INVITATION TO TENDER

OPTICAL CHARACTER RECOGNITION (OCR) SERVICES

DEADLINE FOR SUBMISSION OF TENDER RESPONSES TO
procurement@nationalarchives.gsi.gov.uk

5PM (UK TIME), MONDAY NOVEMBER 13TH, 2017

1. BACKGROUND

The National Archives (TNA) is at the heart of information policy - setting standards and supporting innovation in information and record management across the UK, and providing a practical framework of best practice for opening up and encouraging the re-use of public sector information.

This work helps inform today's decisions and ensure that they become tomorrow's permanent record. The National Archives is also the UK government's official archive, containing 900 years of history, from the Domesday Book to the present, with records ranging from parchment and paper scrolls through to recently created digital files and archived websites. Increasingly, these records are being put online, making them universally accessible.

Further information about TNA can be found on our website at www.nationalarchives.gov.uk

The Arabian Gulf Digital Archive is a joint multi-year project between The National Archives of the UAE (NA) and The National Archives of England and Wales (TNA) to create the leading online resource for historical information that relates to the Arabian Gulf region (UAE), Gulf Cooperation Council states and the Arabian Peninsula). The website will bring together collections from a network of international archives whose records about the Arabian Gulf will improving access to authoritative primary source material for UAE citizens and researchers.

In the first phase of the project, TNA and the NA will partner to build a digital hub that will make available half a million high-quality images of records about the UAE and about the states of the Trucial Coast before it.

These records will be from the collection of TNA and available to view as the original primary source material. Additional cataloguing information will be created in both Arabic and English to allow for dual language searching and increased accessibility to the material.

Future phases of the project will add further content from TNA as well as other archives which have relevant content to the region.

The launch is scheduled for 2 December 2018, to coincide with the celebration of the National Day of the United Arab Emirates.

Digitisation of the aforementioned images has already commenced. **This Invitation to Tender (ITT) relates specifically to the automated data capture of the English language text contained within these images.**

2. REQUIREMENT

This document is an Invitation to Tender (ITT) for the supply of Optical Character Recognition (OCR) services to TNA.

Optical Character Recognition (OCR) software will be applied to all documents where possible to enable deep text searching (i.e. all typed material).

OCR output will be provided in the language of the source document.

We expect that there will be a proportion of English language handwritten records within the series that NA has chosen for inclusion. In addition, certain records or images within records may contain partial or complete Arabic script.

This OCR tender involves the processing of up to an estimated 500,000 images. These images are spread across 4,870 pieces.

For the purpose of the outcome from this tender, TNA has some core 'essential' requirements and some additional 'optional' requirements.

TNA recognises the challenges associated with delivering the additional 'optional' requirements and would urge any and all Respondents not to be put off by these 'optional' requirements if they do not feel they are able to deliver on these.

Our essential requirements comprise:

- Optical Character Recognition (OCR) of all the printed English language text contained within the supplied scanned images.
- Manual editing and enhancement of the raw OCR output mentioned above.
    - A high minimum level of output accuracy is a key priority for TNA. During the course of this tender exercise, TNA would like to establish what levels can be achieved by manual enhancements and at what cost. We understand that this would be dependent on many factors, not least on the expected level of accuracy from the raw output.
    - As part of their responses, Respondents must state what the highest minimum accuracy levels they can guarantee as a result of their OCR and manual enhancement processes.
    - Respondents must provide cost quotations for minimum accuracy levels of 70%, 80%, 90% and above 95%
    - Respondents must also outline how these accuracy levels are calculated and how they propose to reach these standards.
- Output delivered in the specified format.

Our optional requirements comprise:

- Automatic capture of any English language handwritten text present within the images by means of Handwriting Text Recognition (HTR) or any other similar processes. Note that the output from any such process should be unified with the output from the OCR process listed in the core essential requirements above in order to supplement and enhance this OCR output.
- Manual editing and enhancement of the raw HTR output mentioned above.
    - A high minimum level of output accuracy is a key priority for TNA. During the course of this tender exercise, TNA would like to establish what levels can be achieved by manual

enhancements and at what cost. We understand that this would be dependent on many factors, not least on the expected level of accuracy from the raw output.
    - As part of their responses, Respondents must state what the highest minimum accuracy levels they can guarantee as a result of their HTR and manual enhancement processes.
    - Respondents must provide cost quotations for minimum accuracy levels of 70%, 80%, 90% and above 95%
    - Respondents must also outline how these accuracy levels are calculated and how they propose to reach these standards.

TNA does not intend to prescribe any particular processes to achieve the desired outcomes. Rather, we rely on the expertise and innovation of the Respondents to propose a solution that would best meet TNA requirements and achieves the optimal results.

TNA does however require that Respondents are able to articulate their processes in a sufficiently clear manner to allow TNA to evaluate their proposed solution. This means that:

- Respondents must provide a description of the software, hardware and human resources that will be utilised to carry out the HTR and/or OCR process.
- Respondents must detail their OCR/HTR processes. These processes should include steps taken to prepare the master images prior to the OCR/HTR as well as the actual processes and any quality assurance checks and measures that would be put in place.
- Respondents must specify what average output they think they can achieve on the raw output from their HTR and/or OCR Processes.
- Respondents must clearly state how these accuracy levels are calculated.
- Respondents must provide information about their OCR/HTR/manual editing process and capabilities.
- Respondents must describe the expected accuracy achieved using these processes and how the accuracy is calculated.
- Respondents may also describe alternative output formats that can be produced using their recommended processes.
- Respondents must list any industry standards or best practice guidelines that will be applied.

3. HOW TO RESPOND

Please submit your Tender Response to procurement@nationalarchives.gsi.gov.uk ***by 5PM (UK time) on Monday, November 13th, 2017.***

If you have any clarification questions, please submit these to procurement@nationalarchives.gsi.gov.uk ***by 5PM (UK time) on Friday, October 27th, 2017***.

Your Tender Response must be presented using the following headings in the following sequence:

1. Proposed Service
2. Resourcing
3. Delivery Plan
4. References
5. Pricing

Proposed Service

Describe in detail how your proposed service will meet TNA's requirements, as specified in **Annex A** to this ITT. Within your response please ensure that you include an Executive Summary, summarising in no more than two pages the key aspects of your proposed service and why TNA should contract with your organisation. The summary should identify a single point of contact for correspondence and should exclude any pricing information. You should also outline what you believe to be the risks associated with your ability to deliver your proposed service, and how you intend to mitigate them.

Please also supply the test output as specified in **Annex B** to this ITT. The test set of images referred to in Annex B is available upon request to procurement@nationalarchives.gsi.gov.uk

Resourcing

Provide details of how you will resource the delivery of your proposed service. You must provide details of your project organisation as well as detailed Curriculum Vitae for the proposed Project Manager and for other personnel with key project roles.

You must nominate a suitably qualified point of contact for all TNA queries for the operational phase, if you were successful in securing this contract from TNA. You must also outline your method(s) for management of review and change control policies.

Where you are acting for a group of respondents, you must provide a profile for each member in the group, indicating their role within the group.

Delivery Plan

Provide a delivery plan for your proposed service - from setup to completion - ensuring that you highlight any necessary inputs from TNA. You should also outline what you believe to be the risks associated with your ability to deliver your proposed service to the dates specified within the delivery plan, and how you intend to mitigate them.

You may use a GANTT chart or any other means of a visual representation to illustrate your delivery plan, but you must also provide a text based description of all elements of your plan.

Your plan must include, at a minimum, the following elements:

- Date for the commencement of setup.
- Date for the provision of sample output to demonstrate the validity of setup
    - Note that this sample must include the raw OCR output as well as an edited and manually enhanced output which meets all the minimum requirements for the output as set out in this ITT, including such things as minimum accuracy levels, format of output, file names and folder structure etc.
- Date for the completion of setup.
    - Note that the proposed schedule must include at least three iterations of sample provision before the setup can be signed off.
- Date for the commencement of live production.
- Date for the delivery of first batch of output to TNA.
- Expected quantity contained within this first batch of output to TNA.
- Total number of batches to be delivered to TNA.
- Date for each expected batch to be delivered to TNA.
- Quantity of output within each batch to be delivered to TNA.
- A schedule for dealing with rejected output from TNA.
- Date of delivery of last batch (including any rejected output that had to be resubmitted).

References

Provide details of **three** similar contracts that you have delivered in the past three years.

Pricing

Submit your pricing for your proposed service.

This pricing must include the following information, in this order:

- Setup cost – if not included in price per image.
- Cost for pre OCR image preparation – if required.
- Cost for raw OCR output per image.
- Cost for manual enhancement of data in order to bring the minimum accuracy to a certain level - for those Respondents bidding for this additional optional service. (Respondents must clearly set out the minimum level of accuracy they will be able to achieve and state their costs against this level. More than one level of accuracy and associated costs can be suggested).
- Costs for capturing handwritten text by means of HTR or any other automated method - for those Respondents bidding for this additional optional service. (Respondents must clearly set out the minimum level of accuracy they will be able to achieve and state their costs against this level. More than one level of accuracy and associated costs can be suggested).
- Cost of post OCR quality assurance processes per image.
- Cost for editing and manual enhancements to OCR output.
- Cost of media if USB hard drives supplied by TNA not to be used.
- Costs of returning USB drives and output.
- Any other costs.

You must include every price which TNA will be required to bear. If any prices are not specifically included in your Tender Response, it will be assumed that such services will be provided free of charge.

4. EVALUATION CRITERIA

Tender responses will be evaluated using the following criteria:

Proposed Service	40% of total available score
Resourcing	15% of total available score
Delivery Plan	20% of total available score
References	10% of total available score
Pricing	15% of total available score

5. TIMETABLE

Publication of Invitation to Tender	October 13th, 2017
Deadline for submission of clarification questions to procurement@nationalarchives.gsi.gov.uk	October 27th, 2017 (5PM)
Deadline for TNA to provide responses to clarification questions*	November 3rd, 2017
Deadline for submission of Tender Responses to procurement@nationalarchives.gsi.gov.uk	November 13th, 2017 (5PM)
Contract Award Decision	w/c November 20th. 2017

*TNA reserves the right to share the responses to clarification questions received with all potential suppliers

6. CONTRACT TERMS

This Contract will be awarded under our standard terms and conditions.

Please note that the information you supply in your Tender Response may be used, in whole or in part, to populate the Contract. As such, please make clear and unambiguous statements about the commitments you are making.

Please further note that:

- *The successful Supplier must confirm that TNA will retain ownership of all source images and all material derived from them.*
- *The successful Supplier must confirm that they will destroy all source images and derived material, including backup or duplicate digital copies, after an agreed time (e.g. six months after the final payment) or at TNA's request.*
- *Any changes to the agreed Delivery Plan must be agreed with TNA at least 30 working days ahead of the change.*
- *The successful Supplier must issue invoices for all completed work on a monthly basis.*
- *Invoices for the payment of any batches of output must only be issued once the batch has been QA'd by TNA.*
- *Invoices must only be raised for those outputs which have successfully cleared TNA QA.*
- *Any output that fails TNA QA must be corrected and delivered as part of a subsequent batch of output and invoiced for once that subsequent batch has cleared TNA QA.*
- *For any data that is delivered to TNA but does not pass TNA QA, TNA will provide a list of piece and item reference to Supplier for correction.*
- *All supplier invoices issued to TNA must contain at least the following information in order to assist TNA in tracking payments:*

Service Provided	Department	Series	Piece	Date of delivery	Delivery batch no	No of images processed	Price per image	Total
e.g. OCR	e.g. ADM	e.g. 127	e.g. 1	e.g. 01 December 2017	e.g. 01	e.g. 500	e.g. £0.20	e.g. £100

ANNEX A

Source Images: Technical Specifications

All the images will be captured and supplied by TNA with the following technical specifications:
Bit depth: 24 bit colour
File format: TIFF Version 6
Compression: Uncompressed
Resolution: 300dpi
On average, the image size would be around 30KB.

Source Images: Directories and File Names

Most scanned images will be saved under the following naming convention: piece_item_image no

For example:

1551_1_0001

The images in each new folder, whether piece level or item level, will always begin at image 0001. Of the pieces that are being scanned and require OCR, there are some that are itemised, and others that are not. For those images that are contained within non-itemised folders, the naming convention will be as follows:

1552_0001

Non-Itemised Folder Structure:

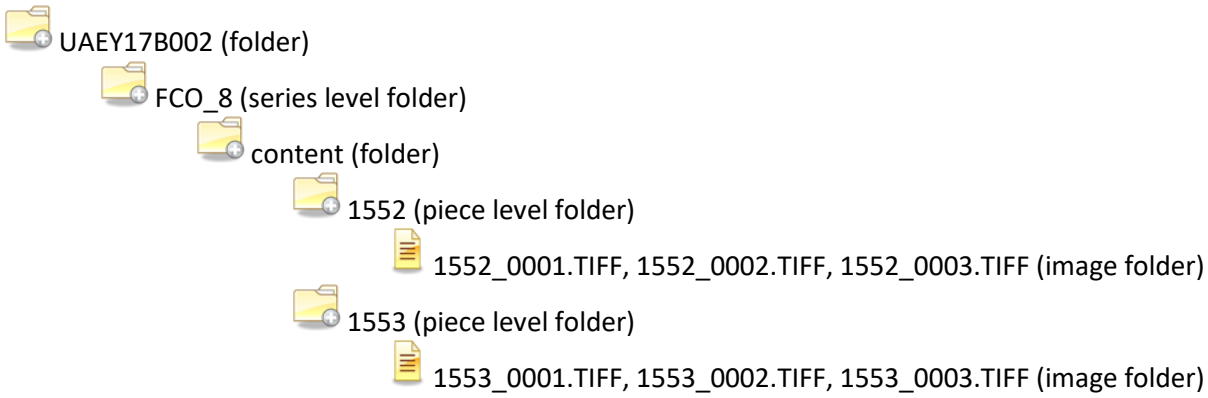
Non-itemised pieces will be saved under the following directory:

Batch code\Series_ref\content folder\Piece\Images

For example:

UAEY17B002\FCO_8\content\1552

The folder structure for the above directory would look like this:



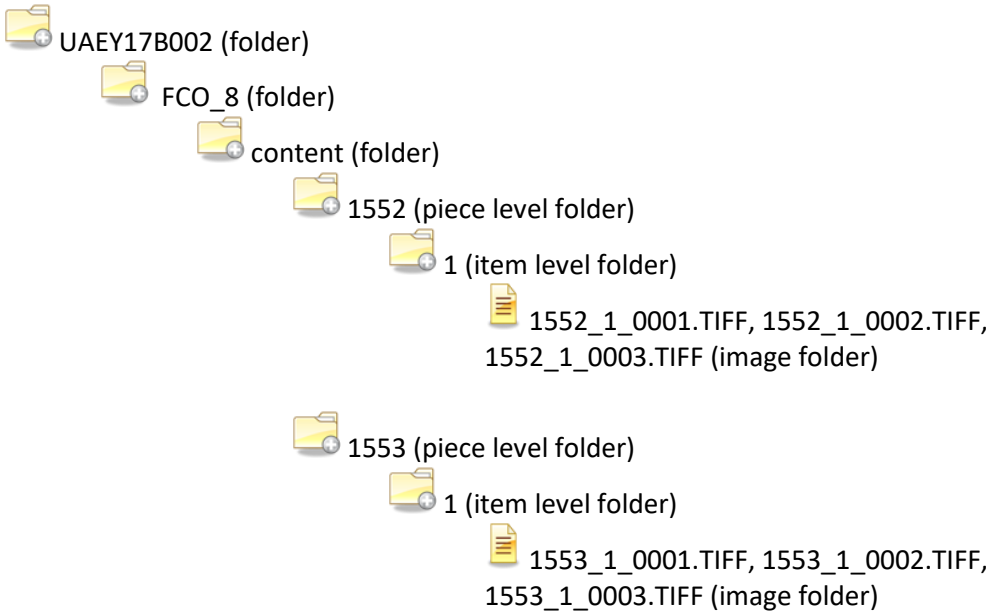
Itemised Folder Directory:

Batch code\Series_ref\content folder\Piece\Item\Images

For Example:

UAEY17B002\FCO_8\content\1552\1

The folder structure for the above directory would look like this:



Source Images: Quality and Content

The images requiring data capture are derived from a number of different series and so can have a wide variety of differing content and conditions.

We have not checked all the images for content, but have carried out random spot checks. Our samples were taken from the following series:

- FCO 8
- FCO 9
- FCO 31
- FO 93
- FO 464
- IR 40
- CAB 158
- ADM 127
- FO 1016

Some of the key features to note from our findings include:

- Colouring: Varies. High number of white, off white or yellowed documents, but we also found blue, green and pink.
- Size: The majority of the pages look to be approximately A4 in size, while there are also square documents, small slips of paper, and documents that are landscape orientated.

- Bound or loose leaf: The majority of images were taken from loose leaf documents, but we also found images relating to a number of bound volumes. Within these latter types of images, there weren't very many instances of text getting lost in the gutter. However, we did see a handful of images where this was the case.
- Date range: The earliest date we found was 1829. The latest was 1970.
- Text types: There is a wide variety of types of text. This includes typewritten text in various ink colours, printed and stamped text in various ink colours, and handwriting in various colours of pen and pencil. The style of handwriting is not consistent, as it seems like these documents have been written on by a number of different people at different times. It isn't at all uncommon to see documents that contain 3 or 4 different types of handwriting as well as printed text. We found a large number of examples of printed text being crossed through by hand, with annotations written in the gaps between lines or in the margins.
- Font sizes: Varies, but in the majority of cases the text is of a readable size. Some references are in smaller type, and certain styles of handwriting are smaller than others. But for the main part the size doesn't tend to be too small to read. (And in the case of article headings or headlines in newspaper clippings there are larger fonts present). Some types of handwriting are rather difficult to decipher, either due to poor penmanship or because the writing style is old fashioned.
- Text bleed through: We only found a few examples of this, generally in documents that are in poor condition in general. The majority of the images we looked at did not contain text bleed through.
- Language: Most of the documents seem to be in English; however there were plenty of examples of documents that were solely in Arabic, as well as some that were in both languages.
- Condition: From the percentage check we did most of the source documents looked to be in a fairly good condition. However, there were some examples that stood out as being of very poor quality. Types of damage included tears (both in the middle of the document and on the edges) which sometimes obscured text, staining which sometimes obscured text, and crease lines. In some cases the scans are very dark, as they are scans of a copy or a poor mimeograph. In many of these cases it is very difficult to make out the text due to the darkness of the image. We also saw several instances of very light scans, where the text is too faded to be readable (this was due to the condition of the original document), and some instances of the text being skewed one way or another. In some documents the text is only patchy in certain places, but this still makes it difficult to read.
- Content of documents: Varies widely. Examples of the types of documents we saw:
    - Front pages of folders: forms that have been filled out by several people.
    - Typewritten letters featuring handwritten annotations.
    - Small slips of paper with a single reference, or only a couple of lines of handwritten text.
    - Reports with additional slips of paper pasted over certain paragraphs.
    - Newspaper clippings pasted to larger pieces of paper.
    - Letters written in Arabic with the English translation either alongside or underneath.
    - Proofread articles with multiple handwritten notes, handwritten insertions, and sections crossed out, etc.
    - Printed tables containing text and numbers in multiple columns, as well as hand drawn tables
    - Reports written in standard paragraphs, in multiple columns and broken up by lists
    - Letters with additional pieces of paper attached via staples or glue
    - Floor plans
    - Flow charts
    - Family trees
    - Drawn maps

- Letters and reports make up the bulk of the material, but each piece will contain documents of other types as well.
- There were also a handful of documents that had some sort of wax seal and ribbon attached to the top corner. In general, if a document had something affixed to it that was obscuring text there would be a following image of the same document just with the obstructing item moved to the side or flipped to show what was underneath. Where this second image isn't present we can assume that whatever was attached cannot be moved.

Output Format

The platform on which these images and data will be loaded has not yet been developed. As such, we require output format(s) that would provide maximum versatility and adaptability. We would thus require the output to be delivered to us in the following formats:

- Searchable PDF V2.0 (resolution to be the same as the source files)
  - The output from each image should be delivered as a separate PDF file.
  - We would require compression to be applied to each PDF. The level of compression can be agreed upon at project setup stage.
- One set of output to be delivered as XML output.
  - TNA will provide a template for this XML output at setup stage.
- Output to also be delivered as a text file in .RTF format.
  - The output from each image should be delivered as a separate .RTF file.
- Any other format the respondents can recommend

In addition to the above output, TNA would also like to receive accuracy statistics relating to the raw output.

We understand that these accuracy statistics will be generated by the OCR software and will be based on confidence level rather than actual confirmed accuracy. Furthermore, these statistics will be out dated by the time the raw output has been through the manual enhancement process. Therefore we only intend to use this as a guide and not as a hard measure of accuracy levels.

Supplier should provide a single spreadsheet with each delivered piece containing accuracy statistics in the following format:

image number	TotalCharacters	CharacterAccuracy (%)	TotalWords	WordAccuracy (%)	SuspectWords
5100_0001	1000	98	80	98	2
5100_0002	9000	76	500	96	3
5100_0003	6000	82	200	84	5

Respondents will endeavour to deliver the above files with the highest possible quality standard.

TNA wishes to establish Respondents' capability in relation to optical character recognition of page content. TNA also wishes to investigate options relating to the additional information that can be encoded within the outputs.

The supplier recommended file format:

- May be a text format
- Should consist of a single file per image page image
- Must encode the full text of the source image (in other words everything recognised)
- May encode an estimate of the accuracy of the OCR process
- May encode any additional metadata that is available (e.g. publication, page number, etc)
- May be an XML format (for example, NDNP ALTO version 1-1-041) appropriate for digital preservation
- May also be another type of PDF
- May include bounding-box coordinates for each word (i.e. the coordinates of the word on the image page)

When suggesting the recommended output format, the Respondent must:

- Specify the advantages and limitations of the suggested format.
- State whether the machine-readable text file format includes bounding-box coordinates for each word.
- State whether the machine-readable text file format includes an estimate of the accuracy of the OCR process. If the machine-readable text file format does not include an estimate of the accuracy of the OCR process, respondents must indicate how information about accuracy is provided.
    - TNA would require accuracy count on character, word and page level.
- List any other metadata encoded using the machine-readable text file format.
    - Metadata may include the publication title, date of publication, page number, and format details.
    - The metadata may include details of the OCR process.
    - The metadata may include details of the digitisation process extracted from TIFF tags.
- Respondents may list any other machine-readable file formats they are capable of supplying.
- Respondents should provide details of the image processing software used to generate the output images and data
- Respondents must describe any web browser requirements for the specified format for primary access images and must specify any necessary plug-ins.

Output File Naming Convention and File Path

It is not possible to define the file naming convention at this stage as it is not clear what file format will ultimately be adopted and how many associated files there are likely to be with each output. The output file naming convention must be agreed upon by mutual consultation between TNA and the Supplier at project setup stage.

The folder structure of the output should remain the same as the input folder structure as far as this is possible.

Again, this must be agreed upon by mutual consultation at setup stage, as it is likely that some additional folders might need to be embedded in this existing structure.

In any case, there must be a completely separate (but similar) folder structure for the output to ensure that the input files and output files are kept separate.

Transfer of Images and Data

Source and output material will be transferred from and to TNA using BitLocker encrypted external USB 3.0 hard drives that will be supplied by TNA. TNA is prepared to consider an alternative delivery and return option.

See below for details of the image delivery schedule.

Delivery Schedule

Images will be delivered to Supplier in batches. The scanning of these images has already commenced. The scanning output schedule is listed below, and the batches can be delivered to Supplier accordingly.

- Batch 1: October 2017 (78,036 images)
- Batch 2: November 2017 (59,086 images)
- Batch 3: December 2017 (90,000 images)
- Batch 4: January 2018 (90,000 images)
- Batch 5: March 2018 (90,000 images)
- Batch 6: May 2018 (Will include the rest of the images: approx. 103,000)

As part of their tender response, and based on the above delivery schedule, the Respondents must outline a similar schedule for the delivery of output to TNA.

OCR Processing

As part of their tender response, Respondents are to provide an overview of the processes they will use to generate the required output files.

- Respondents must describe the processes required to prepare the master images (pre and post OCR processing)
- Respondents must describe in detail their manual editing processes, including how the decision will be made on which output requires editing, how this editing will take place and what results will be achieved as well as any quality assurance processes that will be put in place.
    - Respondents must provide information about the minimum accuracy that will be achieved in both the raw output OCR data as well as the final edited sets using these processes and how this accuracy is calculated.
    - We appreciate that there can be a difference between the accuracy of the actual data and the accuracy of output reported by an OCR software. OCR software reported accuracy is based on confidence levels rather than on an objective analysis by an independent observer – i.e. a human being.
    - Once the OCR data has been delivered to TNA, it will be QA'd by human observers; therefore to avoid any potential conflict it is important that as part of their tender response, Respondents provide a detailed explanation of how they would measure accuracy levels and what in your understanding would constitute an error – either on character or word level.
    - This accuracy of the output data, as it will be assessed by individuals who are independent of the OCR software, must by definition be based on the actual accuracy of the data and not on confidence levels.

    - Over and above the raw output, TNA also has some additional optional requirements that Respondents are encouraged to consider and include in their bid, if they possess the required knowhow and expertise. These additional optional requirements include:
    - Capturing of handwritten English text by HTR or similar automated means.
    - Manual editing and enhancement of the output in order to increase the level of accuracy to specified levels.
- Respondents that wish to take on one or both of these additional optional requirements must outline the processes that will be adopted to achieve them.
- For the requirement relating to manual enhancements, Respondents must specify the minimum level of accuracy that can be achieved through these processes.
- For the differing levels of desired accuracy listed above, Respondents must outline each of these clearly along with any price and time implications
- Respondents must also detail the resources that would be utilised in order to fulfil this requirement.
- Respondents must also clearly state which type of text capture can be included within this process and which, if any will be excluded.
- Respondents must describe in full, the process that will be adapted to capture this handwritten text.
- Respondents must list any industry standards or best practice guidelines that are applied.
- Respondents should provide details of the software used and human resources involved in all of the above processes.

Supplier QA

- Respondents must specify if they are bidding only for the delivery of TNA's 'Core Requirements' or also the 'Optional Requirements'. If the latter, then which of the two elements the Respondents are bidding to deliver.
- For each of the requirement elements, and for each level of service for a given element (if there is more than one), Respondents must specify the minimum level of accuracy that can be achieved.
- This accuracy level must be based on each individual scanned image that TNA will deliver to Supplier.
- Within a single image, Supplier must clarify whether the accuracy levels are based on word or character count, as well as how the accuracy level has been calculated.
- Alongside the output, the Supplier will deliver to TNA an accuracy report generated by the OCR software. However, this accuracy report must be treated both by TNA and the Supplier as a QA guide only and not the definitive level of accuracy.
- Where necessary, the Supplier should have in place a robust plan of QA and editing to ensure that each output file meets the minimum accuracy level proposed within their response.
- As part of their response, all Respondents should provide a detailed account of their QA checks, which should include their methodology of measuring the output accuracy level and the processes they will implement to ensure the minimum level of accuracy is achieved.

TNA QA

Upon receipt of output, TNA will carry out its own quality checks. These checks will be carried out over random samples. The size of these samples will be established once we have received the first batch of delivery from Supplier.

The typical checks during this QA process will include:

- Files are in the agreed format and comply with the specification for that format.
- Directories are correctly named and follow the correct directory structure.
- Files are correctly named and in the appropriate directory.
- The Supplier's output accuracy estimates are correct.
- During this QA process, where the error rate is significantly lower than the agreed margin for any given output file, TNA will reject these files and the Supplier will be asked to correct these files and redeliver at their own expense.
- It is also possible that during the QA process, TNA identifies small errors that do not merit rejection, but nevertheless require correction. In such cases, it would be far more efficient for the TNA QA team to make amendments themselves rather than send back to Supplier. However, it is possible that these small changes might break the link between the data and image (if such a link exists). In anticipation of this risk, Respondents must include the following information within their tender response:
    - Confirm that Respondent agrees that any minor errors to the output data can be corrected by TNA staff on TNA premises.
    - If the above will not be feasible then Respondent must suggest the preferred approach under such a scenario.
    - If Respondents agree that changes can be made by TNA staff at TNA premises then please detail what type of changes would be appropriate and which elements (if any) must not be changed by anyone but the Respondent.
    - If a link exists between the OCR data and the source image (e.g. in a searchable PDF), and this link can be broken by amendments to the data then Respondents must include a methodology within their processes that will ensure that any such link can be re-established once both Respondent and TNA QA is complete and the output is ready to be accepted by TNA.
    - Respondents should state clearly who would be responsible for the re-establishment of any such broken links.

Failed Output

Deliverables that do not meet the requisite level of quality will be rejected, and the Supplier will be required to reprocess and replace at their own expense, some or all of the rejected deliverables. Acceptance by TNA will be a prerequisite for payment, which will be made in arrears (see section 6 above for payment details).

On notification by TNA of quality problems:

- The Supplier will produce a correction plan draft for discussion with TNA within 10 working days, which identifies the cause(s) of the quality problem, and the correction process proposed.
- TNA will accept or reject the correction plan within 10 working days of receipt of the plan at TNA (TNA to acknowledge receipt).
- Both TNA and the Supplier will use their best endeavours to refine and agree the correction plan, including a timescale for the plan.
- The Supplier will execute the agreed plan.
- In the event of a Supplier failure to propose a draft correction plan, or of TNA and the Supplier to agree a correction plan, both sides will have a further period of three months elapsed time - from the date of the failure to agree – in which to discuss and agree other appropriate remedial action. If such actions cannot be agreed by the end of this three month period, the contract will be terminated.

ANNEX B

Test Source Images

TNA will supply, upon request, a test set of images that accompanies this ITT. The test set consists of approximately 100 images, as specified in the following table.

Department	Series	Pieces	No. of Images
ADM	127	1	1
CAB	158	2	10
IR	40	2	10
	8	7	35
	9	1	5
	31	1	5
	93	2	10
	464	1	5
FO	1016	4	20
Total: 5	**Total: 9**	**Total: 21**	**Total: 101**

Respondents should note that the test images, while attempting to show the range of images in the project series, are not truly representative, and therefore should not be used to make generalisations about the quality of the collection as a whole.

Test Output

Respondents must generate from the source images machine-readable output as specified in the Output Format section of Annex A.

Test output delivered by each Respondent should mirror the services they are bidding for. Respondents that are bidding for the optional requirements should provide test output containing not just the manually enhanced OCR output, but also the manually enhanced HTR output. This output will be tested against the levels of accuracy that respondents are committing to.

In order to demonstrate different levels of service, or different pricing structures, the Respondent may provide additional sets of outputs, but each set must be supplied as a complete response to the requirements in this section.

Respondents must supply copy/or copies of the outputs to TNA as part of their Tender Response.

Respondents must warrant that the outputs supplied reflect their submitted pricing.

Test Outputs: Machine-Readable Text

TNA requires that all Respondents provide machine-readable text corresponding to each test image.

TNA will comparatively assess the test output provided by Respondents. Criteria will include OCR accuracy (word and/or character), conformance to specifications for file formats, degree of fit with specifications of TNA requirements as laid out in this ITT and impact on storage requirements.

Respondents must provide machine-readable text file/s for every page in the test set provided. The files must be prepared in accordance with the information provided in Annex A.

Test Outputs: Format, Accuracy levels, Directories and Filenames

Test Output Format

The purpose of this OCR test is to assess respondent capabilities against TNA requirements set out in this ITT. As such, we require output format(s) that would mirror, as closely as possible TNA requirements and respondent's proposed solution. We would thus require the output to be delivered to us in the following formats:

- Searchable PDF V2.0 (resolution to be the same as the source files)
    - The output from each image should be delivered as a separate PDF file.
- Output to also be delivered as a text file in .RTF format.
    - The output from each image should be delivered as a separate .RTF file.
- Any other format the respondents can recommend

Test Output Accuracy Levels

Respondents will endeavour to deliver the test output files with the highest possible quality standard. TNA would expect all files to reflect the high standards of output that the Respondents are committing to within their tender response.

In terms of accuracy level and Respondent QA, all the conditions outlined in the Supplier QA section of Appendix A will apply to this test output.

When the test output data reaches TNA, TNA will check the validity of this output by a combination of automated and manual checks. The output score will then contribute towards the overall tender response evaluation for each respondent.

Test output can only be delivered once along with the tender responses. There will not be an opportunity for Respondents to receive feedback from TNA and make any amendments to the test output before the evaluation is complete. Feedback can be provided at the conclusion of the tendering process at Respondent's written request.

Test Output Directories and File Naming Convention

As mentioned in Annex A, we will agree on the file naming convention for the main project at project setup stage. For the purpose of this test output, and to assist TNA with its evaluation, the Respondent is asked to return the test outputs using the naming scheme used by TNA for the test source material.

Required Information about the Outputs

Respondents must supply the following information about the outputs (in this layout and order).

For the machine-readable text files:

- File format
- Average file size
- The OCR accuracy for each raw output file
- The lowest and highest levels of OCR accuracy amongst the raw output set.

- Average OCR accuracy for all raw output files – broken down per series
- The OCR accuracy for each file post QA and manual editing/enhancement
- The lowest and highest levels of OCR accuracy amongst the edited/enhanced set
- Average OCR accuracy for all files once they have gone through QA and editing/enhancements – broken down per series

For any access images:

- File format
- Bit-depth
- Resolution
- Average file size of the access images

For the overall set of test files:

- Total number of output files
- Combined size of the output files.