information on previous projects.

- Can you confirm what XML output format you would be likely to require?

  This is the suggested schema:

  <agda_project>
  <dept_series>FO_301</dept_series>
  <piece>2501</piece>
  <image_reference >2501_0001</image_reference >
  <ocr_content>
  OCR data from image
  </ocr_content>
  </agda_project >

- We presume you would not require styling applied to .rtf format outputs – e.g. for tabular content. Can you confirm?

  No, we don't need styling

- For the test output would you like to see XML output as well as PDF and RTF?

  No, XML is not a requirement for the test output.

- In the Required Information about the Test Outputs you refer to access images. Does this mean the PDF output?

  Yes.

- How is the volume and series classified in the overall Input? Is this based on document type/category or any other criteria?

  The series classification is as follows:

  Department: *e.g. FO*
  Series: *e.g. 123*
  Piece: *e.g. 1, 2, 3 … (note that a piece can be a bound volume or loose-leaf pages of varying size and quantity)*
  Item: *e.g. 1, 2, 3 …. (not all pieces have a sub category of 'items')*
  Images: *e.g. 1, 2, 3 …*

- Will the Average size of the file exceed 30MB? The file size mentioned in the ITT final document is 30MB per Image provided

  30MB is the average size, not the maximum size of a file so yes, some files will be bigger and some smaller. As digitisation has not yet completed, the 30mb size is our best estimate to date.

- Which part of the image/document can be ignored or not applicable for OCR process? E.g. stamp, signature, strikeout, maps, drawings etc. (File Numbers – FCO_8_5_0004, FO_1016_3_0001, FO_1016_3_0003).

  Only body text within images should be considered for OCR. Any text associated within tables/maps/diagrams etc. is not expected to be processed. In images with both a map and body text, the text is still expected to be processed i.e. the exclusion is not necessarily at image level.

- Documents with text bleed fill fall under "Essential Requirements" or "Optional Requirements". Samples of these documents provided were very poor in quality. (IR_40_2_0005)

  These are classed under essential requirements. When compiling the test images, to the extent that it was possible, we aimed to include images that were reflective of the different type of documents that exist within the larger collection. If you find some test images that are so poor that OCR could not be applied successfully, then provided this is reflected across all of the proposals, our scoring will be consistent and not to the detriment of any single bidder.

- Samples include Images which are very dark and the content is very faded. These will be considered as Essential or Optional as per the scope?

  These are classed under essential requirements. When compiling the test images, to the extent that it was possible, we aimed to include images that were reflective of the different type of documents that exist within the larger collection. If you find some test images that are so poor that OCR could not be applied successfully, then provided this is reflected across all of the proposals, our scoring will be consistent and not to the detriment of any single bidder.

- Should Hand written Arabic content and signatures be digitally presented or needs to be converted to text?

  No, there is no expectation at this stage for any non-English content to be processed.

- TNA has mentioned that we should be supplying total characters for each document. When calculating total characters in a file, should handwritten and signatures be considered with it.

  No, if the content cannot be processed it should not be included within the total character count for that image.

- Are there any segmentation/zoning of images involved in the OCR process? If so, what are they?

  No segmentation or zoning is required, as apart from such things as annotations, maps and stamps etc., all English language data is to be processed.

- How is the data being consumed (or) utilized after the OCR process? Who are the end-users & what are their expectations?

  The OCR output will be used as part of a presentation website for the images. The OCR will be used by the search engine for that website to help users identify relevant content. Users will have a range of experience in researching archival material from academics with years of practice to members of the public with little or no knowledge of the processes involved.