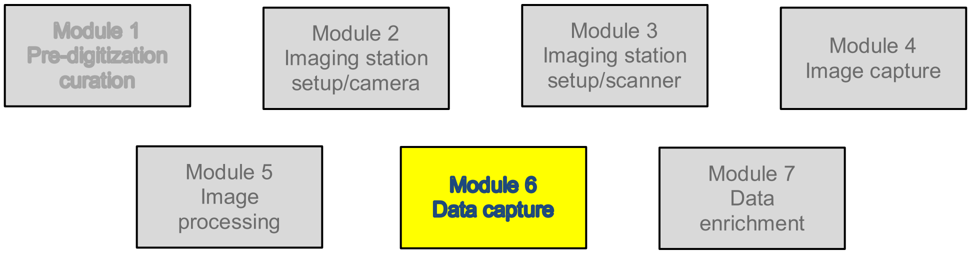
Workflow Detail: Data Capture (for flat sheets and packets)



Module 6: Data Capture

Task ID	Task Description	Explanations and Comments	Resources
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 skeleton (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 scientific name, collector name, collection date, state, and county (or equivalent) or town. Decisions about what to include in a skeleton record are dependent upon numerous factors, including 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 (Filtered-Push, Specify's SGR, Symbiota's Dup Check, etc.), use of political boundary centroids for first-level georeferencing, and/or intended reliance on specimen images to provide first-level serving of complete label data.	Institutional or project policy, intent, and/or goals.

T2	Queue existing image files previously prepared for data capture, or procure physical specimens for data entry.	The underlying focus of the steps throughout these digitization modules is to encourage institutions to follow an object to image to data workflow. Nevertheless, some institutions choose, for various justifiable reasons, to pursue a specimen to data workflow. 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 Module 4 in which specimen folders are moved to a data entry staging area following imaging to eliminate the need to re-file 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. Cart for transporting specimens.
T3	Create new empty database record or find existing database records previously created in Modules 1 or 4.	Some workflows may include creation of a skeleton 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, skeleton record creation might be skipped here, or previously created skeleton records might be more completely populated at this step.	Computer. Database. Image store or physical specimens.
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 T3.	Barcode scanner.
T5	Enter collector name, collector number, and/or collection date.	This minimum initial data entry facilitates electronic search for duplicates.	Database interface.

T6	Attempt search for duplicates.	In software so equipped, this process attempts to discover duplicate specimens from within a regional or global herbarium network based on exact or closely similar matches on several fields (collector, collector number, collection date). Software supporting such duplicate searches currently includes Specify 6 (via Scatter, Gather, Reconcile, currently for herbarium data only) and Symbiota. Even in cases where exact duplicates are not found, closely related records that are found might facilitate more rapid data entry.	Appropriate software. Connection to networked resources,
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. 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 OCR of label data.	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 to refine OCR accuracy and enhance OCR integration.	OCR software or OCR-integrated data entry application.

T9	Attempt automated NLP.	Natural Language Processing (NLP) can be classified mostly as a future tool on which work and research are occurring. It is included here for completeness. When effected, it will follow and depend on OCR execution. Currently, the Salix application, produced at Arizona State, combines OCR and NLP in an external application that creates a spreadsheet suitable for uploading to a database. Steps in the 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 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.

T10	Enter specimen data utilizing speech recognition.	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 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 skeleton 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.	Voice recognition software.

T11	Keyboard transcription.	Select filed-as name from a list. Enter or select from controlled vocabulary pick lists other label data per institutional or project policy, to include but not limited to higher geography, determiner, habitat, etc.	Institutional policies and protocols.
T12	Manually verify results and correct errors.	Regardless of data entry method or combination of methods, data entries should be methodically reviewed for quality control prior to writing the record to the database.	Quality control protocol.
T13	Extract and record annotation label data via keyboard or voice recognition.	Capture of annotation label during initial data entry varies with institution. Some herbaria defer this to second level data entry, others create fully populated records in which annotations are included or populate skeleton records and annotations.	Institutional protocol. Voice recognition software. Computer and database.
T14	Programmatic processing to ensure validity of country, state, county, geographic coordinates, taxonomy, and nomenclature.	Programmatic validation of specific data depends on software and 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 that: • georeferences 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 and repositories can be helpful (e.g., Symbiota, GBIF, iDigBio) with this task.	Validation software.

T15	Inspect specimen for damage.	This applies to specimen images as well as physical specimens. While capturing data from specimens, technicians can inspect the specimen for damage and re-route the specimen through the conservation workflow. If capturing data from images, a protocol should be established for notifying curatorial staff of needed conservation.	Specimen inspection and damage determination protocols.