**APPENDIX D – STAGE 1 REQUIREMENTS AND EVALUATION CRITERIA**

Stage 1 Tender Responses will be evaluated by applying the following criteria:

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Maximum Available ‘Pre-Weighted’ Score** | **Maximum Available ‘Weighted’ Score** | **Minimum ‘Weighted’ Score Required to Progress to Stage 2** |
| **CAPTURE** | 140 | 25 | 15 |
| **ACCESS** | 220 | 25 | 15 |
| **SEARCH** | 100 | 25 | 15 |
| **PRESERVATION** | 60 | 25 | 15 |

For each question in each Category, a (pre-weighted) points score between 0 and 10 is available unless specified otherwise. If your response mainly has the attributes of one score, but also has one or more attributes of a lower score, then you will be awarded the lower score. These points will be allocated applying the following criteria:

|  |  |
| --- | --- |
| **10 Points** | **Outstanding:**   * Potential Supplier has provided a response that addresses all parts of the requirement * Potential Supplier has provided evidence to support all elements of their response * The evidence supplied is convincing and highly relevant to the requirement * Potential Supplier’s response is clear and easy to understand * Where relevant, Potential Supplier has demonstrated a high level of capability to deliver new and innovative service approaches |
| **7 Points** | **Very Good:**   * Potential Supplier has provided a response that addresses all parts of the requirement * Potential Supplier has provided evidence to support most elements of their response * The evidence supplied is good and relevant to the requirement * Potential Supplier’s response is clear and easy to understand * Where relevant, Potential Supplier has demonstrated some level of capability to deliver new and innovative service approaches |
| **5 Points** | **Average:**   * Potential Supplier has provided a response that addresses most parts of the requirement * Potential Supplier has provided evidence to support most elements of their response * The evidence supplied has some relevance to the requirement * Potential Supplier’s response is clear and easy to understand * Where relevant, Potential Supplier has demonstrated limited capability to deliver new and innovative service approaches |
| **3 Points** | **Mediocre:**   * Potential Supplier has provided a response that addresses some parts of the requirement * Potential Supplier has provided evidence to support some elements of their response, but not all * The evidence supplied is weak and has limited relevance to the requirement * Potential Supplier’s response is not always clear and easy to understand * Where relevant, Potential Supplier has demonstrated limited capability to deliver new and innovative service approaches |
| **1 Point** | **Poor:**   * Potential Supplier has provided a response that fails to address most parts of the requirement * Potential Supplier has provided little or no evidence to support most elements of their response * The evidence supplied is very weak and has very limited relevance to the requirement * Potential Supplier’s response is not always clear and easy to understand * Where relevant, Potential Supplier has demonstrated little or no capability to deliver new and innovative service approaches |
| **0 Points** | **Fail:**   * No response provided |

Stage 2 Tender Responses will be evaluated by applying the following criteria:

|  |  |  |
| --- | --- | --- |
| **Category** | **Maximum Available ‘Weighted’ Score** | **Minimum ‘Weighted’ Score Required** |
| **CAPTURE** | 10 | 7 |
| **ACCESS** | 10 | 7 |
| **SEARCH** | 10 | 7 |
| **PRESERVATION** | 10 | 7 |
| **SERVICE LEVELS** | 10 | 7 |
| **COMMITMENT TO CONTINUOUS IMPROVEMENT** | 10 | 7 |
| **STANDARDS** | 10 | 7 |
| **PRICE** | 30 | N/A |

The instructions and requirements for Stage 2 Tender Responses will be shared with short-listed Potential Suppliers once the evaluation of Stage 1 Tender Responses has been completed.

|  |  |  |  |
| --- | --- | --- | --- |
| **STAGE 1 – CAPTURE CATEGORY** | | | |
| **Ref.** | **Requirement** | **Information Required from Potential Suppliers** | **Available Pre-Weighted Score** |
| CA1 | **Messaging** – The Supplier must accept and generate XML files in accordance with TNA’s needs. The message exchange process is as follows:   * TNA generates orders for website crawls in XML format. These are known as XML1 messages. TNA transfers these files to The Supplier via FTP. The Supplier must use these to instruct their crawler(s) to reliably and efficiently capture websites within a defined scope and level of quality. * As The Supplier moves through various phases of the archiving process, The Supplier produces XML2, XML3 and XML4 messages and transfers them to TNA.   Examples of these XML messages and more details about them and the messaging process are supplied in Note 1 to this Appendix. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| CA2 | **Scheduling** – The Supplier must provide a solution through which TNA can request the crawling of multiple websites, in parallel, through a combination of:   * monthly batches of crawls, requested at the beginning of the calendar month, which are progressed through the various quality assurance processes and published separately once these processes are completed. * short notice, ‘exceptional’ crawls, that require immediate launching and prioritising through to publication. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| CA3 | **Crawls** – The Supplier’s solution must be capable of:   * 2,000 (number of) crawls of conventional websites per year. A conventional website is defined as a website that TNA and the Supplier have agreed can be captured using traditional crawling and access software, though it may still need special intervention to make the quality of the capture acceptable. * 800 (number of) crawls of social media services per year. A social media service requires archiving via APIs and then rendering the content through a custom interface. Currently, TNA harvests content from the Twitter and YouTube services using this method. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| CA4 | **Crawls** – The Supplier’s solution must produce very high quality crawls which meet or exceed the “state of the art” in the web archiving field. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| CA5 | **Crawls** – The Supplier’s solution must be capable of:   * launching crawls within a default set of parameters (either encoded in the XML1 messages or, in the case of crawls requiring special intervention, in “special instructions” supplied by TNA) * where default parameters do not result in crawls of sufficient quality, providing TNA with a means of defining the necessary deviations from the default set of parameters to improve quality * identifying all crawls made on behalf of TNA with a valid and agreed user agent, which identifies the crawler as part of the UKGWA. * writing, or storing, all web archive crawl data in the WARC file format, in compliance with ISO 28500:2009 (<http://archive-access.sourceforge.net/warc/>) | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| CA6 | **Crawl Scope** – The Supplier must configure all crawling operations in such a way that content extraneous to the target website is not captured during the crawling phase unless explicitly requested to be so by TNA. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| CA7 | **Social Media Crawls (Twitter)** – The Supplier must:   * capture the JSON Twitter API output of target Twitter accounts, provide access to it and preserve it as a JSON file. * develop and maintain scripts that:   + exclude all retweets, unless the capture of retweets is specifically requested by TNA   + can restrict Tweet captures for a target account to between two dates   + capture the images, videos and shortlinks (from shortlinks, through intermediary redirects to resolution) associated with each tweet * capture the initial, intermediary and final URLs for all URLs in the JSON code. * initially index all content separately for quality assurance purposes * preserve the captured content in standards-compliant WARC files and transfer them with the other WARC content to TNA. * check quality of these within the same QA timescales as with conventional websites. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| CA8 | **Social Media Crawls (YouTube)** – The Supplier must:   * capture all the video files associated with a target YouTube account. * develop and maintain scripts that:   + capture metadata for each video   + capture the associated thumbnails for each video * preserve the captured content in standards-compliant WARC files and transfer them with the other WARC content to TNA. * check quality of these within the same QA timescales as with conventional websites. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| CA9 | **Failed Crawls & Crawl Logs** – The Supplier must:   * alert TNA to crawls that have failed within 24 hours of them failing. * generate crawl logs for all crawls, providing TNA with access to these crawl logs for a minimum period of 3 months after a crawl has completed. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| CA10 | **Quality Assurance (QA)** – The Supplier must perform QA on all crawls to ensure that all archived websites are as close as possible in completeness and quality to the live version of the target website. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| CA11 | **Issue Tracking** – The Supplier must provide an online issue-tracking solution for managing the web archiving workflow with TNA. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| CA12 | **Planning and Relationship Management** – The Supplier must hold regular meetings with TNA, either in person or by telephone, during which progress against priorities and technical developments will be assessed. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| CA13 | **Service Level Agreement** – The Supplier mustadhere to an agreed SLAfor allaspects of the‘Capture’ requirements (messaging, crawl launch SLAs, QA levels, crawl failures, crawl logs, communications between supplier and TNA and quality assurance processing times). | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| CA14 | **Continuous Improvement** – The Supplier must commit to aprogramme of continuous improvement throughout the contract period. TNA is committed to developing the service using iterative approaches and Agile approaches to make sure that the service meets user needs responsively, quickly and to a high level of quality. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| **TOTAL AVAILABLE PRE-WEIGHTED SCORE FOR ‘CAPTURE’ CATEGORY** | | | **140** |
| **TOTAL AVAILABLE WEIGHTED SCORE FOR ‘CAPTURE’ CATEGORY** | | | **25** |
| **MINIMUM WEIGHTED SCORE REQUIRED TO PROGRESS TO STAGE 2 OF THE PROCUREMENT PROCESS** | | | **15** |

|  |  |  |  |
| --- | --- | --- | --- |
| **STAGE 1 – ACCESS CATEGORY** | | | |
| **Ref.** | **Requirement** | **Information Required from Potential Suppliers** | **Available Pre-Weighted Score** |
| AC1 | **Replay** – The Supplier must provide an access system which faithfully replays all UKGWA content captured during crawling processes past, present and future. This will be a mixture of ARC and WARC preservation file formats. Some of these files have deviated slightly from prescribed standards and The Supplier must be able to handle these. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC2 | **Replay** - The size of the UKGWA is currently measured at approximately 4 billion URLs and is growing at approximately 30 million URLs per month. The Supplier must be able to accommodate the current size and this rate of growth going forward. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC3 | **Replay** – The Supplier must configure the UKGWA so that it is hosted on webarchive.nationalarchives.gov.uk, and replicates the current service’s persistent and predictable URL structure and URL-based browsing functionality. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC4 | **Replay** – The Supplier’s solution must serve content of any type and of any file size. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC5 | **Webpages** –The Supplier must provide a front end to the service that complies with W3C standards, the Government Digital Service Design Manual, TNA web guidelines and other applicable standards for web services. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC6 | **Webpages** – The Supplier must:   * provide a number of mechanisms through which the UKGWA is sufficiently signposted as a web archive. This includes initially maintaining the red UK Government Web Archive banner and [ARCHIVED CONTENT] titlebar prefix. It also involves working with TNA to refine, develop and extend these notifications beyond HTML pages to, for example, PDF files, during the course of the contract. * supply ‘index pages’ that, based on original URL, list and link to each dated capture of a resource. * allow TNA-branded access and error pages. * provide responsive designs so that users can successfully access the service on a wide variety of devices. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC7 | **Collection Management** – The Supplier must provide a secure online system with a user interface that allows TNA, and the supplier if necessary, to immediately restrict access to content without removing it from preservation files. This is so that the service meets the requirements of our Takedown Policy (<http://www.nationalarchives.gov.uk/legal/takedown-policy.htm>). The system must support wildcards/regular expressions and time-banded restrictions. The system must be able to hold millions of such rules to restrict access where necessary. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC8 | **Collection Management** – The Supplier must be able to supply a system that selectively gives access to content at URL level. This is essential for TNA to maintain the correct scope for UKGWA. This must support wildcards/regular expressions and time-banded restrictions. The system must be able to hold millions of such rules to grant selective access where necessary. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC9 | **Availability** – The Supplier must maintain the UKGWA and Web Continuity redirection service to the specification (see <http://nationalarchives.gov.uk/webarchive/guidance.htm>) and without interruption. This is an SLA. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC10 | **Memento** – The Supplier must maintain TNA’s Memento implementation. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC11 | **Application Programming Interfaces (APIs)** – The Supplier must provide APIsthat allow exploration of web archive data and metadata programmatically. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC12 | **Social Media Archive** – The Supplier must maintain and provide occasional developments to TNA’s social media archive (see <http://www.nationalarchives.gov.uk/news/929.htm>) | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC13 | **Search engine indexing** – The Supplier must permit indexing of the collection by web search engines. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC14 | **Testing and development** – The Supplier must provide multiple environments, separate from the main collection, for quality assurance, development and testing purposes. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC15 | **Availability** – The Supplier must provide very high service availability. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC16 | **Availability** – The Supplier must provide a solution that can serve high volumes of concurrent users reliably and quickly. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC16 | **Access Logging** – The Supplier must log all traffic to the UKGWA using either the Apache NCSA extended log format or W3C extended log format and supply copies of the logs to TNA (see preservation category). | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC17 | **Traffic Statistics** – The Supplier must provide real-time or near real-time traffic statistics for the UKGWA. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC18 | **Security** – The Supplier must have a DOS/DDOS plan and employ sufficient measures to provide a secure and robust service. TNA may, on occasion, ask The Supplier to accommodate security assurance measures, such as penetration tests, on the systems that make up the service. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC19 | **Standards** – The Supplier must conform to applicable web and ICT standards:   * W3C Web Accessibility Guidelines. * W3C Mobile Web Guidelines (where applicable). * Government Digital Service (GDS) Government Service Design Manual. * TNA’s web team’s style and branding guidelines. * UK Information Commissioner’s Office Cookies Guidance. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC20 | **Usability Testing** – The Supplier must support regular usability testing and user needs analysis exercises by TNA. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC21 | **Service Level Agreement** – The Supplier mustadhere to an agreed SLAfor all‘Access’ requirements (availability, support of access services, traffic logging etc.). | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC22 | **Continuous Improvement** – The Supplier must commit to aprogramme of continuous improvement throughout the contract period. TNA is committed to developing the service using iterative approaches and Agile approaches to make sure that the service meets user needs responsively, quickly and to a high level of quality. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| AC23 | **Hosting** - The Supplier must ensure that the hosting of the UKGWA is secure, highly reliable and capable of serving large volumes of concurrent users at high performance. Some periods may see fluctuations in the numbers of users of the service. TNA is open to suggestions from suppliers of either traditional datacentre, cloud-based hosting, or hybrid models. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | Not Scored at Stage 1 – for information only |
| **TOTAL AVAILABLE PRE-WEIGHTED SCORE FOR ‘ACCESS’ CATEGORY** | | | **220** |
| **TOTAL AVAILABLE WEIGHTED SCORE FOR ‘ACCESS’ CATEGORY** | | | **25** |
| **MINIMUM WEIGHTED SCORE REQUIRED TO PROGRESS TO STAGE 2 OF THE PROCUREMENT PROCESS** | | | **15** |

|  |  |  |  |
| --- | --- | --- | --- |
| **STAGE 1 – SEARCH CATEGORY** | | | |
| **Ref.** | **Requirement** | **Information Required from Potential Suppliers** | **Available Pre-Weighted Score** |
| SE1 | **Scope** – The Supplier must provide a search service to search across the entire textual content of the UKGWA. This must include all formats that are rendered as HTML/XHTML in browsers. These include, but are not limited to:   * Hyper Text Markup Language (HTM, HTML, XHTML) * Dynamically generated content (e.g. asp, aspx, php and jsp etc.) * Portable Document Format (PDF, PDF/A) * Microsoft Office Formats (DOC, DOCX, PPT, PPTX) * Open document formats (ODF, FODT, ODT, ODP) * Text files (TXT) * but provide the option to exclude certain file types as defined by TNA, for example XLS, XLSX, ODS, FODS, robots.txt, .js and, .json extensions, and any files which returned an HTTP status code of 400 or greater when initially captured. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| SE2 | **Scope** – The Supplier must index, and maintain an index of, all content matching the criteria specified in SE1 above. Some content will remain indexed, but will be hidden from search results if the original domain does not appear on TNA’s whitelist. TNA’s whitelist is a list of URL paths that are in scope for TNA. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| SE3 | **Scope** – The Supplier must exclude boilerplate content (i.e. menus, navigation bars and page footers) from the index in order to minimise ‘noise’. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| SE4 | **Functionality** – The Supplier must provide a search service that:   * supports English language stemming. * supports special characters such as (&,( , ), £, ? etc.). * includes technology which can potentially support the future creation and maintenance of user accounts to allow users to save searches for later. * provides functionality to match keyword search or exact phrases of up to 255 characters. * provides exact phrase search when the user encloses the search term in “”. * supports search by a known URL. * supports the export of a results set into reusable formats such as CSV and XML. * displays search result snippets only when they belong to an original domain that is on a list of domains supplied by TNA. This list will be regularly updated and acts as a ‘whitelist’ by displaying all URLs that match a pattern while hiding those that do not. The patterns on the whitelist will be analogous to the takedown rules in that they can:   + invoke wildcards. These can be brought to bear at sub-domain, domain and below sub-domain level.   + hide or reveal based on specificity of pattern.   + apply only between two dates (of crawl). * gives results from particular domains a weighting, or ranking, to elevate them in a results set. This can be modified periodically based on user needs analysis. * supports the sorting of results in a result set by relevance and by crawl date. * provides post-search filtering functionality of a result set, allowing filtering by file type (html, pdf, doc etc.), year of crawl, category (with capability to support more detailed taxonomies as they arise), government department, domain/sub-domain, and refinement within results set with additional terms. UKGWA crawl metadata or TNA domain knowledge can provide the necessary data to meet this requirement. When it is the latter, for example the government department or category, this data will be provided by TNA in CSV format. * in advanced search mode, supports entity coreference. So, for example, all the acronyms “DfES” OR “DfE” OR “DCSF” are queried when just one of them is queried. This functionality can be enabled or disabled by the user and is only available as an advanced search feature. A list of equivalents or analogous terms will be supplied by TNA. * is hosted at webarchive.nationalarchives.gov.uk and must point to results there * provides basic support for languages other than English; this must allow for the querying of terms with non-English language special characters and accents. * applies Optical Character Recognition (OCR) to .pdf files to enable them to be indexed in search. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| SE5 | **Indexing** – The Supplier must:   * index content in monthly batches based on the websites published in the preceding month. * index content, and provide access to that content through search to TNA, within 10 working days of the publication of the last crawl of the monthly batch in question. * before providing access to TNA, hide any content on a takedown master list, provided by TNA. * search the content using a set of keywords, supplied by TNA and designed to identify inappropriate material in the UKGWA, and provide a list of the resultant URLs to TNA. * provide, to TNA only, the updated search in a search environment with restricted access, thus allowing TNA to test the update. * improve the quality of the search index, based on feedback from TNA and user needs analysis. * merge the updated index with the existing publically-accessible index, thus ‘publishing’ it and therefore updating the index. This will happen within 5 working days of TNA accepting the update following testing by TNA in the restricted search environment. * exclude error messages from the search index. For this purpose, an error message is a page with an HTTP status code of 400 or greater. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| SE6 | **Takedown** – The Supplier must:   * be able to quickly take down search snippets from domains and specific URLs within an agreed takedown SLA. * be able to occasionally restore snippets previously taken down from search, on TNA request. * index, but filter outfrom the search service, content that is not on an agreed whitelist. The whitelist will change over time and will be updated periodically by TNA. * provide an online tool through which TNA can control which results should not be displayed in search snippets, depending on:   + a wildcard or specific URL (i.e. where a string or portion of a string matches, it is hidden from results)   + date of capture of the resource (all dates, specific date, or within range)   + file format of the resource (e.g. all .jpg images on the domain example.gov.uk) | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| SE7 | **Performance and availability** – The Supplier must provide a search tool that has very high availability and supports multiple concurrent requests. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| SE8 | **Standards** – The Supplier’s search solution must comply with best practice standards such as the OpenSearch 1.1 specification. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| SE9 | **Service Level Agreement** – The Supplier mustadhere to an agreed SLAfor all‘Search’ requirements (takedown, concurrent use, updates etc.). | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| SE10 | **Continuous Improvement** – The Supplier must commit to aprogramme of continuous improvement throughout the contract period. TNA is committed to developing the service using iterative approaches and Agile approaches to make sure that the service meets user needs responsively, quickly and to a high level of quality. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| **TOTAL AVAILABLE PRE-WEIGHTED SCORE FOR ‘SEARCH’ CATEGORY** | | | **100** |
| **TOTAL AVAILABLE WEIGHTED SCORE FOR ‘SEARCH’ CATEGORY** | | | **25** |
| **MINIMUM WEIGHTED SCORE REQUIRED TO PROGRESS TO STAGE 2 OF THE PROCUREMENT PROCESS** | | | **15** |

|  |  |  |  |
| --- | --- | --- | --- |
| **STAGE 1 – PRESERVATION CATEGORY** | | | |
| **Ref.** | **Requirement** | **Information Required from Potential Suppliers** | **Available Pre-Weighted Score** |
| PR1 | **Ingest** – The Supplier must:   * decrypt historic data delivered using LUKS (Linux Unified Key Setup) encryption. * ingest large volumes (over 100TB) of historic ARC and WARC files from a range of sources, with the ability to cope with minor variations from agreed standards. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| PR2 | **Storage** – The Supplier must:   * provide methods to ensure that data will not be lost in the event of software and/or hardware failure; the supplier must be able to evidence that data has not been lost. * maintain metadata which links harvested data with the relevant crawl instructions. * retain data derived from interaction with TNA (such as crawl instructions and quality assurance notes) and with users (such as web server log files), even where not required for ongoing business purposes, until that data has been exported to TNA. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| PR3 | **Export** – The Supplier must:   * provide the ability to export all harvested data in the form of standard-conformant WARC files (ISO 28500:2009) together with per-WARC CDX index files with an agreed set of metadata fields. * provide the ability to export additional, pre-agreed data (such as web server log files and crawl management logs) in standard, pre-agreed formats (such as NCSA log format, CSV, XML with schema, documented SQL, documented json). * provide data exports to TNA for long-term preservation at regular intervals; where the data export travels through public physical or electronic space, the data must be encrypted using a pre-agreed standard. * provide a means to verify that the transmission of data exports to TNA has been both successful and complete. * provide metadata documenting the data exported, to agreed standards. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| PR4 | **Service Level Agreement** – The Supplier mustadhere to an agreed SLAfor all‘Preservation’ requirements. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| PR5 | **Continuous Improvement** – The Supplier must commit to aprogramme of continuous improvement throughout the contract period. TNA is committed to developing the service using iterative approaches and Agile approaches to make sure that the service meets user needs responsively, quickly and to a high level of quality. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| PR6 | **Standards** – The supplier must supply preservation files that conform to the WARC file standard (ISO 28500:2009) and the relevant agreed log standard. | What approach would you take to meeting this requirement? What experience do you have in meeting this requirement for other clients with similar needs? | 10 |
| **TOTAL AVAILABLE PRE-WEIGHTED SCORE FOR ‘PRESERVATION’ CATEGORY** | | | **60** |
| **TOTAL AVAILABLE WEIGHTED SCORE FOR ‘PRESERVATION’ CATEGORY** | | | **25** |
| **MINIMUM WEIGHTED SCORE REQUIRED TO PROGRESS TO STAGE 2 OF THE PROCUREMENT PROCESS** | | | **15** |

**Note 1 – CA1: Messaging**

**Transfer of Files**

The message files must be arranged in a logical directory structure, preferably on The Supplier’s FTP site, for example Year/Month/XML1.

**XML1m and XML1e (crawl orders, from TNA to Supplier)**

There are two different types of XML1 message - XML1m and XML1e:

* XML1m: the ‘monthly’ crawl message is generated at the beginning of the month as a regular crawl of a website, and normally one of the crawls that makes up the schedule of crawls of a particular website.
* XML1e: the ‘exceptional’ crawl message can be generated any day and is intended for crawling high-priority websites where the crawling, quality assurance and publication processes need to be expedited because, for example, the website is closing, an important event in government has just occurred, or large quantities of information are about to be removed from the website.

These XML messages, generated from TNA’s archiving management database, the Government Websites Database (GWDB), are to be regularly transferred to The Supplier, preferably by FTP, and represent a “crawl order”, which triggers the launch of a crawl within an agreed timeframe SLA.

Each XML1 message has a unique ‘Crawl ID’ and defines a number of parameters that must be followed during the crawling process.

The name of the file will identify whether it is an XML1m or XML1e file. The filenames will be in the following format:

XML1m: YYYYMMDD-(am/pm)-TNA-1m.xml (e.g. 20160301-am-TNA-1m.xml)  
 XML1e: YYYYMMDD-(am/pm)-TNA-1e.xml (e.g. 20160317-pm-TNA-1e.xml)

Sample XML1m and XML1e are attached.

 

|  |  |  |
| --- | --- | --- |
| **Element** | **Role** | **Values** |
| CrawlID | Unique crawl ID | Numeric |
| CrawlSchedule | Date of crawl order | Date (YYYY-MM-DD format) |
| CrawlMode | Specifies method of crawl | Either “Automatic” or “Manual” |
| CrawlType | Specifies type of crawl | Either “Regular” or “Exception” |
| DomainCatRef | TNA use only | Alphanumeric |
| DomainID | Unique domain/website ID | Numeric |
| DomainType | The type of website it is | Either “Domain”, “Subdomain” or “Microsite” |
| DomainName | The name of the website (may not be unique) | Alphanumeric |
| DomainURL | The URL of the ‘parent’ domain or subdomain | Alphanumeric (URL) |
| EntryPoint | The URL at which the crawler should start crawling the target website | Alphanumeric (URL) |
| SpecialInstructions | Any instructions that must be applied in the case of a “Manual” crawl, for example to also include a list of additional URLs supplied by TNA | Alphanumeric (e.g. “Crawl with” |
| SyndicationID | TNA use only | Alphanumeric |
| SyndicationURL | The location of the target website’s XML sitemap URL, to be used by The Supplier to assist with capture. | Alphanumeric (URL) |
| AssociatedURLs | Additional URLs that should be considered in scope for the crawl are in the <AssociatedURL> child element of AssociatedURLs. They may or may not be microsites or subdomains of the DomainURL. | Alphanumeric (URL) |
| ObeyRobots | Instructs The Supplier to either explicitly obey robots.txt, or not. Where TNA is unaware of any special requirement, it will be “Use Default Settings” and this is the default setting. | Use Default Settings OR Yes OR No |
| AcceptCookies | Instructs The Supplier to either explicitly accept cookies, or crawl with cookies disabled. Where TNA is unaware of any special requirement, it will be “Use Default Settings” and this is the default setting. | Use Default Settings OR Yes OR No |
| CrawlInterval | Instructs The Supplier at which rate to crawl the target website. Where TNA is unaware of any special requirement, it will be “Use Default Settings” and this is the default setting. | Use Default Settings OR numeric, expressed in milliseconds |
| AlsoInScopeUrls | This lists, in <AlsoInScope> child elements of AlsoInScopeUrls, the protocol and www. variants of the domain URL, to ensure that both http://, https://, and with “www.” as well as without are all included. For example, if the EntryPoint is one of the four possible variants, the other 3 will be listed here and must be considered in scope for the crawl. | Alphanumeric (URL) |
| PaginationUrls | This defines where pagination is known to exist on a target website and may exisit in an XML1 crawl order so that sufficient depth of crawl can be achieved, which avoiding crawler traps.  This element has two child elements: <PaginationUrl> and <Range>. <PaginationUrl> contains the “path” attribute and specifies the URL that contains the variable portion that the <Range> child element defines. <Range> consists of “last”, “first” and “index” attributes. | Alphanumeric (URL) and numeric |