OCR Workflow at RBGE

8 April 2014

|  |  |  |  |
| --- | --- | --- | --- |
| **Task #** | **Task Name** | **Explanations and Comments** | **Resources**  **(human, physical tool, software)** |
| 1 | Select OCR software | ABBYY Recognition Server |  |
| 2 | Install OCR software | We had some issues with permissions, with our antivirus software and with our institutional data back-up processes. |  |
| 3 | Test and tune installation | We use the following settings:  Always active  Quality = maximum  Languages = English (we add other languages depending on the current imaging project, eg Turkish for the specimens from SW Asia and the Middle East, Spanish from specimens from South Temperate America)  Verification = off  Automatic quality control = off  Indexing = not specified  Output = text document with encoding type set to simple (not utf8 but maybe we should be using utf8?) & pdf document with ocr on top |  |
| 4 | Pre-sort specimens prior to imaging | We routinely OCR all our images immediately after the specimen is imaged. There is a level of presorting due to the digitisation processes which follows two main channels. Large digitisation projects, eg Mellon and Scottish Government funded have often concentrated on particular regions of the world or particular taxonomic groups. Our core digitisation is triggered by loan/image requests from researchers and these tend to be taxon-based. |  |
| 5 | Image specimens | We routinely image specimens at 600dpi or equivalent resolution. When we process the raw file to a tiff file we apply a minimal amount of sharpening. |  |
| 6 | Migrate or copy images to OCR workflow/staging directory | The images are processed automatically through an image management system which creates a copy of the tiff and sends it to the ABBYY dropbox folder where the OCR workflow automatically picks it up and processes it. We do not carry out any additional processing before the OCR step of the workflow. |  |
| 7 | Run OCR on Images | We use a “watched” dropbox folder that is permanently active. The output text file is automatically picked up from an “output” folder and is saved to an SQL database, linked to the specimen data and the specimen image. |  |
| 8 | Evaluate Quality of OCR Text | We do not persistently evaluate the quality of the OCR text, although we have carried out several tests. |  |
| 9 | Run any sorting algorithms? | We have scripts which allow us to search the OCR text for keywords. The resulting file locations are used to create a slideshows of the specimen images. The same results are also used to create database lists which open an identically ordered batch of records for data entry. We routinely use Word Clouds to help select suitable keywords. |  |
| 10 | Batch data entry | We have sometimes used batch data entry for collector names, country names. This is particularly useful when any additional data entry is not an option due to lack of resources. |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Task ID** | **Task Name** | **Explanations and Comments** | **Resources**  **(human, physical tool, software)** |
|  |  | Get help from folks already doing OCR (iDigBio aOCR wg!) |  |
|  | Select OCR software | Factors to use in making selection might include:   * Cost * Batch processing capability * Output types (XML, Text, etc.) * Language Support * SDK availability * OS platform (Windows, Linux, Mac etc) * APIs * Authority File Support * Watched Folders - allowing use across multiple users.   It will be helpful to review the next task to ensure adequate hardware/OS support. | Available OCR software and tools include:  Proprietary   * ABBYY Finereader * ABBYY Finereader Pro Edition * ABBYY Finereader Corporate Edition * ABBYY Recognition Server * OmniPage * Prime Recognition   Open Source   * Tesseract * J(G)OCR * OCRopus |
|  | Install OCR software | **Hardware/OS issues**   * Ensure that hardware meets requirements of the OCR software * Ensure sufficient memory capacity * Permission issues * Firewall settings * Server selection;  some servers might also be running multiple automatic processes or CRON jobs that might interfere with the OCR software * Adequate storage available * Ensure adequate IT support   **Software configuration**   * Character set configuration (e.g. UTF-8) * dictionaries * Set overnight scripts / processes * example: Tesseract dependent libraries for dealing with jpgs |  |
|  | Test and tune installation | Hardware tweaking   * memory allocation * scratch disk allocation * storage capacity   Software tweaking   * settings / options * image recognition * data dictionaries enhancement * font training (in ABBYY referred to as Pattern Editor - note that the image dictionary is compiled at the same resolution of the image you’re looking at.  If you try to apply the same image dictionary at a higher or lower resolution, this could cause errors in the OCR output) * pc vs server, running on a pc can be challenging | The tuning process often requires revisiting, changing, and optimizing settings made during installation, which may mean retracing the steps of one or more installation protocols.  Example: ScioTR configuration with local or shared OmniPage Ultimate installation (under development): <http://sciotr.com/Home/Help#ConfigureOmniPage> |
|  |  | transition row-core OCR |  |
|  | Pre-sort specimens prior to imaging | Some institutions might pre-sort based on language, geography, etc..   * Language & Geography (RBGE): Turkish. Separating specimens before Imaging for OCR gives better results because you set the language setting specifically. * Collectors (Symbiota Lichen and Bryophyte example): - allows efficient use of collector name data dictionaries. * physical nature of the herbarium * scope of the current digitization project   If pre-sorting is used, this task should be included within the task list of [Module 1](https://www.idigbio.org/sites/default/files/sites/default/files/Module_1_Pre-digitization_Curation_Tasks_0.pdf): Pre-digitization Curation module of the flat sheets and packets workflows. | Existing dictionary sources include:   * For location * Taxa, are there resources that could be placed here? * [Harvard list of collectors](http://kiki.huh.harvard.edu/databases/botanist_index.html) * [Names, places, abbreviations by Jason Best](https://github.com/jbest/darwin-score/tree/master/dicts) |
|  | Image specimens | Follow relevant tasks in Modules [2](https://www.idigbio.org/sites/default/files/sites/default/files/Module_2_Imaging_Station_Setup_Camera.pdf), [3](https://www.idigbio.org/sites/default/files/sites/default/files/Module_3_Imaging_Station_Setup_Scanner_0.pdf), and [4](https://www.idigbio.org/sites/default/files/sites/default/files/Module_4_Imaging_Tasks_0.pdf) of the flat sheets and packets workflow.  Some institutions elect to select label regions within the master image and create separate images. Some OCR software facilitate isolation of text blocks by supporting creation of templates that point to portions of the image that contain text.  Some OCR software (e.g. ABBYY) isolate text blocks automatically, rendering the manual selection of text blocks superfluous.   * make sure images are in crystal clear focus * resolution of image * use **pixel x-height** instead of resolution: see <https://www.idigbio.org/wiki/index.php/Technical_Issues> * Thoroughly test images with software since software will differ in tolerance (e.g. a black border around the label will interfere with Tesseract output) * example: Tesseract has more noise if specimen is in the image with the label * label takes up less space on the computer (file size) * running ocr on the label is faster than on the entire herbarium sheet. |  |
|  | Prepare images | * convert to JPGs * convert to grayscale (reduces pixel layers) * image compression * trim off label * contrast/brightness * may create a derived image for the OCR process (space, processing time issue) | software: Photoshop, Lightroom, etc |
|  |  |  |  |
|  |  |  |  |
|  | Migrate or copy images to OCR workflow/staging directory | some institutions move images into an OCR folder  JPG vs TIFF  most copy and maintain the original | Storage space for images |
|  | Modify image (partially software dependent) | Create image to enhance for OCR; e.g. Higher contrast. (proves helpful with Tesseract). AABBY does this automatically, so no image adjustment necessary. Some folks have to **crop out black borders if using Tesseract.** | {{RR}}: I think T2 and T3 can be combined - T3 is part of T2.  Omnipage has an optional ‘Enhance Images’ step that can be scripted and configured with an ‘enhancement template file’. |
|  | Run OCR on Images | Many institutions place images in a “hot” or “watched” folder that gets processed at night. This creates a collection of text files or database entries. |  |
|  | Evaluate Quality of OCR Text | May be programmatic or human evaluation; | Insert Github or other pointers to helpful software  <https://github.com/jbest/darwin-score>  LBCC scripts - Robert Anglin |
|  | Run any sorting algorithms? | * send OCR output to Google API for language detect to create datasets by language if labels are in multiple languages and not pre-sorted. * try sort by hand-writing detection method (see Ben Brumfield’s code) so images with handwriting go to humans; images mostly text and print have OCR that goes to parsing algorithms * search OCR output for keywords to create datasets * compile OCR output and create a word-frequency list (a word cloud) to inform what strings are available for creating data sets for entry and / or parsing (Elspeth’s G67 example) or **editing too!** |  |
|  | Run ML / NLP to populate database records | After OCR the text elements of the OCR should be assigned to different database elements. (eg. the scientific name on the label can be assigned to the DwC:ScientificName field) | this needs an evaluation and cleanup step too |
|  | Normalization of data prior to import into a relational database structure. | Many labels will have descriptions of locations, habitats, taxon names, etc that should perhaps be normalized prior to import into a relational collections database. For example, if you have digitized 10 specimens that were collected at the same locality and would like to merge these into your relational database, which has a locality table, with a unique list of localities. You would want to link these 10 new records with the appropriate locality in your locality table upon import, rather than importing 10 new locality records. The same would need to be done with events, taxon names, etc. | Much of this should be done with database scripts on import, but oftern it is not that easy. If ignored, this step can result in a whole lot of database cleaning later. If over-done, user will reply on label images to later sort out problems. |
|  |  |  |  |
|  | Bibliography |  |  |
|  | Websites |  |  |