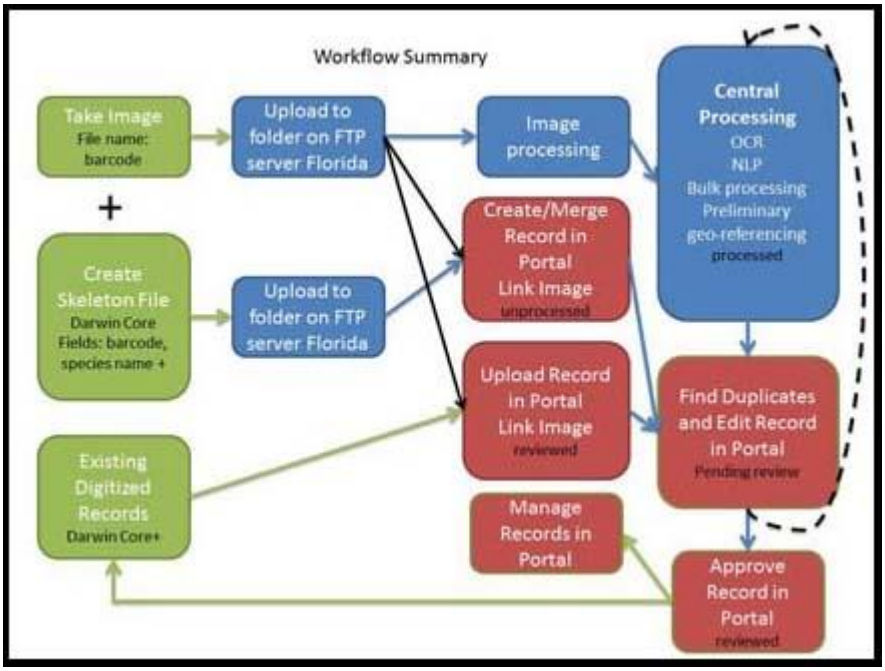
Lichens Bryophytes and Climate Change

Data capture will involve a semi-automated process consisting of the following steps (figure 1):

1. Imaging all specimen labels
2. Automated scripts will prepare images for label processing, storage and web access
3. Images of labels will be converted to text using Optical Character Recognition (OCR)
4. Text will be parsed into appropriate database fields employing Natural Language data Processing (NLP)
5. Human-assisted review with the ability to manually edit or enter the data as necessary
6. Semi-automated geo-referencing
7. Data quality control procedures

A team of at least two people collaborating with the local curator(s) or collections manager(s) will be responsible for imaging specimens at each imaging institution. A central team at WIS consisting of an IT coordinator and a transcription/geo-referencing specialist will be responsible for maximal automation of the image preparation, and the initial transcription and geo-referencing processes. Final editing of label and location information including entering information for handwritten labels will be handled in collaboration between the central institution, the imaging institutions and the institutions owning the specimens, employing transcription assistants and frequently help from trained volunteers. Local imaging teams will collaborate with the central team to develop standard protocols, which will be available on-line as video training modules and other documentation. For a more detailed description of how the web portals will be utilized to process the specimen labels, visit the documentation covering the web portals (http://lbcc.limnology.wisc.edu/node/6).



Institutional imaging teams will consist of a minimum of two people working together to capture images of labels, annotations, and notes for all specimens included in the proposal (for additional imaging of actual specimens, see below). Lichens and bryophytes frequently are archived in paper packets, multiples of which may be affixed to a larger herbarium sheet or they may be stored upright individually in boxes or drawers. Labels usually are on the outside of the packet with additional information like annotations, chemistry etc. in other places, including, within the packet. Therefore, imaging all information frequently will involve opening the packet and in some cases taking several images for each specimen. For increased speed and efficiency, cameras and light stands will be used to capture the images instead of using a flatbed scanner; this approach has been used successfully in similar projects by some of the participants (e.g., ASU, F, WTU). All specimens will be barcoded. Depending on the institution, the barcode will identify the specimen using a variable combination of a Globally Unique Identifier (GUID; InstitutionCode : CollectionCode : CatalogNumber). The barcode number will later be linked to the specimen in the database during the transcription process. Ideally, all images for one specimen will have the barcode label visible. This is possible if the barcode label has not been attached to the package, which should happen after images are captured. Images will be renamed using the barcode identifier and a letter suffix to denote multiple images of the same specimen. Typically, renaming the images using a barcode reader as part of the imaging workflow tends to be the most reliable method. However, the barcode can be captured using Optical Character Recognition (OCR) software to read the barcode as well as the text of the barcode. Experience has shown that a team of two people can capture images of labels, annotations, and notes for 300 and 400 specimens within a work day, which is less than reported for imaging vascular plant specimens due to the different storage methods. Since images will be taken in groups of specimens from each species, images will be stored in folders labeled with the institution and species name. At the end of each day, images organized in these folders will be uploaded to a central storage facility (central server in figure 1). For more information, visit the image page (http://lbcc.limnology.wisc.edu/node/4)

Once images are uploaded to the central server, processing scripts will manipulate them in preparation for web access, label transcription, and archiving. The barcode will be obtained from the image file name or directly from the image using OCR. In cases where OCR is employed to capture the barcode, the barcode as well as the text of the barcode will be read and compared as a verification step. The image file will be renamed to contain the barcode as part of its name.

The barcode will be obtained from the image file name and is used as the primary key in the database while the species name and holding institution will be obtained from the folder name and parsed into the appropriate database fields; thereby linking a database record to the image(s) and specimen. OCR is then used on typed labels to transcribe the label text from the image. Extensive experience at ASU (where the PI and co-PI worked until recently) and WTU assures that widely used fonts have an acceptable transcription rate with some known problems (e.g., the numeral one is frequently confused with the letter l). Initially this information is stored as one text block in the database record and then parsed applying Natural Language Processing (NLP) algorithms. To optimize NLP, extensive lookup tables will be developed containing collector names (including abbreviations and common misspellings), collection number formats, and date formats. Large thesauri are available and will be expanded to include additional species names with authorities, and geographic names. Within the CNALH, such a thesaurus

already exists containing all lichen species names from the Index Fungorum, Integrated Taxonomic Information System (ITIS), and the list of lichens for North America published by Ted Esslinger (2010) among other international sources. The CNABH has been building up a thesaurus of species names starting with ITIS taxonomy and augmenting the liverwort and hornwort data with a taxonomy supplied by F and the Early Land Plants Today project (http://www.mapress.com/phytotaxa/content/2010/f/pt00009p021.pdf). A moss thesaurus is currently being compiled from collaborating institutional resources with major inputs from DUKE, MO and NY. Extensive experience with this approach at ASU has shown that species name, collector name, collection number, and date can be parsed reliably in about 75% of labels. The species name will also be available from the folder name. Comparing the two and checking against entries in an authority table will provide a high degree of reliability. Additional improvements will be made on batches of labels that have the same layout, e.g., from the same collector, or when a herbarium has used pre-printed label forms. The system will be trainable and certain information may be parsed into appropriate database fields based on its location on the label. At this point of automation, considerable information is computer accessible and searchable. Records meeting minimum requirements can be published although still marked as needing manual proofing. These records are then made available in the consortia for limited analytical purposes. Duplicate specimens will be further processed using the FilteredPush approach (Macklin et al. 2009), which will enable the recognition of these specimen as low priority for manual checking and link all duplicates of a specimen to avoid redundancy in manual editing efforts. Geographic information can then be searched and specimens will be grouped for rapid geo-referencing.

To ensure data of high quality, the OCR and NLP results for each specimen label will be reviewed by personnel of the institution owning the specimen, a national network of volunteers, or hourly workers coordinated by a central volunteer coordinator. Handwritten labels and those that failed transformation via OCR for other reasons will have to be keyed into the database at this point. A web application is currently being developed for this purpose as part of the SYMBIOTA package. This application will allow the editor to view the label image and the OCR/NLP results in the database. He/she can then edit the database fields accordingly. Having this editing step on the web provides the opportunity for remote access to editing tasks. This will allow for major volunteer involvement comparable to the successful British program 'Herbaria@Home' (see outreach section for more detail).

TDWG-ratified geo-referencing protocols and standards (http://wiki.tdwg.org/Geospatial, Chapman and Wieczorek 2006) will be followed. Existing scripts will be used to obtain decimal coordinates for records with UTM and Township, Range, Section (TRS) information. BioGeoMancer and GEOLocate geo-referencing services will be used as appropriate. Geo-referencing will only need to be completed for ca. 30% or less of the specimens that have not been originally geo-referenced by the collector because collectors typically collect multiple specimens at any one location. This process will take place primarily during the final year of the project, thus increasing efficiency by allowing geo-referencing to be done as a batch process performed on the project's combined dataset. In this manner, coordinates can be accurately and efficiently assigned to multiple specimens that share matching locality descriptions. Based on earlier correspondence, we infer that the HUB will be involved in this process, as this will be central to all collections being digitized in the ADBC program.

Due to the unique storage of lichens and bryophytes, we cannot at this time provide specimen images together with the label images for the bulk of collections proposed to be digitized here. Since both lichens and bryophytes are usually rather small organisms that require a lens-view and often also microscopic details for taxonomic purposes, either high-resolution scans or direct macroscopic and microscopic imaging are required; high-resolution specimen imaging even with expensive cameras will not result in sufficient resolution to elucidate the necessary specimen details (this has been tested by participating institutions). Thus, meaningful specimen imaging will involve a time investment of at least 5 minutes per specimen (macroscopic shots only) and up to 15 minutes or more if microscopic images are necessary; in addition, such imaging can only be done by personnel with taxonomic training as it is necessary to take images that show representative details of a specimen. This effort would represent a taxonomic review and verification of specimen identification, which is not part of this funding request. Therefore, instead of bulk specimen imaging, we propose a mixed strategy where participants place already existing images at the disposal of the project and one or two selected and representative specimens will be imaged for each species by taxonomic experts of each participating institution as part of existing synergistic research programs. The consortium web sites allow for uploading such images and linking them to a specific specimen using the barcode and from there to the species in general using the species name.

1. **Imaging of Specimen Labels**
2. **Label Image Processing**
3. **Label Information Pre-Processing**
4. **Manual Label Processing**
5. **Geo-referencing**
6. **Specimen Imaging**