Module 6: Data Capture

The underlying focus of the steps throughout these digitization modules is to encourage institutions to follow an object to image to data workflow through which all specimens are first imaged and data recorded from these images. Nevertheless, some institutions choose, for various justifiable reasons, to pursue a specimen to data workflow and we try to accommodate both approaches below.

Task ID	Task Description	Explanations and Comments	Resources
T0	Perform any preparatory steps.	Determine application to be used for data capture, taking into consideration community standards (especially the Darwin Core standard) and project and institutional informatics environment, requirements, and policies. The data capture application (and underlying database) may not be your primary institutional database. For example, one might capture data via an intermediate application/database (e.g., a web-based application underlying database) and then later import the data into the primary institutional database. Load images and skeletal records (if relevant) into application being used for data capture. Perform other tasks that facilitate data capture such as OCR. ● It is strongly recommended that a specimen record with minimal data (a skeletal record) be created in a prior module (most logically the Imaging Module, Module 4). This record must at least contain a barcode and preferably taxonomic and/or geographic data. This facilitates sorting and filtering	Data entry application. Images. Skeletal data. OCR software or OCR-integrated data entry application. See: TDWG Darwin Core Standard, http://rs.tdwg.org/dwc/index.htm.

		of records at later steps in this module. ● OCR processing should occur before manual data capture begins. Consider particular OCR software being used and how this integrates with software being used for manual data capture and other databases where the records will reside. When included, Optical Character Recognition (OCR) constitutes a subtask of at least the following steps: ● Ingest specimen or label image(s) into an OCR tool. ● Execute OCR on image(s). ● Import or insert OCR results into the data entry application. ● Process OCR results within the data entry application: ● Delineate regions of interest within the OCR output and identify the fields into which the text should be imported (e.g, Apiary), ● Clean, parse, format, and import text into a spreadsheet for later upload to the database (e.g., Salix), ● Display and copy text from visible OCR output (e.g., Symbiota). ● Verify and correct OCR errors (typically via manual keystroking). ● Archive corrected, unparsed verbatim text. It should be noted that OCR execution and processing (with the	

		exception of Symbiota's integrated and largely seamless OCR implementation) is often a batch process independent and external to an inline data capture workflow, the results of which are imported into a database to update existing records. Work is underway in the community to refine OCR accuracy and enhance OCR integration.	
T1	Determine extent of record level data fields to capture into the database.	The extent of data captured from specimens in a first pass ranges from skeletal (short) records that include a restricted set of elements to fully populated (long) records that include all label data, including annotations. Institutional policy varies widely in this regard, with some institutions restricting capture to a subset of fields to create a skeletal record and others capturing all data on the sheet, including determinations, annotations, etc. Decisions about what to include in a skeletal record are dependent upon numerous factors, including the composition and arrangement (e.g., geographic, taxonomic, collector) of the collection, an institution's expected plans for future processing and data completion (e.g., OCR, NLP, automated georeferencing), anticipation of additional data entry over time from images, commitments made to funding agencies (e.g. numbers and levels of records to be digitized, project intent, etc.), institutional focus (e.g., quantity or records completed vs. record robustness), potential use of current and developing search technologies for automated or assisted record completion, use of political boundary	Institutional or project policy, intent, and/or goals.

		centroids for first-level georeferencing, and/or intended reliance on specimen images to provide first-level serving of complete label data.	
T2	Queue existing image files previously prepared for data capture, or procure physical specimens for data entry.	Queuing images can take many forms. For example, record sets of skeletal data or OCR'd datasets associated with images could be filtered by criteria catered to the data entry person's interests, the project's focus, etc. If data are to be entered from specimen sheets rather than images, time must be allowed to move specimens to the data entry station(s). This may necessitate an additional terminating task in Imaging Tasks Module in which specimen folders are moved to a data entry staging area following imaging to eliminate the need to refile specimens then re-pull them at data entry time. Alternatively, if data entry precedes or occurs parallel with imaging, additional terminating steps may be needed in Module 1 or other modules for moving specimens to the data entry station. It should be noted that in some institutions both of these strategies are used concurrently, effectively accommodating a variety of pathways for specimens to arrive at the data entry station(s).	Computer. Software. Cart or cabinet for transporting specimens. 'Swing' cabinet. Drop tags. White boards with magnets. Images.
T3	Create new empty database record or find existing database records previously created in the Pre-digitization Curation or Imaging Tasks modules.	Some workflows may include creation of a skeletal record within an earlier module equivalent to what is detailed here, or such previously created records might include only a catalog number (e.g., barcode value). Hence, skeletal record creation might be skipped here, or previously created skeletal records	Computer. Database. Images or physical specimens.

		might be more completely populated at this step.	
T4	Enter catalog number or other identifier via keystroke or barcode scanner.	This task may have been completed during one or more previous modules, as suggested in T0,T3. If working from a queue, this step will not be necessary, for example.	Barcode scanner.
T5	Enter collector name, collector number, and/or collection date OR the exsiccati title and number, if applicable.	This data facilitates electronic search for duplicates. Attempt to use standardized look-up lists, when appropriate.	Database interface.
T6	Attempt search for duplicates.	In software so equipped, this process attempts to discover duplicate specimens or duplicate collecting events from within a regional or global herbarium network based on exact or closely similar matches on several fields (collector, collector number, collection date, exsiccati title and number). Software supporting such duplicate searches currently includes Specify 6 (via Scatter, Gather, Reconcile) and Symbiota. Even in cases where exact duplicates are not found, duplicate collecting events that are found might facilitate more rapid data entry.	Appropriate software. Connection to networked resources. A database that is a source of duplicates.
T7	Parse and move data from found duplicates into the data record.	This step is dependent upon completion of T6 and assumes discovery of a duplicate record or duplicate collecting event. Results might be used to fully populate–via keystroke or automatic transfer–previously partially completed records or to import discovered data into all empty fields.	Appropriate connection to networked resources.

T8	Attempt automated NLP.	Steps in the Natural Language Processing (NLP) process might include: ● Training/setup/configuration of grammars and parsing rules using training sets based on predefined formats and cases (e.g., dates, duplicates). This task is likely to be performed once or only periodically. ● Ingestion of data into the NLP tool (data to typically be the results of OCR, but possibly from keyboard input). ● Output of parsed data and subsequent upload into a database.	NLP software or NLP-integrated data entry application. Data source to be parsed.
T9	Enter specimen data for remaining fields being targeted.	Consider institutional or project policy when choosing target fields, including but not limited to higher geography, determiner, habitat, etc. Enter or select from controlled vocabulary pick lists. Currently keystroking is the most popular method. Some applications have NLPs and duplicate harvesting integrated into data entry form for assisted automated data entry techniques. Voice or speech recognition software is not yet widely used, but has important consequences for biological database data capture. Several institutions are currently using this technology and others are refining it for use with biological and paleontological collections. Using this technology requires training VR software to recognize and parse individual technicians' speech patterns (a one-time, repetitive, and potentially somewhat time-intensive endeavor). Following initial training and setup, steps in using VR mirror	Institutional policies and protocols. Voice recognition software. Computer and database.

		those of keyboard entry and sometime depend upon keyboard-controlled navigation among data fields. To capture data, technicians view the label, navigate to the appropriate data field in the database interface, and read the label data into a microphone. When used, VR allows data entry for filed-as name and other relevant label data, including the population of skeletal data referenced in T5. Significant time investment in training the software for rapid turnover of technicians is a potential deficiency of VR, especially in light of the time that is sometimes required to train the software.	
T10	Extract and record annotation label data via keyboard or voice recognition.	Capture of annotation label data during initial data entry varies with institution. Some herbaria defer this to a later data entry step, others create fully populated records in which annotations are included or populate skeletal records and annotations.	Institutional protocol. Voice recognition software. Computer and database.
T11	Check for specimens in need of repair or filed incorrectly.	Establish and follow protocol for repairing and rerouting specimens in the digitization process.	Specimen handling protocols.
T12	Manually verify results and correct errors.	After the data capture session and regardless of data entry method or combination of methods, data entries should be methodically reviewed for quality control. This task should be carried out on batches of records on a periodic basis (e.g., daily, weekly, etc.).	Quality control protocol.

T13	Record enhancement or secondary digitization.	There are several tasks that entail deriving data from the label or specimen. Such tasks include georeferencing, assessing phenology, obtaining DNA sequence data, etc.. Some of these tasks are covered in the Proactive Digitization and Georeferencing modules.	
T14	Programmatic processing to ensure validity of captured data.	Programmatic validation of specific data depends on software and automated electronic processes that can rapidly check for and alert technicians to inaccuracies. Such validations can occur in batch following entry of a set of records, or can be integrated via automatic processing at data entry time. Ideally, validation should be executed at various stages within the data entry process. Examples include validating country, state, county, geographic coordinates, taxonomy, and nomenclature. More specifically: ● geographic coordinates applied to records are within the appropriate geographic scope, ● taxonomy and nomenclature reflect appropriate spelling and are derived from standard sources, ● geographic names reflect correct spelling and are derived from standard sources. Automated data validation tools offered by data aggregators (e.g., Symbiota, GBIF, iDigBio) and repositories can be helpful with this task. The Kepler based Kurator Project promises to be helpful in this regard.	Quality control software.