GLOBAL Digitization Standards - Minimum Expectations

1. **Imaging - PRIORITY ACTIVITY: Jan 2021**
    1. **Start with non-North American bryophytes and lichens.** The focus of the grant is specimens from outside Canada, Mexico, and the USA. Once these specimens are completed, remaining resources and time can be used to work on North American specimens.
    2. **Labels/Annotations (Label Image)** - Image the label information for specimens that are not already imaged or databased. No need to image labels that have already been entered manually into either of the portal databases.
        i. The barcode should be visible in every label image.
    3. **Specimen Image** - If the specimen has not already been photographed in the same image as the label, then collect one high quality image of each physical specimen.
        i. The barcode, barcode number, or another unique number (e.g. accession number) should be visible in the specimen image. If the paper with the barcode/number is loose and archival quality then the best practice is to retain it in the specimen packet/box for quality control purposes.
    4. **Institutions must keep their own archival files** [strongly recommended: JPG+original camera RAW file, not DNG (if DNG, do not discard original raw or embed it within the DNG)]
    5. **Recommendation for online images in CNABH and CNALH: file size max. 6MB, JPG, high quality (photoshop >8).**
        i. For **labels *only***: 1000px x 979px (can be color, but grayscale has smaller file size and is sufficient for OCR)
        ii. For **specimens *only*** and/or **specimens+labels**: typically no bigger than 4500px x 3600px (again 6MB maximum; thus for most cameras no down-sizing necessary, but a full size 45 megapixel jpeg file / 8256 x 5504 pixel might have to be downsized/cropped (rule of thumb: the bigger the images, the slower they load) as an example: this file is 3MB: https://serv.biokic.asu.edu/imglib/seinet/ASU/ASU0128/ASU0128577_lg.jpg
        iii. Example: a JPEG 20 x 15 cm, 300dpi has pixel dimensions of 12MB, but at a JPG compression setting of >8 scales down to just 727 KB (e.g., https://lichenportal.org/imglib/lichens/ASU/ASUL006/ASUL006158-014_print_1601394755.jpg) This means: the 6MB max file size is HUGE compared to most files currently served online. [example vascular plants: https://swbiodiversity.org/imglib/h_seinet/seinet/ASU/ASU0051/ASU0051478.jpg 6MB pixel dimensions of full herbarium sheet = 10 x15cm 300dpi saved at quality 8 = 925KB]
        iv. For optimal OCR output, the pixel height of lower case text characters should be a minimum of 15 pixels tall. If you zoom into the text, you can count the pixels. Pixel height or text is independent of the DPI.

6. **Recommendations for batch file image naming conventions are listed below).**
    i. File naming template: **barcode+separator+sequence#.jpg**,
        1. e.g., **ASUL006165-001.jpg** Generally, the new batch-upload tool will accept all previously established image upload routines and profiles.
    ii. The most IMPORTANT aspects are:
        1. File names should **start with the barcode** (= catalog number); that barcode should be a fixed amount of digits, e.g. ASUL123456 (not mixed amount of digits, one time ASUL123456, the next time ASUL1234567). Leading zeros (e.g., ASUL001234) can be used to accomplish this.
        2. **Barcodes and the remainder of the file name must be separated** by a unique character, e.g. an underscore "_", hyphen "-", if possible blank spaces should be avoided (e.g., "ASUL006165-001.jpg" and "ASUL006165_001.jpg" are both acceptable, but not " ASUL006165 001.jpg
        3. **File names must be unique**; uploading several images from the same specimen should all be using the same barcode, followed by a unique identifier [this can be a sequence number, time-stamp (date+time), or any other unique set of characters, e.g. **all these are permissible**:
        ASUL006165-001.jpg
        ASUL006165_001.jpg
        ASUL006165-A.jpg
        ASUL006165-2020-10-09_3-51pm.jpg
        ASUL006165-001_perianth.jpg
7. **Minimum recommended Metadata:** copyright (Creative Commons licenses recommended), author (xmp::creator)
8. **Image hosting policies: batch uploading all images to the new portal is encouraged**; institutions that prefer (or are obliged) to host all images on their own server are strongly advised to configure their own servers to **serve images using a secure protocol** (i.e., an image URL starting with https),; unsecure protocols are problematic because Google Chrome Browsers per default block insecure image links and most users will not be able to see these images in the Symbiota portals if they use Google Chrome at the default settings).

2. **Pace of work.** We estimated an average imaging rate of 50 specimens per hour when we submitted the grant. This includes labor paid for by the grant, estimates of volunteer hours, and savings through work study programs. We understand that this will be highly variable by herbarium and specimen storage type. We will be combining data from across all collaborators to calculate a rate that applies across the whole project. Please

make your imaging workflows as efficient and cost effective as possible, so that we can meet our deliverables for the grant.

1. Links to Collaboration Resources:
    i. Workflow Development Plan: https://docs.google.com/document/d/1_VRoTvq7VB4x6xOVzD2vgYftV16oWOvjMjpdMUibxq8/edit?usp=sharing
    ii. Groups with Shared Storage Types: https://docs.google.com/document/d/1K852Odk-Vt2NxrIxuC2dEGwuW5ktMBaHRp-6xC7X5C4/edit?usp=sharing

3. **Skeletal Data Capture - PRIORITY ACTIVITY: Jan 2021**
    1. Capture during imaging.
        i. **Barcode** - Many people have limited access to specimens due to COVID restrictions. Focusing on only capturing the barcode may increase imaging efficiency. Any of the other skeletal data listed below (Section ii) can also be captured during imaging, but this is left up to the individual institution to determine what is best for their workflow.
    2. Capture during imaging, prior to, or ASAP after loading the image and barcode into the portal. *Why are these fields important to include as skeletal data?* (1) Having the information listed below attached to all new non-North American specimens added in the portal as skeletal data will enable everyone to search for duplicate specimens already transcribed and/or georeferenced in our portals and in other international databases using new tools developed by the IT team. This will increase our transcription/georeferencing efficiency by leveraging specimen records that are already complete and attaching these data to our specimens. (2) These data will improve the efficiency for our centralized georeferencing at WIS and will enable Julie to organize CoGe communities of students, volunteers, and staff to focus on particular geographic regions. (3) These data will also enable us to have people with particular language expertise to help with transcribing non-English specimen labels.
        i. **Scientific Name** from the “filed as name” or most recent annotation; possible exception type specimens where basionym is preferred.
        ii. **Collector Name/s & Collection Number** (Capture **Collection Date** if collection number is not present.)
        iii. **Country** as listed on the label (Optional: **State**, only if easy capture due to geographic organization of specimens.)
        iv. **Language** as best can be determined.
        v. **Indicate if the specimen is part of an exsiccate.**
            1. Additional lists of international exsiccate are being added to the portals, so that there will be more options available to choose from a drop down list.
    3. Once all the skeletal data listed above is included in the portal, **update the record’s Processing Status to Stage 1.**

4. **Skeletal Data Tools** to add the Skeletal Data listed above either into a file that can be uploaded or directly entered into the portal.
    1. **In the Portals: Occurrence Skeletal Record Submission** {Recommended for adding skeletal data during imaging OR after imaging but before image upload to portal.}
        i. Video demo Skeletal Data Entry Tool in a Symbiota Portal - YouTube
        ii. Written instructions - can be downloaded and adapted based on how much data was captured during the imaging process at your institution. https://drive.google.com/file/d/1RcoPqsVv5leN7ID8HgpSuR9rDWjVJjii/view?usp=sharing
    2. **Image Collection Workflow Tool (LBCC Imaging App)** {Recommended for adding skeletal data during imaging.}
        i. Files needed to setup the App. https://drive.google.com/file/d/1EhY_CG7utBgEbCuac3fFvcPMiDuvYVIa/view?usp=sharing
        ii. Instructions for setting up the App. https://drive.google.com/file/d/1LFvpvGBxuc-NnFkXPDhdCFImXMWi-UHC/view?usp=sharing
        iii. Instructions for using the App begin on the bottom of page 5. https://drive.google.com/file/d/1iK8p8icR8wvECtY6QKsplhxb1jBcBZ1E/view?usp=sharing
        iv. Or see this video. https://drive.google.com/file/d/17d4tgf-Zha6NW0b5ipRgFsYgbyJ1N919/view?usp=sharing
    3. **In the Portals: Edit Existing Occurrence Records** {Recommended for adding skeletal data after image upload to the portal.}
    4. PhotoWatcher {Program to be developed by ASU.}
    5. Spreadsheet with dropdown lists {Not recommended.}
        i. Draft example adapted from TENN's SERNEC skeletal data sheet: https://docs.google.com/spreadsheets/d/1-R7hKUoq81FzeCCL2HLxON0Alq91o03hN6Vc4lr4jeQ/edit?usp=sharing

5. **Transcription Strategies** we will use to leverage the energy from institutions across the collaboration. Additional ideas will be discussed on Basecamp.
    1. First search for duplicate specimens already transcribed and/or georeferenced in our portals and other international databases using new tools developed by the IT team.
    2. Coordinating transcription across the collaboration.
        i. Focus on transcribing specimens from particular collectors or particular exsiccate that have duplicates deposited at multiple institutions, so that these are only transcribed once.
            1. Examples: TENN could transcribe all unique A.J. Sharp specimens. WIS has transcribed all their lichen and bryophyte exsiccate and are in the process of georeferencing these specimens. Then everyone with duplicates could connect these data to their specimens.
        ii. Focus on transcribing specimens from particular geographic regions to help Julie at WIS with centralized georeferencing.

        1. Example: All collaborators could focus for a period of time on transcribing specimens from southeast Asia, then Julie can create a CoGe community to engage volunteers from communities who have expertise in that region of the world to help with georeferencing.
    iii. Coordinate teams of non-English speakers to help with label transcription.
3. **GOAL** - Transcribe the label in full, including names from previous annotation labels.